How Uncertain Is the Grade? A Benchmark of Uncertainty Quantification for LLM-Based Automatic Assessment

Anonymous ACL submission

Abstract

Large language models (LLMs) are rapidly transforming automatic assessment in education by enabling flexible grading across diverse question types and response formats. However, their probabilistic nature introduces output uncertainty, which can undermine the reliability of downstream pedagogical decisions such as feedback generation and instructional support. To better understand this challenge, we benchmark a broad range of uncertainty quantification (UQ) methods for LLM-based automatic assessment. Although these methods have demonstrated effectiveness in domains such as question answering, their reliability in educational grading remains underexplored, particularly due to the ordinal nature of grading scores. To address this challenge, we propose an evaluation framework that considers not only grading correctness but also the magnitude of deviation from ground-truth scores. Through extensive analyses across multiple assessment datasets, LLM families, and decoding settings, we characterize uncertainty behaviors in grading scenarios and investigate how factors such as model family, task type, and decoding strategy influence uncertainty estimation. Our findings highlight both the strengths and limitations of existing UQ methods and provide practical insights for developing more reliable and uncertainty-aware grading systems. Code is available at `https://anonymous.4open.science/r/UQBench-1FF4/`.

1 Introduction

With the surge of Large language models (LLMs), LLM-powered automatic assessment has become increasingly popular in many domains (Gu et al., 2024; Li et al., 2024a). Automated assessment, also called automated grading, is a system that takes questions and candidate solutions, optionally with rubrics and other context information, and generates text to present the correctness of the evaluated solution (Ala-Mutka, 2005; Emirtekin, 2025). Its inputs include questions, answers, rubrics, and optionally demonstrations of solution-grade examples, while its outputs may include grades, aspect-level assessment (Low et al., 2019; Song et al., 2024a), and reasons for assessments (Cohn et al., 2024; Li et al., 2024b). Existing automated assessments are supported by rule-based (Braun et al., 2006; Cline et al., 2010) or supervised machine learning models (Ahmed et al., 2022; Lu and Cutumisu, 2021), which require intensive labor from domain experts for rule and feature design, dataset annotation, validation, and handling anomalies or out-of-distribution cases (Ala-Mutka, 2005; Paiva et al., 2022). Owing to LLMs' broad prior knowledge and instruction-following ability, LLM-based grading can handle diverse question domains, understand rubrics, and produce structured evaluations with interpretable rationales in natural language (Li et al., 2025a; Kortemeyer, 2023; Hashemi et al., 2024). Compared with traditional task-specific grading systems, LLM-based systems offer a unified pipeline capable of handling diverse inputs without training or adaptation, enabling their usage across domains and problem formulations in practical educational workflows.

Although automated assessment is efficient, the uncertainty issue has been a central challenge, and this is even more severe in LLM-based assessment. Uncertainty is the metric denoting how sure the models are when making judgments, with the causes of randomness, absence of precise boundaries of concepts, inaccurate base knowledge, and ambiguity between different objects (Mena et al., 2021). It leads to decreased performance and misleading outputs, especially when information or model abilities are inadequate. This issue becomes more severe in LLM-based frameworks because LLMs hide the internal mechanisms in large-scale parameters, rely on token probability and sampling, and tend to produce seemingly correct responses even without enough information and capabilities.

In conclusion, certainty is highly prioritized by automated assessment tasks, and responses of low certainty are harmful to both the learning procedures and the equality between students.

Uncertainty quantification (UQ) has been studied in various tasks, including classification (Kurz et al., 2022; Chlaily et al., 2023), regression (Levi et al., 2022; Abadie et al., 2020), and Bayesian inference settings (Mena et al., 2021), covering data in text and image modalities. However, many existing methods rely heavily on access to model structure and hidden states, including prediction probabilities represented by neural networks (Sensoy et al., 2018) and the feasible space of possibility distributions (Chlaily et al., 2023). These methods are inapplicable for LLM-based auto-assessment systems as the internal states of LLMs are computationally expensive, hard to handle, or inaccessible when using proprietary LLMs. Existing UQ methods for LLMs mainly include white-box methods that require logits, internal layer outputs, and model parameters (Vashurin et al., 2025); consistency-based methods that compare multiple generations (Zhang et al., 2024); and reflection-based methods that elicit LLMs to self-report or analyze certainty (Tian et al., 2023). Because their model access and settings differ, these works are not systematically compared, and few insights exist into how well they perform. As a prosperous research field, LLM-based auto-assessment has developed many works on modeling and implementing uncertainty, while a comprehensive investigation to validate and compare them is still ignored.

To bridge this gap, in this work, we propose a benchmark on uncertainty quantification for the auto-assessment tasks. The challenges of applying UQ methods to auto-assessment derive from the specificity in the wide range of domain knowledge, the finite set of ordinal output features, and the influences of input prompts. First, auto-assessment deploys general-use LLMs as a judge in a specialized domain, requiring benchmarking work to consider heterogeneous subjects rather than domains of LLMs' expertise only. Second, the expected output, i.e., the grade, is an ordinal variable whose labels have a relative order, but the distances between categories are not comparable. This prevents arithmetic operations on the grade labels and makes numeric metrics such as differences and mean squared error invalid. Similarly, metrics designed for semantic distance may not adequately capture output differences because both correct and wrong gradings are centered on the semantic representations of question contexts. Third, LLMs are sensitive to the input text, making many UQ methods incomparable when they use different templates along with the question and rubrics.

Taking these points in mind, we design the benchmark with multiple learning subjects, different forms of output grades, and unified templates to build the input prompts from questions and rubrics. We collect datasets that are both open-sourced and private, covering essay writing, natural science, chemistry, and math education. For the output grades, the grade sets include binary, ordered, multi-class, etc. In addition, we elicit the LLMs to judge with the text of reasons for the grades, incorporating the semantic features into the analysis. Finally, we use a set of curated-designed templates to build the prompts, forcing the input text similar to prior works or other datasets in this benchmark. To satisfy the need for broad usage and flexibility, we focus on repetition-based (or ensemble-based) methods in this work, which take several runs of generation to estimate uncertainty, without using internal states of LLMs in any way.

2 Related Work

2.1 LLM-based Automatic Assessment

Automatic assessment has long been studied before the emergence of LLMs. Earlier systems were typically built on task-specific machine learning models trained for particular grading formats, such as automated essay scoring, short-answer evaluation, or rubric-based classification (Ala-Mutka, 2005; Paiva et al., 2022; Higgins et al., 2005). These approaches generally relied on engineered linguistic features and supervised neural architectures for fixed domains or tasks. While effective in narrow settings, they require annotated datasets for each task, making it costly to scale assessment systems to new contexts. Besides, rigorous prompt tuning is needed to align the assessment with fine-grained evaluation criteria, limiting transferability across subjects, question types, and scoring criteria.

The advent of large language models has shifted this paradigm toward general-purpose evaluators that can assess with instructions and rubrics. A representative framework is LLM-as-a-Judge (Gu et al., 2024), where a prompted model assigns correctness labels (Zhu et al., 2023b; Liu et al., 2023; Gao et al., 2023) or reviews written works (Brake and Schaaf, 2024; Hosseini et al., 2024). The LLM-

as-a-Judge framework has been used for diverse tasks, including mathematics (Luo et al., 2023), natural and medical science (Zhao et al., 2024; Krolik et al., 2024), programming (Kumar et al., 2024), and language (Song et al., 2024b). Recent studies show that LLM evaluators can produce stable grading judgments across prompts and reasoning settings (Raina et al., 2024); rubric-guided prompting supports fine-grained scoring aligned with evaluation criteria (Deng et al., 2025); and aggregation frameworks such as SURE (Korthals et al., 2025) use self-consistency and majority voting to improve grading reliability. Along with prompt engineering, studies also refine rubrics to achieve fine-grained grades (Singh et al.; Chu et al., 2024). However, these studies reveal limitations: evaluation outcomes may change under minor input perturbations, including synonyms and meaning-invariant rephrasing (Zhu et al., 2023a; Raina et al., 2024; He et al., 2024a), symbols indicating grade labels (Ye et al., 2024), and the order of provided grade labels (Li et al., 2025b; Shi et al., 2024). These findings suggest that although LLM-based assessment improves coverage and flexibility, existing methods remain vulnerable to undesired variation in generated judgments, highlighting the need for uncertainty estimation to complement assessment decisions and mitigate unreliable outputs.

2.2 Uncertainty Estimation in LLMs

With the surge of LLM applications, inaccurate or non-factual generations have motivated uncertainty estimation as a practical signal of when outputs should or should not be trusted. A common starting point is to derive confidence from observable generation behavior, especially via repeated sampling. (Cole et al., 2023) shows that repetition-based metrics from multiple samples provide a reliable criterion for abstention decisions against uncertain responses, outperforming likelihood-based and self-verification alternatives. (Lyu et al., 2024) formalizes sample consistency into confidence measures based on agreement or entropy, demonstrating effectiveness for confidence calibration across models and reasoning tasks, while also exploring the influence of sampling size and explanation prompting. In addition, (Xiong et al., 2023) proposes self-probing, which elicits verbalized confidence from LLMs and helps with calibration and failure prediction, while also showing overconfident statements like humans. Complementing these studies, (Huang et al., 2023) conducts a large-scale empirical comparison across uncertainty estimators, LLMs, and domain tasks, showing that uncertainty signals can surface risky generations while exposing practical limitations. Beyond surface features, more sophisticated approaches refine the notion of equivalence used to aggregate samples. For open-ended generation, (Kuhn et al., 2023) introduces semantic entropy, which clusters responses by meaning invariance and quantifies uncertainty over semantic modes rather than token sequences. Recently, (Xia et al., 2025) synthesizes inference-time uncertainty estimation for LLMs by defining key uncertainty sources and organizing methods into a taxonomy including verbalized confidence, latent-information, consistency-based, and semantic clustering approaches.

In addition, the ordinal characteristic of grade labels makes the assessment task different from other tasks. Grades are usually presented as a label or value from a finite set; for example, in SemEval 2013 - Task 7 (Dzikovska et al., 2013), the grade labels indicate whether and to what degree a student answer aligns with the given reference: `3` for fully correct, `2` for partially correct, `1` for contradictory with reference, and `0` for irrelevant. Although the labels are ordered, with higher scores indicating a stronger tie with the reference, the distance between labels is incomparable or even nonsensical (Tellamekala et al., 2023; Xie et al., 2024). This body of work motivates our benchmark design. When evaluating these uncertainty signals in automated assessment, we focus on grading tasks whose predictions lie in a constrained label space.

3 Benchmark Design

3.1 Overview

As described in § 1, we focus on uncertainty estimation approaches based on repeated generation, since such methods can be uniformly applied to both open-source and proprietary LLMs without requiring access to internal model states. Following prior work (Kuhn et al., 2023; Vashurin et al., 2025), we define uncertainty as the variability in a model's predictions. In LLM-based automatic assessment, this task aims to measure how stable or inconsistent the model is when judging the same question, where outputs may include final assessment scores, intermediate reasoning steps, and supporting evidence. Depending on the uncertainty quantification method, our benchmark parses LLM-generated responses in different ways

and computes corresponding uncertainty values. We then evaluate these uncertainty estimates with seven metrics from three perspectives related to their downstream usage. By comparing the ranking results of different uncertainty quantification methods across multiple automatic assessment datasets and evaluation metrics, our benchmark provides a comprehensive analysis of uncertainty quantification methods for LLM-based automatic assessment tasks. An overview of the benchmark workflow is illustrated in Fig. 1.

3.2 Uncertainty Quantification Methods

Existing uncertainty quantification methods for LLMs are often designed for open-ended question answering or reasoning tasks (Xia et al., 2025), where uncertainty is estimated from the variability or equivalence of free-form generations. In this work, we study uncertainty quantification in LLM-based automatic assessment, where repeated generations for the same input usually include both final grading labels and intermediate grading rationales. Accordingly, we organize the evaluated methods into two groups: categorical-based methods, which operate on the distribution of predicted score labels, and relation-based methods, which estimate uncertainty from the relationships among generated responses. Detailed definitions and formulas of all methods are provided in Appx. C.

3.2.1 Categorical Based Methods

Categorical-based methods use the predefined and discrete output label space of automatic assessment. Given multiple grading outputs for the same question, these methods first extract the final score label from each output and then compute uncertainty from the empirical frequency distribution of the labels. We evaluate four methods in this category. **Numset** (Lyu et al., 2024) measures the number of unique labels produced across repeated generations, where more distinct labels indicate larger variability. **Max-Agree-Rate (MAR)** (Xiong et al., 2023) measures the dominance of the most frequent label and converts it into an uncertainty score, so lower agreement corresponds to higher uncertainty. **Categorical-Entropy (CE)** (Lyu et al., 2024; De Boer et al., 2005) computes entropy over the empirical label distribution and therefore considers both majority and minority responses. **First–Second Distance (FSD)** (Lyu et al., 2024) focuses on the gap between the most frequent and second most frequent labels, making it sensitive to cases where the model oscillates between two plausible scores.

3.2.2 Relation Based Methods

Categorical-based methods are efficient, but they only use the extracted final labels and may be unstable when the number of repeated generations is limited. Relation-based methods additionally exploit the textual content of generated outputs, especially the reasoning traces produced by CoT prompting (Wei et al., 2022) and reasoning-oriented models. These methods follow a two-step paradigm. First, they construct a relation graph over repeated outputs, where each node is a generated grading response, and each edge represents the pairwise relationship between two responses. We consider three relation functions: **Jaccard Similarity**, which captures lexical overlap; **Embedding Cosine Similarity** (Reimers and Gurevych, 2019), which captures continuous semantic relatedness; and **Entailment Score** (Kuhn et al., 2023; Lewis et al., 2020; Zhang et al., 2019), which captures logical consistency through NLI-based sentence-pair entailment.

Second, relation-based methods compute graph-level uncertainty from the constructed relation graph. We evaluate four property-driven uncertainty measures. **Normalized Average Degree (NAD)**, also known as variational ratio (VR) (Huang et al., 2023), measures the overall connectivity density of the graph, where weaker connectivity indicates higher uncertainty. **Graph Eccentricity (GE)** measures the global spread of responses through shortest-path distances. **Graph Eigenvalues (Eigen)** use spectral connectivity of the graph Laplacian to quantify how tightly the response graph is connected. **Discrete Semantic Entropy (DSE)** (Kuhn et al., 2023) clusters semantically equivalent responses and computes entropy over the induced cluster distribution. Together, these relation-based methods provide continuous uncertainty estimates by measuring whether repeated grading outputs form a compact and coherent response set or a dispersed and inconsistent one.

3.3 Evaluation Metrics

To comprehensively assess UQ methods for automatic assessment, we design our evaluation from three key perspectives: effectiveness, stability and correlation. For each perspective, we develop various evaluation metrics.

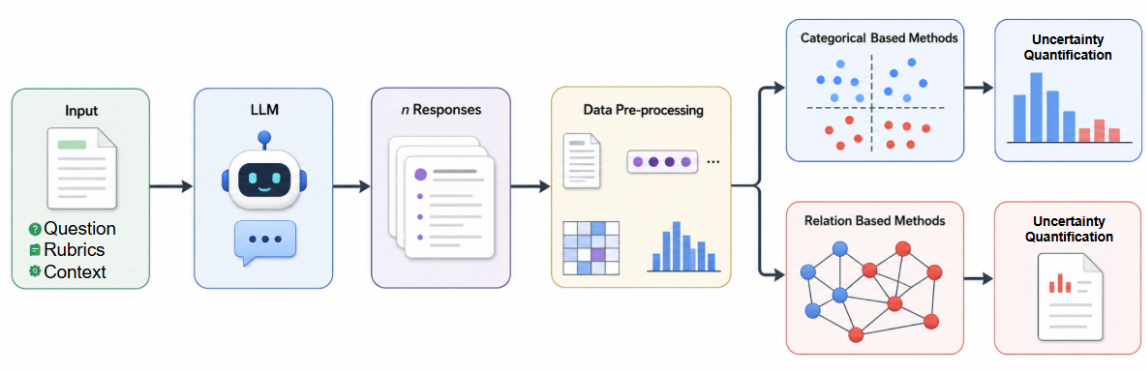


Figure 1: An overview of the benchmark workflow where n denotes the number of generation rollouts in this work.

3.3.1 Effectiveness

Uncertainty values are expected to indicate the reliability of responses produced by LLMs: responses with lower uncertainty should exhibit higher accuracy than those with higher uncertainty. In automatic assessment, this allows uncertainty to serve as a signal for human intervention and prevent low-reliability grading results from being released for downstream use. Based on this premise, we evaluate the effectiveness of each UQ method as an accuracy indicator using four metrics.

AUROC (Xia et al., 2025) evaluates whether uncertainty scores can distinguish correct and incorrect predictions. By treating incorrect ones as positive cases and uncertainty scores as ranking signals, AUROC measures the probability that an incorrect response receives higher uncertainty than a correct response. Higher AUROC indicates better discrimination between correct and incorrect predictions.

C-Index (Steck et al., 2007) is a rank-based metric for assessing consistency between predicted risk scores and observed outcomes. Since assessment scores are ordinal rather than purely binary, we use the C-index to evaluate whether responses with larger true errors receive higher uncertainty scores. Higher C-index indicates better alignment between uncertainty and error severity.

AUARC (Xia et al., 2025) measures the trade-off between prediction accuracy and rejection rate when selectively abstaining from low-confidence responses. The accuracy–rejection curve reports the accuracy of retained predictions as high-uncertainty responses are rejected. Higher AUARC indicates that rejecting a small fraction of uncertain responses yields larger accuracy gains, reflecting the utility of uncertainty for selective prediction and human-in-the-loop workflows.

AUERC (Simeone et al., 2011) is the error-based counterpart of AUARC, where the error-rejection curve plots the mean absolute error of retained predictions against the rejection rate. It evaluates whether high-uncertainty responses correspond to larger errors. Lower AUERC values are better, indicating that rejecting uncertain responses effectively reduces the remaining prediction error.

3.3.2 Stability

All uncertainty quantification methods in this work are derived from stochastic responses generated by LLMs, so the estimated uncertainty values are inherently noisy. Increasing the number of sampled responses can reduce this variance, but repeatedly generating many responses for the same grading request is computationally expensive and impractical in real-world deployment. To study this trade-off, we analyze how uncertainty estimates evolve as the response set size n increases and measure how quickly the values stabilize. Since uncertainty is mainly used for ranking responses, such as flagging high-uncertainty cases for human review, we also evaluate the stability of uncertainty-induced rankings using Spearman's rank correlation coefficient (Bishara and Hittner, 2012) between estimates computed from different response set sizes.

3.3.3 Correlation

Although different uncertainty quantification methods assess response uncertainty from diverse perspectives, they may still be correlated because of shared underlying evidence sources, such as categorical frequency or relational similarity. To summarize relationships among methods and help future work avoid redundant measures, we analyze pairwise correlations between methods using the Pearson correlation coefficient.

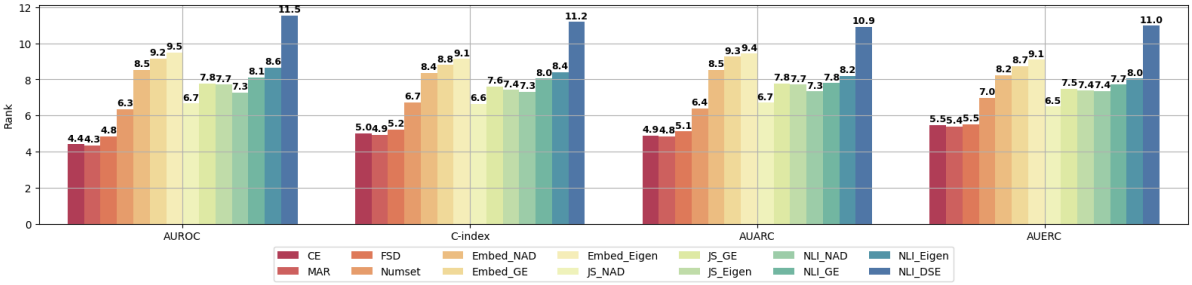
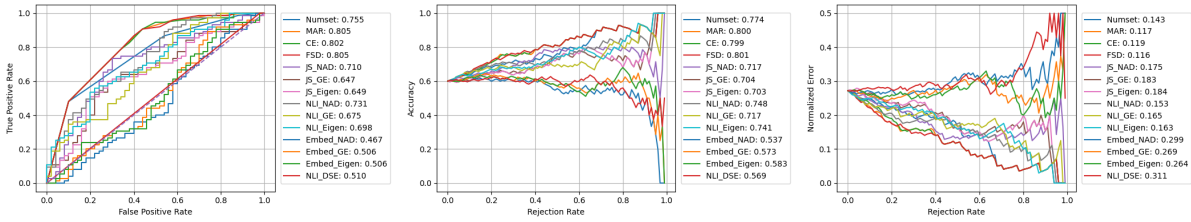
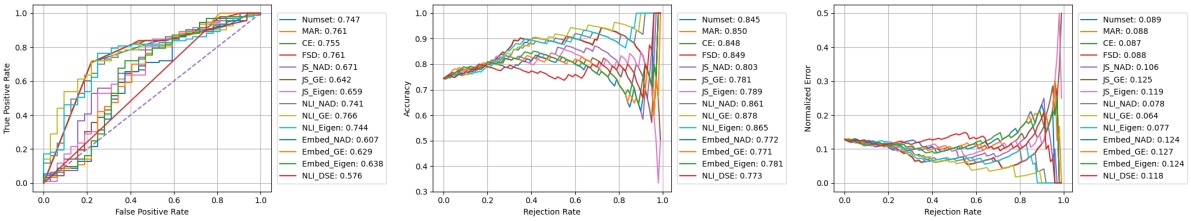


Figure 2: Average ranks of methods across questions, datasets, and generation strategies on effectiveness metrics.



(a) Ministral-3-14B (AUROC) (b) Ministral-3-14B (AUARC) (c) Ministral-3-14B (AUERC)



(d) gemini-3-flash (AUROC) (e) gemini-3-flash (AUARC) (f) gemini-3-flash (AUERC)

Figure 3: UQ Method Comparison between Ministral and Gemini over different metrics on the T1-PCK question.

4 Experiment

4.1 Datasets

To ensure the robustness of our conclusions, we benchmark the uncertainty quantification (UQ) methods on 14 grading questions sampled from five different sources, covering topics including essay writing, classroom Q&A, chemistry, earth science, and pedagogical knowledge. For each question, we collect 125 student responses spanning all score categories for subsequent experiments, resulting in a total of 1,750 samples. Detailed information about each question is provided in Appx. A.

4.2 LLM Settings

We benchmark a comprehensive set of mainstream LLMs, including three proprietary models: Claude-Haiku-4.5, Gemini-3-Flash, and GPT-5-Nano, and four open-source models: Qwen3-4B, Llama-3.1-8B, Ministral-3-14B, and DeepSeek-R1-Distill-Qwen-32B. Additional details about each model are provided in Appx. B. To ensure broad coverage of uncertainty quantification methods under realistic automatic assessment settings, we implement three commonly used generation strategies (Chu et al., 2024): (1) Zero-shot, (2) Zero-shot + Chain-of-Thought (CoT), and (3) Few-shot + Chain-of-Thought (CoT). During grading, each model and generation strategy is used to grade every response 15 times under the model's default generation settings. The resulting solution sets are then used to estimate uncertainty values with different UQ methods. More details are provided in Appx. D.

4.3 Effectiveness

Due to differences in score levels and response distributions across questions, the scale of uncertainty values produced by the same UQ method can vary substantially across datasets and tasks. To avoid large-magnitude uncertainty values dominating the evaluation, we rank each UQ method within each experimental configuration (model, dataset, and generation strategy) and aggregate results by averaging rankings across settings. Fig. 2 summarizes the aggregated rankings across the four evaluation metrics, while detailed results are provided in Appx. E. Overall, the categorical-based method CE achieves the best average ranking across all

metrics. This contrasts with prior UQ studies on open-ended Q&A tasks (Xia et al., 2025), where relation-based methods often perform better. Other categorical-based methods, including MAR, FSD, and Numset, also consistently outperform relation-based approaches, suggesting that the discrete output space of automatic assessment is particularly compatible with categorical uncertainty measures. To further investigate the influence of different factors, we conduct case studies on model choice, question characteristics, and generation settings. Fig. 3 compares Gemini and Ministral on the same dataset and generation condition. Although rankings remain relatively consistent across evaluation metrics within the same model, the preferred UQ methods differ substantially between models: Gemini favors relation-based methods, while Ministral shows stronger performance for category-based methods. Similar patterns are observed across datasets, whereas generation strategies have relatively limited influence. These findings suggest that UQ methods should be selected according to the underlying model and task conditions.

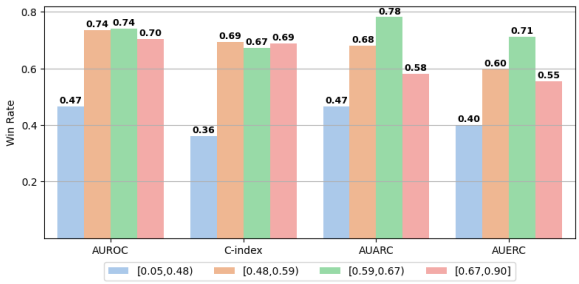


Figure 4: Win rates of category-based UQ over relation-based UQ across different task accuracy groups.

Finally, we explore the relationship between task accuracy and the preference for category- versus relation-based UQ methods. Based on Fig.2, we select MAR and JS_MAD as the best-performing category- and relation-based methods, respectively, and compare their win rates across different accuracy-level groups. As shown in Fig.4, relation-based methods achieve win rates above 50% when task accuracy is low, suggesting that relational information provides more reliable signals under difficult conditions. However, as accuracy increases, category-based methods begin to dominate, reaching win rates above 70%. Under high-accuracy conditions, the performance gap narrows again, with win rates around 55% for metrics such as AUARC and AUERC, indicating that both approaches become similarly effective.

4.4 Stability

In addition to effectiveness, we also evaluate different UQ methods from the perspective of stability, since obtaining many repeated generations for uncertainty estimation can be computationally expensive, especially for recent studies relying on high-cost proprietary LLMs. In Fig.5, we report the average number of repeats required to achieve 95% stability, defined as the change in the UQ value from n to n+1 repeats being less than 5%. More details about the stability metrics across different repeat numbers are provided in Appx.F. From the figure, we observe that relation-based methods using Jaccard similarity and embedding distance require fewer than five repeats on average to achieve stable uncertainty estimates, demonstrating the advantage of leveraging additional relational information compared to score-only UQ methods. In contrast, although category-based UQ methods exhibit relatively larger fluctuations in uncertainty values, their ranking differences across student responses stabilize quickly, requiring fewer than seven repeats on average. Considering practical deployment scenarios, if uncertainty decisions are made based on the relative ranking of responses within the same batch, category-based UQ methods still maintain strong overall performance. However, when uncertainty is determined using a fixed threshold, the higher variance in category-based uncertainty values may reduce consistency, making relation-based methods more stable and cost-efficient in practice.

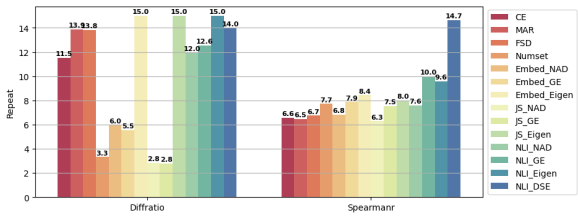


Figure 5: Stability comparison of uncertainty estimates produced by different UQ methods across repeated runs.

Similar to the effectiveness analysis, we further investigate how factors such as model choice, dataset, and generation strategy influence these conclusions. Specifically, Fig. 6 presents box plots illustrating the variance introduced by each factor. From the results, we draw conclusions consistent with the effectiveness study: model choice has the largest influence on UQ behavior, followed by the dataset, while the three generation-control strategies considered in this work have relatively limited

impact. Based on these observations, we suggest that UQ methods should be re-evaluated whenever the underlying model or dataset changes.

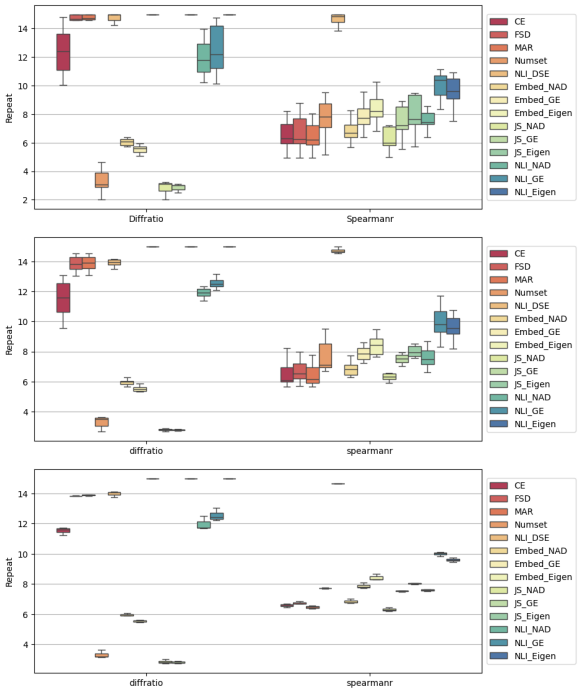
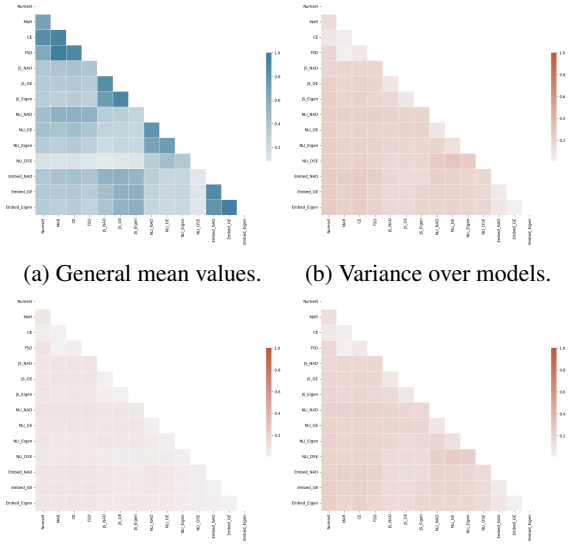


Figure 6: Distribution of the minimum number of repeats required to achieve stable UQ values across different models, questions, and generation strategies.

4.5 Correlation

To analyze correlations among methods, we compute the Pearson correlation coefficient for each pair of uncertainty quantification methods and aggregate these correlations across questions, models, and generation strategies by averaging. The resulting aggregated correlation matrix is shown in Fig. 7a. From the figure, we observe strong correlations among categorical-based methods, reflecting their reliance on similar frequency-based evidence. Among relation-based methods, those that share the same graph construction strategy exhibit particularly high correlations. In addition, methods derived from embedding-based and Jaccard similarity graphs show moderate correlation, as both capture semantic relatedness between responses, albeit at different representational levels. Similar to our analyses of effectiveness and stability, we examine the robustness of our correlation findings under variations in questions, models, and generation strategies. As shown in Fig. 7c, Fig. 7b and Fig. 7d, although variability exists across different models, questions, and strategies, the magnitude of this variance is relatively small (typically < 0.2) compared to the average correlation values (generally > 0.5). This indicates that the overall correlation patterns reported in our general results are robust to these sources of variation.



(a) General mean values. (b) Variance over models.

(c) Variance over questions. (d) Variance over strategics.

Figure 7: Mean and variance of aggregated pairwise correlation values among methods across factors.

5 Conclusion

In this work, we present a comprehensive benchmark of UQ methods for LLM-based automatic assessment. Although UQ has been widely studied in domains such as question answering, its application to automated assessment remains underexplored, particularly given the unique characteristics of assessment tasks, including categorical and ordinal outputs as well as rubric-driven evaluation settings. To address this gap, we establish a novel benchmark framework that categorizes existing UQ methods into two groups and systematically compares them from the perspectives of effectiveness, stability, and interrelationships. Through experiments across multiple assessment datasets spanning different subjects and LLMs of varying scales, our study reveals several important findings. Category-based methods demonstrate strong effectiveness in capturing uncertainty in assessment judgments, while relation-based methods exhibit greater robustness to sampling variation and provide more stable uncertainty estimates. Overall, our results suggest that no single UQ method is universally optimal, and that selecting an appropriate UQ metric should depend on the underlying LLM capability, assessment task, and evaluation setting.

Limitations

This work has three main limitations. First, our evaluation is limited to proprietary models and small-to-medium-scale open-source LLMs. While this choice reflects current computational constraints and realistic K–12 deployment conditions where large GPU resources are scarce, benchmarking larger open-source models will be essential as inference costs decrease. We argue that auto assessment by propriety and smaller-scale open-sourced LLMs is more practical, which is the reason we focus on LLMs out of large open-sourced models. Second, our assessment tasks are limited to questions with no more than 5 score levels and clear grading boundaries. Real-world assessment often involves more complex rubric settings (e.g., percentage scores, deduction-based, additive, or rubric items with dependency) that introduce different uncertainty patterns. However, we find the selected datasets are representative enough and of high quality, and these tasks are challenging enough for current LLMs. Thus, benchmarking the LLMs selected by this work is sufficient to showcase the auto-assessment advances in educational scenarios, and we leave these harder tasks for future work. Third, we focus on fixed grading criteria rather than personalized assessment based on student characteristics or teacher-specific rubrics. Although personalization is crucial for realistic scenarios, it requires substantial human-designed rubrics and gold labels, which increases grading inconsistency and raises privacy and ethical concerns. As the auto-assessment is still primarily deployed with fixed rubrics, and personalized assessment inherently shares the same capabilities of LLMs, we believe the results of evaluation with fixed rubrics are insightful enough, and will leave these personalized requirements for future exploration.

References

Alberto Abadie, Susan Athey, Guido W Imbens, and Jeffrey M Wooldridge. 2020. Sampling-based versus design-based uncertainty in regression analysis. *Econometrica*, 88(1):265–296.

Abbirah Ahmed, Arash Joorabchi, and Martin Hayes. 2022. On deep learning approaches to automated assessment: Strategies for short answer grading.

Kirsti M Ala-Mutka. 2005. A survey of automated assessment approaches for programming assignments. *Computer science education*, 15(2):83–102.

Anthropic. 2025. Introducing claude haiku 4.5. https://www.anthropic.com/news/claude-haiku-4-5. Large language model.

Anthony J Bishara and James B Hittner. 2012. Testing the significance of a correlation with nonnormal data: comparison of pearson, spearman, transformation, and resampling approaches. *Psychological methods*, 17(3):399.

Nathan Brake and Thomas Schaaf. 2024. Comparing two model designs for clinical note generation; is an llm a useful evaluator of consistency? In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 352–363.

Henry Braun, Isaac I Bejar, and David M Williamson. 2006. Rule based methods for automated scoring: Application in a licensing context. *Automated scoring of complex tasks in computer-based testing*, pages 83–122.

Saloua Chlaily, Debanshu Ratha, Pigi Lozou, and Andrea Marinoni. 2023. On measures of uncertainty in classification. *IEEE Transactions on Signal Processing*, 71:3710–3725.

Yucheng Chu, Hang Li, Kaiqi Yang, Harry Shomer, Hui Liu, Yasemin Copur-Gencturk, and Jiliang Tang. 2024. A llm-powered automatic grading framework with human-level guidelines optimization. *arXiv preprint arXiv:2410.02165*.

Ben E Cline, Carlyle C Brewster, and Richard D Fell. 2010. A rule-based system for automatically evaluating student concept maps. *Expert systems with applications*, 37(3):2282–2291.

Clayton Cohn, Nicole M. Hutchins, Tuan Le, and Gautam Biswas. 2024. A chain-of-thought prompting approach with llms for evaluating students' formative assessment responses in science. In *AAAI Conference on Artificial Intelligence*.

Jeremy Cole, Michael Zhang, Dan Gillick, Julian Eisenschlos, Bhuwan Dhingra, and Jacob Eisenstein. 2023. Selectively answering ambiguous questions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 530–543.

Yasemin Copur-Gencturk and Tammy Tolar. 2022. Mathematics teaching expertise: A study of the dimensionality of content knowledge, pedagogical content knowledge, and content-specific noticing skills. *Teaching and Teacher Education*, 114:103696.

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. 2019. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9268–9277.

Pieter-Tjerk De Boer, Dirk P Kroese, Shie Mannor, and Reuven Y Rubinstein. 2005. A tutorial on the cross-entropy method. *Annals of operations research*, 134(1):19–67.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Haotian Deng, Chris Farber, Jiyoon Lee, and David Tang. 2025. Rubric-conditioned llm grading: Alignment, uncertainty, and robustness. *arXiv preprint arXiv:2601.08843*.

Myroslava O Dzikovska, Rodney Nielsen, Chris Brew, Claudia Leacock, Danilo Giampiccolo, Luisa Bentivogli, Peter Clark, Ido Dagan, and Hoa Trang Dang. 2013. Semeval-2013 task 7: The joint student response analysis and 8th recognizing textual entailment challenge. In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 263–274.

Emrah Emirtekin. 2025. Large language model-powered automated assessment: A systematic review. *Applied Sciences*.

Mingqi Gao, Jie Ruan, Renliang Sun, Xunjian Yin, Shiping Yang, and Xiaojun Wan. 2023. Human-like summarization evaluation with chatgpt. *arXiv preprint arXiv:2304.02554*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, and 1 others. 2024. A survey on llm-as-a-judge. *The Innovation*.

Ben Hamner, Jaison Morgan, lynnvandev, Mark Shermis, and Tom Vander Ark. 2012. The hewlett foundation: Automated essay scoring. https://kaggle.com/competitions/asap-aes. Kaggle Competition.

Helia Hashemi, Jason Eisner, Corby Rosset, Benjamin Van Durme, and Chris Kedzie. 2024. Llm-rubric: A multidimensional, calibrated approach to automated evaluation of natural language texts. In *Annual Meeting of the Association for Computational Linguistics*.

Jia He, Mukund Rungta, David Koleczek, Arshdeep Sekhon, Franklin X Wang, and Sadid A. Hasan. 2024a. Does prompt formatting have any impact on llm performance? *ArXiv*, abs/2411.10541.

Peng He, Namsoo Shin, and Joseph Krajcik. 2024b. School level. *Handbook of Research on Science Learning Progressions*.

Colin A Higgins, Geoffrey Gray, Pavlos Symeonidis, and Athanasios Tsintsifas. 2005. Automated assessment and experiences of teaching programming. *Journal on Educational Resources in Computing (JERIC)*, 5(3):5–es.

Pedram Hosseini, Jessica M Sin, Bing Ren, Bryceton G Thomas, Elnaz Nouri, Ali Farahanchi, and Saeed Hassanpour. 2024. A benchmark for long-form medical question answering. *arXiv preprint arXiv:2411.09834*.

Yuheng Huang, Jiayang Song, Zhijie Wang, Shengming Zhao, Huaming Chen, Felix Juefei-Xu, and Lei Ma. 2023. Look before you leap: An exploratory study of uncertainty measurement for large language models. *arXiv preprint arXiv:2307.10236*.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, and 1 others. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

Gerd Kortemeyer. 2023. Performance of the pre-trained large language model gpt-4 on automated short answer grading. *Discover Artificial Intelligence*, 4.

Luke Korthals, Emma Akrong, Gali Geller, Hannes Rosenbusch, Raoul Grasman, and Ingmar Visser. 2025. Towards reliable llm grading through self-consistency and selective human review: Higher accuracy, less work.

Jack Krolik, Herprit Mahal, Feroz Ahmad, Gaurav Trivedi, and Bahador Saket. 2024. Towards leveraging large language models for automated medical q&a evaluation. *arXiv preprint arXiv:2409.01941*.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *arXiv preprint arXiv:2302.09664*.

Abhishek Kumar, Sonia Haiduc, Partha Pratim Das, and Partha Pratim Chakrabarti. 2024. Llms as evaluators: A novel approach to evaluate bug report summarization. *arXiv preprint arXiv:2409.00630*.

Alexander Kurz, Katja Hauser, Hendrik Alexander Mehrtens, Eva Krieghoff-Henning, Achim Hekler, Jakob Nikolas Kather, Stefan Fröhling, Christof Von Kalle, and Titus Josef Brinker. 2022. Uncertainty estimation in medical image classification: systematic review. *JMIR Medical Informatics*, 10(8):e36427.

Dan Levi, Liran Gispan, Niv Giladi, and Ethan Fetaya. 2022. Evaluating and calibrating uncertainty prediction in regression tasks. *Sensors*, 22(15):5540.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 7871–7880.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, Kai Shu, Lu Cheng, and Huan Liu. 2024a. From generation to judgment: Opportunities and challenges of llm-as-a-judge. In *Conference on Empirical Methods in Natural Language Processing*.

Hai Li, Chenglu Li, Wanli Xing, Sami Baral, and Neil T. Heffernan. 2024b. Automated feedback for student math responses based on multi-modality and fine-tuning. *Proceedings of the 14th Learning Analytics and Knowledge Conference*.

Hang Li, Yucheng Chu, Kaiqi Yang, Yasemin Copur-Gencturk, and Jiliang Tang. 2025a. Llm-based automated grading with human-in-the-loop. *2025 IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE)*, pages 1–8.

Qingquan Li, Shaoyu Dou, Kailai Shao, Chao Chen, and Haixiang Hu. 2025b. Evaluating scoring bias in llm-as-a-judge. *ArXiv*, abs/2506.22316.

Alexander H Liu, Kartik Khandelwal, Sandeep Subramanian, Victor Jouault, Abhinav Rastogi, Adrien Sadé, Alan Jeffares, Albert Jiang, Alexandre Cahill, Alexandre Gavaudan, and 1 others. 2026. Ministral 3. *arXiv preprint arXiv:2601.08584*.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 2511–2522.

Daniel Mark Low, Kate H Bentley, and Satrajit S. Ghosh. 2019. Automated assessment of psychiatric disorders using speech: A systematic review. *Laryngoscope Investigative Otolaryngology*, 5:96 – 116.

Chang Lu and Maria Cutumisu. 2021. Integrating deep learning into an automated feedback generation system for automated essay scoring. *International Educational Data Mining Society*.

Haipeng Luo, Qingfeng Sun, Can Xu, Pu Zhao, Jianguang Lou, Chongyang Tao, Xiubo Geng, Qingwei Lin, Shifeng Chen, and Dongmei Zhang. 2023. Wizardmath: Empowering mathematical reasoning for large language models via reinforced evol-instruct. *arXiv preprint arXiv:2308.09583*.

Qing Lyu, Kumar Shridhar, Chaitanya Malaviya, Li Zhang, Yanai Elazar, Niket Tandon, Marianna Apidianaki, Mrinmaya Sachan, and Chris Callison-Burch. 2024. Calibrating large language models with sample consistency. *ArXiv*, abs/2402.13904.

José Mena, Oriol Pujol, and Jordi Vitrià. 2021. A survey on uncertainty estimation in deep learning classification systems from a bayesian perspective. *ACM Computing Surveys (CSUR)*, 54(9):1–35.

José Carlos Paiva, José Paulo Leal, and Álvaro Figueira. 2022. Automated assessment in computer science education: A state-of-the-art review. *ACM Transactions on Computing Education (TOCE)*, 22(3):1–40.

Vyas Raina, Adian Liusie, and Mark Gales. 2024. Is llm-as-a-judge robust? investigating universal adversarial attacks on zero-shot llm assessment. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7499–7517.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 3982–3992.

Murat Sensoy, Lance Kaplan, and Melih Kandemir. 2018. Evidential deep learning to quantify classification uncertainty. *Advances in neural information processing systems*, 31.

Lin Shi, Chiyu Ma, Wenhua Liang, Xingjian Diao, Weicheng Ma, and Soroush Vosoughi. 2024. Judging the judges: A systematic study of position bias in llm-as-a-judge.

Paolo Simeone, Claudio Marrocco, and Francesco Tortorella. 2011. Shaping the error-reject curve of error correcting output coding systems. In *International Conference on Image Analysis and Processing*, pages 118–127. Springer.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, and 1 others. 2025. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*.

Juhi Singh, Ziqiao Ao, and Sebastian Antinome. Optimizing prompt refinement: Algorithmic strategies for large language model-based text classification. *Available at SSRN 5525975*.

Hwanjun Song, Hang Su, Igor Shalyminov, Jason (Jinglun) Cai, and Saab Mansour. 2024a. Finesure: Fine-grained summarization evaluation using llms. *ArXiv*, abs/2407.00908.

Yishen Song, Qianta Zhu, Huaibo Wang, and Qinhua Zheng. 2024b. Automated essay scoring and revising based on open-source large language models. *IEEE Transactions on Learning Technologies*, 17:1880–1890.

Harald Steck, Balaji Krishnapuram, Cary Dehing-Oberije, Philippe Lambin, and Vikas C Raykar. 2007. On ranking in survival analysis: Bounds on the concordance index. *Advances in neural information processing systems*, 20.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan

Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Qwen Team. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Mani Kumar Tellamekala, Shahin Amiriparian, Björn W Schuller, Elisabeth André, Timo Giesbrecht, and Michel Valstar. 2023. Cold fusion: Calibrated and ordinal latent distribution fusion for uncertainty-aware multimodal emotion recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(2):805–822.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D Manning. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5433–5442.

Roman Vashurin, Ekaterina Fadeeva, Artem Vazhentsev, Lyudmila Rvanova, Daniil Vasilev, Akim Tsvigun, Sergey Petrakov, Rui Xing, Abdelrahman Sadallah, Kirill Grishchenkov, and 1 others. 2025. Benchmarking uncertainty quantification methods for large language models with lm-polygraph. *Transactions of the Association for Computational Linguistics*, 13:220–248.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Zhiqiu Xia, Jinxuan Xu, Yuqian Zhang, and Hang Liu. 2025. A survey of uncertainty estimation methods on large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 21381–21396.

Zhuyang Xie, Yan Yang, Jie Wang, Xiaorong Liu, and Xiaofan Li. 2024. Trustworthy multimodal fusion for sentiment analysis in ordinal sentiment space. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(8):7657–7670.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2023. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. *arXiv preprint arXiv:2306.13063*.

Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, Nitesh V. Chawla, and Xiangliang Zhang. 2024. Justice or prejudice? quantifying biases in llm-as-a-judge. *ArXiv*, abs/2410.02736.

Caiqi Zhang, Fangyu Liu, Marco Basaldella, and Nigel Collier. 2024. Luq: Long-text uncertainty quantification for llms. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5244–5262.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Tianjie Zhao, Sheng Wang, Chaojun Ouyang, Min Chen, Chenying Liu, Jin Zhang, Long Yu, Fei Wang, Yong Xie, Jun Li, and 1 others. 2024. Artificial intelligence for geoscience: Progress, challenges, and perspectives. *The Innovation*, 5(5).

Yanqi Zhou, Tao Lei, Hanxiao Liu, Nan Du, Yanping Huang, Vincent Zhao, Andrew M Dai, Quoc V Le, James Laudon, and 1 others. 2022. Mixture-of-experts with expert choice routing. *Advances in Neural Information Processing Systems*, 35:7103–7114.

Kaijie Zhu, Jindong Wang, Jiaheng Zhou, Zichen Wang, Hao Chen, Yidong Wang, Linyi Yang, Weirong Ye, Neil Zhenqiang Gong, Yue Zhang, and Xingxu Xie. 2023a. Promptrobust: Towards evaluating the robustness of large language models on adversarial prompts. *Proceedings of the 1st ACM Workshop on Large AI Systems and Models with Privacy and Safety Analysis*.

Lianghui Zhu, Xinggang Wang, and Xinlong Wang. 2023b. Judgelm: Fine-tuned large language models are scalable judges. *arXiv preprint arXiv:2310.17631*.

A Dataset Detail

To provide a comprehensive evaluation, we collect 14 grading questions from five different sources for our experiments, including two open-source datasets and three private datasets. Specifically, the two open-source datasets are the Automated Student Assessment Prize (ASAP) dataset (Hamner et al., 2012), which focuses on short-essay grading (150–550 words) written by students in Grades 7–10, and SemEval-2013 Task 7 (SemEval-13-T7) (Dzikovska et al., 2013), which scores student responses to questions collected from in-class exercises, tests, and tutorial dialogues. For the three private datasets, we manually collect two sets of questions from elementary and middle school classes covering chemistry and general science following the 3DLP standard (He et al., 2024b), and one set from adult learners focusing on pedagogical skills (Copur-Gencturk and Tolar, 2022). For the open-source datasets, we use the original ground-truth labels to evaluate the correctness of LLM-based automatic grading. For the private datasets, the ground-truth labels are obtained via agreement between two expert human graders. In cases of disagreement, a third grader adjudicates to determine the final label. Based on these labels, we compute the correctness of the automatic assessment results. In Tab. 1, we summarize the statistics and characteristics of each dataset and the corresponding grading questions. For each question, we randomly sample 125 answers covering all score categories for evaluation.

Table 1: List of questions evaluated by our benchmark.

Question	Score	Subject	Question	Score	Subject
ASAP	0-3	Essay Writing	Semveal	0-3	Mixture
T1-DCI	0-3	Chemistry	T1-SEP	0-3	Chemistry
T2-DCI	0-3	Chemistry	T2-SEP	0-3	Chemistry
T1-CK	0-3	Math Pedagogy	T1-PCK	0-3	Math Pedagogy
T2-CK	0-3	Math Pedagogy	T2-PCK	0-3	Math Pedagogy
U41Q1	0-4	Earth Science	U41Q2	0-4	Earth Science
U42Q1	0-4	Earth Science	U42Q2	0-4	Earth Science

B LLM Detail

To ensure the robustness of our conclusions, we benchmark a comprehensive set of mainstream LLMs, including seven closed-source API-based models from three leading vendors, i.e., OpenAI, Anthropic, and Google, and 4 open-source models released by research teams worldwide. For the closed-source models, we choose the most common used models by prior LLM-based automatic assessment studies (Chu et al., 2024) to ensure the validenss of the conclusion. For the open-source models, we cover recent representative variants of standard LLMs, including mixture-of-experts models (Zhou et al., 2022) and explicitly reasoning-oriented models (Jaech et al., 2024). With this broad model coverage, we aim to ensure that our benchmark results generalize across diverse deployment scenarios of automatic assessment. Tab. 2 summarizes the details of all evaluated models.

Table 2: List of LLMs evaluated by our benchmark.

Proprietary Models		
Model	**Model ID**	**Scale**
Calude Haiku 4.5 (Anthropic, 2025)	claude-haiku-4-5-20251001	/
Gemini 3 Flash (Team et al., 2023)	gemini-3-flash-preview	/
GPT-5 nano (Singh et al., 2025)	gpt-5-nano	/
Open-source Models		
Model	**Model ID**	**Scale**
Qwen3-4B (Team, 2025)	Qwen3-4B-Instruct-2507	4B
Llama-3.1-8B (Grattafiori et al., 2024)	Llama-3.1-8B-Instruct	8B
Ministral-3-14B (Liu et al., 2026)	Ministral-3-14B-Instruct-2512	14B
DeepSeek-R1-Distill-Qwen-32B (DeepSeek-AI, 2025)	DeepSeek-R1-Distill-Qwen-32B	32B

C Details of Uncertainty Quantification Methods

C.1 Categorical Based Methods

Existing studies on uncertainty quantification have primarily focused on open-ended question answering (Q&A) or reasoning tasks (Xia et al., 2025), where model responses are free-form. In contrast, automatic assessment is typically formulated as a classification or scoring task, in which model outputs are constrained to a fixed label or score space. Therefore, we select uncertainty quantification methods (Xia et al., 2025; Lyu et al., 2024; Kuhn et al., 2023) that can capture uncertainty from categorical outputs. Meanwhile, with the recent adoption of chain-of-thought (CoT) prompting (Wei et al., 2022) and the emergence of reasoning-oriented models, model outputs in automatic assessment may include not only final grading scores but also intermediate grading rationales. To capture uncertainty manifested in these intermediate representations, we further incorporate relation-based uncertainty quantification methods.

Let $\mathcal{O} = \{o_i\}_{i=1}^{N}$ denote the set of repeated grading outputs for the same input, where N is the

number of repeated generations. For categorical-based methods, each o_i is mapped to its final assessment score label, and $\mathcal{O}_u$ denotes the set of unique labels observed across repeated generations. Owing to the pre-defined and discrete output label space, methods in this category are relatively simple to compute and computationally efficient. In general, they derive uncertainty from the empirical frequency distribution of score labels observed across repeated queries to the same request.

Numset. Numset (Lyu et al., 2024) measures uncertainty as the number of unique answers observed across multiple samples:

$$U_{\text{Numset}} = |\mathcal{O}_u|. \tag{1}$$

A larger $|\mathcal{O}_u|$ indicates greater variability in the model's outputs and thus higher uncertainty.

Max-Agree-Rate (MAR). MAR (Xiong et al., 2023) captures uncertainty by incorporating the full distribution of answer frequencies. Compared to Numset, which only counts the number of distinct answers, MAR reflects the dominance of the most frequent answer among all sampled responses. A larger agreement rate indicates that the model consistently produces the same answer and is therefore more confident. To express this as an uncertainty measure, we use the complement of the maximum empirical frequency:

$$U_{\text{MAR}} = 1 - \max_{o \in \mathcal{O}_u} \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}[o_i = o]. \tag{2}$$

A larger U_{MAR} indicates lower agreement and higher uncertainty.

Categorical Entropy (CE). Categorical Entropy (Lyu et al., 2024) is a classical distribution chaos quantification method over output class probabilities for classification problems (De Boer et al., 2005). As automatic assessment is also commonly treated as a classification task, we compute entropy over the empirical distribution of predicted grade labels:

$$U_{\text{CE}} = \sum_{o \in \mathcal{O}_u} -p_o \log(p_o), \quad p_o = \frac{1}{N} \sum_{i}^{N} \mathbf{1}[o_i = o]. \tag{3}$$

Here, p_o is the normalized frequency of each unique answer $o \in \mathcal{O}_u$. Compared to MAR, CE not only captures the majority class but also accounts for all categories, so changes in minority responses are reflected in the uncertainty value. A higher U_{CE} indicates a more chaotic answer distribution and thus higher uncertainty.

First-Second Distance (FSD). FSD (Lyu et al., 2024) is a frequency-based uncertainty measure that addresses limitations of commonly used methods such as MAR and Entropy. It quantifies the gap between the proportions of samples supporting the most frequent answer $\bar{o}$ and the second most frequent answer $\bar{\bar{o}}$. To align this measure with our uncertainty formulation, where larger values indicate higher uncertainty, we use the complement of this gap:

$$U_{\text{FSD}} = 1 - \frac{1}{N} \left(\sum_{i=1}^{N} \mathbf{1}[o_i = \bar{o}] - \sum_{i=1}^{N} \mathbf{1}[o_i = \bar{\bar{o}}] \right). \tag{4}$$

Compared with MAR and CE, FSD focuses on the relative dominance between the top two categories. It is particularly informative when the model is uncertain between two plausible answers, since it avoids overconfidence induced by considering only the most-voted answer.

C.2 Relation Based Methods

Although evaluating uncertainty directly in the categorical label space is computationally convenient, it has notable limitations. Due to the discrete nature of label categories in automatic assessment tasks, estimated uncertainty values can exhibit large fluctuations when the number of sampled responses N is small. For example, entropy-based measures such as CE are particularly sensitive to sampling noise before the response distribution stabilizes (Cui et al., 2019). In practical automatic assessment settings, repeatedly querying an LLM for the same input is constrained by efficiency and cost considerations, making such instability especially problematic.

The recent emergence of reasoning-path generation in large language models provides rich intermediate signals that help mitigate the limitations of category-based uncertainty quantification. Specifically, by modeling semantic similarity (Reimers and Gurevych, 2019) or logical entailment (Kuhn et al., 2023) among the reasoning traces associated with grading outputs, relation-based uncertainty quantification methods enable more fine-grained and continuous assessments of model uncertainty. These methods construct a relation graph $\mathcal{G} = (V, E)$ over the output set $\mathcal{O}$, where each ver-

tex $v \in V$ corresponds to one generated grading response and each edge $e_{ij} \in E$ is weighted by the pairwise relationship between responses o_i and o_j. A compact and coherent graph indicates lower uncertainty, whereas a dispersed graph indicates higher uncertainty.

C.2.1 Relation Graph Construction

Inspired by graph-based clustering analysis, recent studies (Huang et al., 2023; Xia et al., 2025) construct a relation graph over generated outputs and measure the tightness of relationships among responses. A key factor influencing the analysis lies in how this relation graph is constructed: different definitions of pairwise relationships yield different notions of answer similarity and capture uncertainty at varying semantic depths and computational costs. In this work, we consider three representative relation functions.

Jaccard Similarity. Jaccard Similarity measures lexical overlap between two responses based on their token sets:

$$s_{ij}^{\text{jac}} = \frac{|T(o_i) \cap T(o_j)|}{|T(o_i) \cup T(o_j)|}, \tag{5}$$

where $T(o)$ denotes the set of tokens in response o. This approach captures surface-level similarity and is computationally efficient, but it is limited in its ability to reflect deeper semantic equivalence between paraphrased or lexically diverse answers.

Embedding Cosine Similarity. Embedding Cosine Similarity (Reimers and Gurevych, 2019) computes the cosine similarity between vector representations of responses obtained from a sentence embedding model:

$$s_{ij}^{\text{emb}} = \frac{\mathbf{h}_i^\top \mathbf{h}_j}{\|\mathbf{h}_i\|_2 \, \|\mathbf{h}_j\|_2}, \quad \mathbf{h}_i = f_{\text{emb}}(o_i). \tag{6}$$

Here, $\mathbf{h}_i \in R^d$ denotes the semantic embedding of response o_i produced by a pre-trained encoder f_{emb}. By operating in a continuous semantic embedding space, this method captures higher-level semantic relatedness beyond exact lexical overlap, at the cost of additional computational overhead for embedding extraction.

Entailment Score. Entailment Score (Kuhn et al., 2023) models pairwise relationships using natural language inference (NLI) by estimating the degree to which one answer entails another. Off-the-shelf NLI models output class probabilities over entailment, neutral, and contradiction (Lewis et al., 2020). However, the limited input context window makes them unsuitable for directly processing long, multi-step reasoning traces produced by LLMs. To address this, we adopt a BERTScore-inspired strategy (Zhang et al., 2019): each response is split into sentences, entailment scores are computed over sentence pairs, and a mean–max aggregation is applied. Following prior work (Kuhn et al., 2023), we use the entailment probability to construct the relation graph. Since entailment is directional, we symmetrize the pairwise relationship by averaging both directions:

$$\begin{aligned} s_{i\to j}^{\text{nli}} &= \frac{1}{M} \sum_{m=1}^{M} \max_k P(\text{entail} \mid o_{i,m}, o_{j,k}), \\ s_{ij}^{\text{nli}} &= \frac{1}{2}\big(s_{i\to j}^{\text{nli}} + s_{j\to i}^{\text{nli}}\big). \end{aligned} \tag{7}$$

Here, $P(\cdot)$ denotes the entailment probability predicted by the NLI model; M and K are the numbers of sentences after splitting answers o_i and o_j, respectively; and $o_{i,m}$ and $o_{j,k}$ denote the m-th and k-th sentences of answers o_i and o_j. Entailment scores capture fine-grained semantic and logical consistency between responses, but they are computationally expensive due to NLI inference over sentence pairs.

C.2.2 Property-driven Uncertainty

Given the relation graph $\mathcal{G}$, different graph-theoretic properties can be leveraged to quantify its tightness. Below, we describe four representative property-driven uncertainty measures that have been widely adopted in prior work on uncertainty quantification for Q&A and reasoning tasks (Xia et al., 2025). Let s_{ij}^* denote the pairwise similarity between responses o_i and o_j obtained from one of the relation graph construction methods.

Normalized Average Degree (NAD). Normalized Average Degree, also known as the variational ratio (VR) (Huang et al., 2023), measures the connectivity density of the relation graph. Intuitively, if each answer is closely connected to many others, the response set forms a dense and compact cluster, indicating low uncertainty. We define the NAD-based uncertainty as:

$$U_{\text{NAD}} = 1 - \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{i \neq j} s_{ij}^*. \tag{8}$$

Larger values of U_{NAD} indicate weaker local connectivity and thus higher uncertainty.

Graph Eccentricity (GE). Graph Eccentricity quantifies the maximum shortest-path distance from each node to all other nodes in the graph and summarizes the global spread of responses:

$$U_{\mathrm{ECC}} = \frac{1}{N} \sum_i \max_j [d^{\mathrm{sp}}_{ij}]. \tag{9}$$

Here, d^{sp} denotes the all-pairs shortest-path distance induced by the complement of similarity, $1 - s^*_{ij}$. Smaller values indicate that all answers are mutually close in the relation graph, whereas larger values imply that some answers are distant from the rest, reflecting higher uncertainty.

Graph Eigenvalues (Eigen). Graph Eigenvalues use spectral properties of the graph Laplacian to provide a global measure of connectivity. In particular, the second smallest eigenvalue, $\lambda_2(L)$, known as algebraic connectivity, reflects how tightly connected the graph is. We define the eigenvalue-based uncertainty as:

$$U_{\mathrm{Eigen}} = \frac{1}{\lambda_2(L)}, \quad L = D - A, \quad A_{ij} = s^*_{ij}. \tag{10}$$

Here, L is the graph Laplacian constructed from the degree matrix D and adjacency matrix A. A more tightly connected graph exhibits larger $\lambda_2(L)$ and thus lower uncertainty, whereas smaller $\lambda_2(L)$ corresponds to higher uncertainty.

Discrete Semantic Entropy (DSE). Discrete Semantic Entropy extends entropy-based uncertainty measures to the relation-graph setting by computing entropy over semantically clustered responses (Kuhn et al., 2023). By grouping answers into equivalence classes based on pairwise relations in $\mathcal{G}$ and computing the entropy of the induced discrete distribution, DSE captures both the diversity of semantic modes and their relative frequencies. Specifically, we define:

$$U_{\mathrm{DSE}} = -\sum_{m=1}^{M} p_m \log p_m, \quad p_m = \frac{|\mathcal{C}_m|}{N}. \tag{11}$$

Here, $\{\mathcal{C}_m\}_{m=1}^{M}$ denotes the clusters obtained from the relation graph $\mathcal{G}$, and $|\mathcal{C}_m|$ is the size of cluster $\mathcal{C}_m$. As clustering over $\mathcal{G}$ incurs additional computational cost, we follow prior work (Kuhn et al., 2023) and compute this metric only for the NLI-based relation graph using bidirectional entailment relations. Larger U_{DSE} indicates a more fragmented semantic landscape and thus higher uncertainty.

D Setting Detail

To ensure broad coverage of uncertainty quantification methods under realistic automatic assessment settings, we implement three widely used generation strategies adopted in recent automatic assessment studies (Chu et al., 2024).

Zero-shot is the most straightforward way to leverage LLMs for automatic assessment. Given a question, a student answer, and a grading rubric, the LLM is prompted to directly produce a score without any demonstrations or intermediate reasoning steps.

Zero-shot + Chain of Thought (COT) extends zero-shot by eliciting intermediate reasoning steps from the LLM. By decomposing the grading process into a step-by-step reasoning trajectory, the model is better able to align with the grading criteria specified in the rubric and to analyze student responses more comprehensively, often resulting in improved grading performance.

Few-shot + Chain of Thought (COT) further augments the prompt with exemplar answers corresponding to different score levels. These demonstrations provide concrete interpretations of the abstract criteria in the rubric, helping the model better internalize grading standards and produce more accurate and consistent assessment outcomes. Specifically, in our evaluation, we include two demonstration examples for each score level.

Finally, to ensure the robustness of our conclusions, we set the number of repeated grading samples for each answer to fifteen and use the default generation parameters recommended by each model for every generation. The prompts used for each query setting are presented as follows: Fig. 8 shows the zero-shot prompt, Fig. 10 presents the zero-shot + CoT prompt, and Fig. **??** illustrates the few-shot + CoT prompt.

E Effectiveness Result

We present the detailed metric values and ranking results of different UQ methods across datasets, LLMs, and generation controls in the following tables. Due to the large volume of results, we report the results for each LLM separately. Specifically, Tab.3 presents the results for GPT-5-Nano, Tab.4 for Gemini-3-Flash, Tab.5 for Claude-Haiku-4.5, Tab.6 for Qwen3-4B, Tab.7 for Llama-3.1-8B, Tab.8 for Ministral-3-14B, and Tab. 9 for

```
As a virtual teaching assistant, your role
involves assisting in the evaluation of
student assignment submissions.  Please be
impartial and only use the grading guidelines,
the student`s submission, and the task in your
evaluation.
Analyze the submission and evaluate how well
the student`s submission answers the task.
Grade this submission of a student in 10th,
with a score in {grade_range}
# COT Command:
  In this task, you should give the response
directly WITHOUT using the chain-of-thought or
solving the problem step by step.
# Question (the students will work on):
  {question}
# Prompt
  {grading_rubrics}
# Submission (to be evaluated):
  {student_answer}
# Output Format:
  `text (reasons for your decisions)...  ##
Final: final_score`.
```

Figure 8: Prompt template for Zero-shot without COT Setting.

DeepSeek-R1-Distill-Qwen-32B. In addition to the detailed results, we further visualize the influence of different factors, including model choice, dataset, and generation control, on the ranking behavior of UQ methods. Specifically, Fig. 11a illustrates the ranking variance introduced by model differences, Fig. 11b shows the variance caused by datasets, and Fig. 11c presents the variance under different generation controls. From these plots, we observe that data choice is the most influential factor affecting UQ method rankings, while model also introduce noticeable variation. In contrast, the influence of generation-control strategies is relatively limited.

F Stability Result

To visualize how stability changes with the number of repetitions, we present the two stability metrics in Fig. 12. From the figure, we observe that for the DiffRatio metric, UQ methods based on embedding relations achieve less than a 5 change ratio at a very early stage, while methods using NLI-based relations exhibit a much slower convergence speed. This finding highlights the importance of selecting appropriate relation metrics in UQ studies. For ranking-based metrics such as Spearman correlation, most UQ methods converge within a similar number of repetitions, and all methods reach the

Table 3: Metric value and ranking results of each uncertainty quantification methods with GPT-5-nano over 14 questions.

AUROC

Control	Question	CE	Embed_Eigen	Embed_GE	Embed_NAD	FSD	JS_Eigen	JS_GE	JS_NAD	MAR	NLI_DSE	NLI_Eigen	NLI_GE	NLI_NAD	Numset
shot=0, cot=0	ASAP	0.742(2)	0.647(10)	0.636(12)	0.586(14)	0.733(5)	0.667(9)	0.673(8)	0.638(11)	0.737(4)	0.601(13)	0.741(3)	0.725(6)	0.704(7)	0.752(1)
	Semeval	0.706(1)	0.696(6)	0.703(3)	0.704(2)	0.699(5)	0.677(10)	0.681(8)	0.686(7)	0.700(4)	0.575(13)	0.644(11)	0.611(12)	0.565(14)	0.680(9)
	T1-CK	0.658(1)	0.612(6)	0.603(8)	0.593(9)	0.650(4)	0.559(12)	0.540(14)	0.607(7)	0.654(3)	0.548(13)	0.631(5)	0.590(10)	0.560(11)	0.656(2)
	T1-DCI	0.586(5)	0.433(13)	0.431(14)	0.434(12)	0.587(3)	0.537(10)	0.546(9)	0.570(7)	0.586(4)	0.481(11)	0.593(2)	0.550(8)	0.639(1)	0.577(6)
	T1-PCK	0.543(11)	0.565(6)	0.563(7)	0.562(8)	0.537(13)	0.602(3)	0.604(2)	0.661(1)	0.540(12)	0.554(9)	0.543(10)	0.567(5)	0.508(14)	0.578(4)
	T1-SEP	0.656(3)	0.623(9)	0.633(8)	0.645(5)	0.639(7)	0.617(10)	0.659(2)	0.651(4)	0.644(6)	0.392(14)	0.488(11)	0.432(12)	0.429(13)	0.669(1)
	T2-CK	0.630(2)	0.536(6)	0.529(10)	0.533(7)	0.608(4)	0.532(8)	0.529(9)	0.543(5)	0.614(3)	0.477(13)	0.512(11)	0.501(12)	0.438(14)	0.632(1)
	T2-DCI	0.594(6)	0.514(12)	0.503(13)	0.452(14)	0.595(5)	0.551(10)	0.550(11)	0.558(9)	0.595(4)	0.603(3)	0.561(8)	0.623(2)	0.657(1)	0.580(7)
	T2-PCK	0.614(5)	0.590(9)	0.591(8)	0.622(4)	0.609(7)	0.676(3)	0.696(2)	0.707(1)	0.610(6)	0.543(14)	0.590(10)	0.582(12)	0.575(13)	0.584(11)
	T2-SEP	0.465(13)	0.550(1)	0.529(2)	0.488(7)	0.467(11)	0.505(5)	0.520(3)	0.477(10)	0.467(12)	0.493(6)	0.485(8)	0.481(9)	0.506(4)	0.459(14)
	U41Q1	0.636(4)	0.541(10)	0.554(9)	0.590(6)	0.651(1)	0.486(13)	0.499(12)	0.479(14)	0.650(2)	0.511(11)	0.583(7)	0.576(8)	0.637(3)	0.600(5)
	U41Q2	0.664(3)	0.603(13)	0.613(12)	0.616(11)	0.660(4)	0.626(8)	0.628(7)	0.588(14)	0.669(2)	0.626(10)	0.626(8)	0.656(5)	0.673(1)	0.634(6)
	U42Q1	0.595(2)	0.580(3)	0.573(4)	0.568(6)	0.558(9)	0.469(14)	0.470(13)	0.477(12)	0.559(8)	0.519(11)	0.541(10)	0.559(7)	0.570(5)	0.621(1)
	U42Q2	0.588(2)	0.435(9)	0.434(10)	0.404(12)	0.535(4)	0.402(13)	0.376(14)	0.413(11)	0.557(3)	0.514(6)	0.479(8)	0.522(5)	0.483(7)	0.605(1)
shot=0, cot=1	ASAP	0.697(4)	0.575(9)	0.577(8)	0.544(14)	0.697(3)	0.557(12)	0.559(11)	0.572(10)	0.699(1)	0.548(13)	0.697(5)	0.614(6)	0.610(7)	0.698(2)
	Semeval	0.727(1)	0.706(6)	0.711(5)	0.727(2)	0.713(4)	0.648(11)	0.656(10)	0.691(8)	0.718(3)	0.574(14)	0.666(9)	0.631(13)	0.633(12)	0.705(7)
	T1-CK	0.701(2)	0.572(13)	0.594(7)	0.590(8)	0.700(3)	0.589(9)	0.581(10)	0.597(6)	0.703(1)	0.534(14)	0.647(5)	0.577(11)	0.574(12)	0.684(4)
	T1-DCI	0.527(9)	0.456(13)	0.451(14)	0.459(12)	0.527(11)	0.632(4)	0.640(3)	0.648(1)	0.527(9)	0.545(7)	0.643(2)	0.581(6)	0.624(5)	0.545(8)
	T1-PCK	0.540(5)	0.507(12)	0.517(10)	0.476(14)	0.536(7)	0.579(3)	0.581(2)	0.622(1)	0.537(6)	0.524(9)	0.531(8)	0.509(11)	0.480(13)	0.554(4)
	T1-SEP	0.752(1)	0.671(7)	0.677(6)	0.680(5)	0.735(4)	0.558(10)	0.553(11)	0.563(9)	0.737(3)	0.551(12)	0.643(8)	0.540(13)	0.516(14)	0.738(2)
	T2-CK	0.558(3)	0.473(12)	0.470(13)	0.481(11)	0.556(4)	0.515(8)	0.525(7)	0.506(10)	0.558(2)	0.529(6)	0.532(5)	0.508(9)	0.405(14)	0.593(1)
	T2-DCI	0.598(9)	0.496(12)	0.489(13)	0.477(14)	0.606(6)	0.589(10)	0.599(8)	0.620(4)	0.601(7)	0.610(5)	0.678(1)	0.642(3)	0.668(2)	0.568(11)
	T2-PCK	0.608(8)	0.584(10)	0.604(9)	0.612(4)	0.610(7)	0.642(3)	0.659(2)	0.679(1)	0.611(6)	0.523(14)	0.575(11)	0.556(13)	0.611(5)	0.564(12)
	T2-SEP	0.517(4)	0.543(1)	0.541(2)	0.520(3)	0.517(8)	0.513(9)	0.517(4)	0.490(10)	0.517(6)	0.456(14)	0.489(12)	0.474(13)	0.517(6)	0.489(11)
	U41Q1	0.649(3)	0.534(9)	0.534(10)	0.577(7)	0.656(2)	0.502(12)	0.515(11)	0.479(13)	0.658(1)	0.476(14)	0.582(6)	0.567(8)	0.642(4)	0.597(5)
	U41Q2	0.657(4)	0.532(12)	0.558(9)	0.576(8)	0.660(3)	0.530(13)	0.533(11)	0.508(14)	0.662(2)	0.554(10)	0.579(7)	0.619(5)	0.676(1)	0.619(6)
	U42Q1	0.607(6)	0.576(11)	0.596(9)	0.612(3)	0.614(2)	0.611(4)	0.615(1)	0.599(8)	0.610(5)	0.544(14)	0.554(13)	0.582(10)	0.602(7)	0.574(12)
	U42Q2	0.598(3)	0.389(14)	0.416(13)	0.458(12)	0.606(2)	0.507(9)	0.497(11)	0.535(5)	0.608(1)	0.551(4)	0.502(10)	0.509(8)	0.523(7)	0.533(6)
shot=2, cot=1	ASAP	0.678(4)	0.610(10)	0.617(9)	0.586(13)	0.667(6)	0.573(14)	0.586(12)	0.590(11)	0.669(5)	0.631(8)	0.683(2)	0.682(3)	0.658(7)	0.686(1)
	Semeval	0.790(3)	0.716(8)	0.722(5)	0.719(6)	0.793(1)	0.686(10)	0.698(9)	0.718(7)	0.793(2)	0.542(14)	0.650(11)	0.631(12)	0.624(13)	0.730(4)
	T1-CK	0.806(2)	0.654(12)	0.644(13)	0.671(9)	0.805(3)	0.674(8)	0.702(6)	0.721(4)	0.810(1)	0.590(14)	0.668(10)	0.664(11)	0.687(7)	0.717(5)
	T1-DCI	0.499(7)	0.347(13)	0.340(14)	0.350(12)	0.500(4)	0.412(11)	0.414(10)	0.457(9)	0.499(6)	0.499(5)	0.558(2)	0.556(3)	0.648(1)	0.492(8)
	T1-PCK	0.602(9)	0.663(6)	0.670(4)	0.660(7)	0.587(11)	0.690(3)	0.704(2)	0.710(1)	0.593(10)	0.505(14)	0.606(8)	0.546(12)	0.532(13)	0.669(5)
	T1-SEP	0.762(4)	0.754(9)	0.756(8)	0.772(1)	0.760(7)	0.771(2)	0.762(5)	0.734(10)	0.761(6)	0.494(14)	0.691(11)	0.613(13)	0.638(12)	0.770(3)
	T2-CK	0.818(1)	0.666(11)	0.670(9)	0.703(6)	0.814(3)	0.675(8)	0.699(7)	0.711(5)	0.817(2)	0.525(14)	0.668(10)	0.595(13)	0.638(12)	0.746(4)
	T2-DCI	0.672(3)	0.480(13)	0.466(14)	0.488(12)	0.663(5)	0.639(10)	0.643(9)	0.661(7)	0.669(4)	0.603(11)	0.734(1)	0.691(2)	0.663(6)	0.659(8)
	T2-PCK	0.601(2)	0.559(10)	0.567(8)	0.576(5)	0.600(4)	0.569(7)	0.575(6)	0.609(1)	0.600(3)	0.525(13)	0.513(14)	0.535(12)	0.555(11)	0.566(9)
	T2-SEP	0.575(8)	0.585(6)	0.573(9)	0.554(11)	0.589(3)	0.622(2)	0.625(1)	0.589(4)	0.585(5)	0.583(7)	0.532(14)	0.542(12)	0.569(10)	0.540(13)
	U41Q1	0.737(2)	0.566(14)	0.584(11)	0.630(8)	0.737(3)	0.572(13)	0.596(10)	0.612(9)	0.743(1)	0.581(12)	0.659(6)	0.676(5)	0.730(4)	0.647(7)
	U41Q2	0.645(5)	0.644(6)	0.652(4)	0.639(7)	0.654(2)	0.567(11)	0.572(10)	0.575(9)	0.658(1)	0.481(14)	0.564(12)	0.543(13)	0.653(3)	0.601(8)
	U42Q1	0.646(2)	0.571(10)	0.570(11)	0.594(8)	0.608(6)	0.565(12)	0.562(13)	0.574(9)	0.627(4)	0.512(14)	0.606(7)	0.620(5)	0.635(3)	0.647(1)
	U42Q2	0.645(3)	0.374(12)	0.373(13)	0.361(14)	0.663(2)	0.439(10)	0.420(11)	0.496(9)	0.668(1)	0.532(7)	0.516(8)	0.568(6)	0.572(4)	0.569(5)

C-index

Control	Question	CE	Embed_Eigen	Embed_GE	Embed_NAD	FSD	JS_Eigen	JS_GE	JS_NAD	MAR	NLI_DSE	NLI_Eigen	NLI_GE	NLI_NAD	Numset
shot=0, cot=0	ASAP	0.650(4)	0.603(10)	0.590(11)	0.535(14)	0.642(6)	0.611(9)	0.617(8)	0.566(13)	0.644(5)	0.589(12)	0.666(1)	0.659(3)	0.639(7)	0.664(2)
	Semeval	0.706(1)	0.696(6)	0.703(3)	0.704(2)	0.699(5)	0.677(10)	0.681(8)	0.686(7)	0.700(4)	0.575(13)	0.644(11)	0.611(12)	0.565(14)	0.680(9)
	T1-CK	0.615(3)	0.584(6)	0.572(9)	0.565(10)	0.606(5)	0.532(13)	0.517(14)	0.572(8)	0.611(4)	0.549(11)	0.616(2)	0.574(7)	0.544(12)	0.622(1)
	T1-DCI	0.586(5)	0.433(13)	0.431(14)	0.434(12)	0.587(3)	0.537(10)	0.546(9)	0.570(7)	0.586(4)	0.481(11)	0.593(2)	0.550(8)	0.639(1)	0.577(6)
	T1-PCK	0.549(10)	0.565(6)	0.564(8)	0.564(7)	0.543(13)	0.606(3)	0.607(2)	0.660(1)	0.546(11)	0.551(9)	0.544(12)	0.572(5)	0.515(14)	0.584(4)
	T1-SEP	0.590(7)	0.607(6)	0.617(2)	0.632(1)	0.577(10)	0.584(8)	0.616(3)	0.615(4)	0.581(9)	0.407(13)	0.470(11)	0.414(12)	0.386(14)	0.608(5)
	T2-CK	0.634(1)	0.543(6)	0.535(9)	0.542(7)	0.613(4)	0.536(8)	0.533(10)	0.544(5)	0.618(3)	0.492(13)	0.517(11)	0.507(12)	0.451(14)	0.632(2)
	T2-DCI	0.546(7)	0.498(12)	0.486(13)	0.433(14)	0.550(5)	0.520(11)	0.520(10)	0.523(9)	0.549(6)	0.622(2)	0.559(4)	0.621(3)	0.647(1)	0.535(8)
	T2-PCK	0.611(4)	0.588(10)	0.585(12)	0.611(4)	0.607(7)	0.682(3)	0.700(2)	0.709(1)	0.608(6)	0.558(14)	0.598(8)	0.588(11)	0.590(9)	0.580(13)
	T2-SEP	0.472(13)	0.551(1)	0.530(2)	0.492(7)	0.474(11)	0.513(4)	0.529(3)	0.487(9)	0.474(12)	0.502(6)	0.490(8)	0.486(10)	0.513(4)	0.462(14)
	U41Q1	0.616(4)	0.540(10)	0.553(9)	0.587(5)	0.630(1)	0.465(13)	0.480(12)	0.462(14)	0.629(2)	0.520(11)	0.580(7)	0.569(8)	0.628(3)	0.586(6)
	U41Q2	0.632(3)	0.612(13)	0.622(6)	0.622(8)	0.623(5)	0.614(11)	0.620(9)	0.583(14)	0.632(4)	0.617(10)	0.622(6)	0.650(1)	0.649(2)	0.612(12)
	U42Q1	0.550(4)	0.563(2)	0.560(3)	0.545(5)	0.523(9)	0.470(12)	0.468(13)	0.462(14)	0.521(10)	0.517(11)	0.533(7)	0.527(8)	0.545(6)	0.583(1)
	U42Q2	0.591(2)	0.433(9)	0.430(10)	0.405(13)	0.543(4)	0.415(12)	0.389(14)	0.426(11)	0.564(3)	0.518(6)	0.487(8)	0.531(5)	0.495(7)	0.603(1)
shot=0, cot=1	ASAP	0.627(5)	0.552(10)	0.556(8)	0.522(14)	0.628(4)	0.554(9)	0.550(12)	0.546(13)	0.629(3)	0.550(11)	0.642(1)	0.596(6)	0.581(7)	0.633(2)
	Semeval	0.727(1)	0.706(6)	0.711(5)	0.727(2)	0.713(4)	0.648(11)	0.656(10)	0.691(8)	0.718(3)	0.574(14)	0.666(9)	0.631(13)	0.633(12)	0.705(7)
	T1-CK	0.642(3)	0.548(10)	0.568(6)	0.566(7)	0.642(2)	0.558(8)	0.547(11)	0.556(9)	0.644(1)	0.484(14)	0.596(5)	0.531(12)	0.531(12)	0.630(4)
	T1-DCI	0.511(9)	0.445(13)	0.441(14)	0.448(12)	0.511(9)	0.620(4)	0.627(3)	0.635(1)	0.511(9)	0.538(7)	0.632(2)	0.568(6)	0.609(5)	0.529(8)
	T1-PCK	0.543(5)	0.507(12)	0.516(10)	0.474(14)	0.539(7)	0.583(3)	0.585(2)	0.626(1)	0.541(6)	0.521(9)	0.536(8)	0.515(11)	0.488(13)	0.554(4)
	T1-SEP	0.716(1)	0.651(7)	0.658(6)	0.662(5)	0.700(4)	0.537(10)	0.533(11)	0.547(9)	0.702(3)	0.509(12)	0.603(8)	0.502(13)	0.487(14)	0.706(2)
	T2-CK	0.570(3)	0.493(12)	0.492(13)	0.503(11)	0.568(4)	0.530(8)	0.540(5)	0.515(10)	0.570(2)	0.539(6)	0.536(7)	0.517(9)	0.426(14)	0.600(1)
	T2-DCI	0.547(10)	0.476(12)	0.470(13)	0.460(14)	0.554(8)	0.565(7)	0.569(6)	0.591(4)	0.549(9)	0.579(5)	0.628(1)	0.605(3)	0.628(2)	0.527(11)
	T2-PCK	0.607(7)	0.585(10)	0.603(9)	0.605(8)	0.610(6)	0.641(3)	0.658(2)	0.678(1)	0.611(5)	0.538(14)	0.578(11)	0.560(13)	0.619(4)	0.561(12)
	T2-SEP	0.523(6)	0.548(1)	0.546(2)	0.527(3)	0.523(6)	0.522(9)	0.526(4)	0.499(10)	0.523(6)	0.466(14)	0.496(11)	0.483(13)	0.525(5)	0.491(12)
	U41Q1	0.656(3)	0.549(9)	0.549(10)	0.588(6)	0.658(2)	0.504(12)	0.515(11)	0.477(13)	0.662(1)	0.470(14)	0.587(7)	0.579(8)	0.651(4)	0.606(5)
	U41Q2	0.587(4)	0.506(10)	0.526(9)	0.546(7)	0.594(2)	0.475(12)	0.475(12)	0.435(14)	0.592(3)	0.495(11)	0.531(8)	0.564(5)	0.612(1)	0.557(6)
	U42Q1	0.562(8)	0.548(11)	0.568(6)	0.591(1)	0.569(5)	0.574(2)	0.573(4)	0.556(10)	0.564(7)	0.535(14)	0.542(12)	0.558(9)	0.573(3)	0.541(13)
	U42Q2	0.544(5)	0.391(14)	0.412(13)	0.447(12)	0.566(1)	0.520(6)	0.510(7)	0.550(3)	0.558(2)	0.545(4)	0.501(9)	0.491(10)	0.509(8)	0.483(11)
shot=2, cot=1	ASAP	0.647(4)	0.621(10)	0.629(8)	0.603(12)	0.635(7)	0.596(14)	0.606(11)	0.603(13)	0.636(6)	0.624(9)	0.665(1)	0.659(3)	0.639(5)	0.659(2)
	Semeval	0.790(3)	0.716(8)	0.722(5)	0.719(6)	0.793(1)	0.686(10)	0.698(9)	0.718(7)	0.793(2)	0.542(14)	0.650(11)	0.631(12)	0.624(13)	0.730(4)
	T1-CK	0.785(3)	0.633(12)	0.625(13)	0.654(9)	0.786(2)	0.656(8)	0.683(6)	0.699(5)	0.790(1)	0.569(14)	0.650(10)	0.641(11)	0.663(7)	0.700(4)
	T1-DCI	0.499(7)	0.347(13)	0.340(14)	0.350(12)	0.500(4)	0.412(11)	0.414(10)	0.457(9)	0.499(5)	0.499(5)	0.558(2)	0.556(3)	0.648(1)	0.492(8)
	T1-PCK	0.607(8)	0.658(6)	0.665(4)	0.656(7)	0.593(11)	0.682(3)	0.697(2)	0.704(1)	0.598(10)	0.502(14)	0.604(9)	0.543(12)	0.536(13)	0.665(5)
	T1-SEP	0.729(7)	0.730(5)	0.733(4)	0.747(1)	0.729(8)	0.735(3)	0.726(9)	0.704(10)	0.729(6)	0.487(14)	0.658(11)	0.587(13)	0.619(12)	0.737(2)
	T2-CK	0.791(1)	0.649(11)	0.652(9)	0.684(6)	0.786(3)	0.664(8)	0.684(6)	0.697(5)	0.791(2)	0.516(14)	0.651(10)	0.584(13)	0.621(12)	0.718(4)
	T2-DCI	0.612(5)	0.451(12)	0.437(14)	0.442(13)	0.609(7)	0.595(10)	0.600(8)	0.614(4)	0.612(6)	0.589(11)	0.664(1)	0.641(3)	0.647(2)	0.595(9)
	T2-PCK	0.607(2)	0.549(11)	0.557(10)	0.566(8)	0.603(4)	0.575(6)	0.581(5)	0.621(1)	0.604(3)	0.521(14)	0.523(13)	0.544(12)	0.568(7)	0.565(9)
	T2-SEP	0.581(8)	0.588(7)	0.576(9)	0.560(11)	0.595(3)	0.628(2)	0.631(1)	0.595(3)	0.591(5)	0.588(6)	0.539(14)	0.549(12)	0.576(10)	0.540(13)
	U41Q1	0.718(3)	0.587(12)	0.605(9)	0.648(6)	0.712(4)	0.568(14)	0.590(11)	0.595(10)	0.721(1)	0.582(13)	0.648(6)	0.666(5)	0.721(1)	0.639(8)
	U41Q2	0.598(7)	0.625(2)	0.635(1)	0.615(3)	0.600(6)	0.548(11)	0.550(10)	0.555(9)	0.601(4)	0.448(14)	0.530(12)	0.505(13)	0.601(5)	0.569(8)
	U42Q1	0.611(5)	0.559(9)	0.557(10)	0.571(8)	0.586(7)	0.534(12)	0.530(13)	0.545(11)	0.598(6)	0.514(14)	0.614(4)	0.621(2)	0.626(1)	0.615(3)
	U42Q2	0.577(3)	0.340(12)	0.335(13)	0.313(14)	0.607(1)	0.390(10)	0.378(11)	0.451(9)	0.596(2)	0.504(7)	0.466(8)	0.514(6)	0.517(5)	0.519(4)

AUARC

Control	Question	CE	Embed_Eigen	Embed_GE	Embed_NAD	FSD	JS_Eigen	JS_GE	JS_NAD	MAR	NLI_DSE	NLI_Eigen	NLI_GE	NLI_NAD	Numset
shot=0, cot=0	ASAP	0.528(5)	0.456(11)	0.440(12)	0.395(14)	0.526(7)	0.496(9)	0.504(8)	0.488(10)	0.526(6)	0.425(13)	0.564(1)	0.559(2)	0.542(3)	0.531(4)
	Semeval	0.781(1)	0.747(6)	0.744(7)	0.748(5)	0.780(3)	0.740(9)	0.740(8)	0.739(10)	0.781(2)	0.683(11)	0.681(12)	0.672(13)	0.638(14)	0.771(4)
	T1-CK	0.682(1)	0.675(5)	0.663(8)	0.630(11)	0.679(4)	0.628(12)	0.610(14)	0.656(9)	0.680(2)	0.619(13)	0.671(6)	0.670(7)	0.631(10)	0.680(3)
	T1-DCI	0.672(3)	0.565(12)	0.555(14)	0.559(13)	0.672(2)	0.620(10)	0.628(9)	0.650(8)	0.672(3)	0.613(11)	0.666(6)	0.656(7)	0.698(1)	0.672(5)
	T1-PCK	0.570(6)	0.568(8)	0.558(10)	0.582(5)	0.567(9)	0.634(3)	0.645(2)	0.683(1)	0.569(7)	0.546(11)	0.525(13)	0.532(12)	0.515(14)	0.589(4)
	T1-SEP	0.735(1)	0.700(9)	0.708(8)	0.723(7)	0.729(3)	0.697(10)	0.726(5)	0.725(6)	0.731(2)	0.572(12)	0.589(11)	0.558(13)	0.544(14)	0.727(4)
	T2-CK	0.636(1)	0.601(10)	0.602(9)	0.602(8)	0.628(3)	0.611(6)	0.605(7)	0.622(5)	0.630(2)	0.545(12)	0.548(11)	0.533(13)	0.495(14)	0.622(4)
	T2-DCI	0.501(6)	0.436(12)	0.427(13)	0.403(14)	0.501(5)	0.469(11)	0.478(10)	0.485(9)	0.501(7)	0.511(4)	0.532(3)	0.591(1)	0.571(2)	0.494(8)
	T2-PCK	0.575(7)	0.610(4)	0.605(6)	0.607(5)	0.573(9)	0.640(3)	0.648(2)	0.669(1)	0.574(8)	0.490(14)	0.551(12)	0.565(11)	0.545(13)	0.566(10)
	T2-SEP	0.374(14)	0.434(3)	0.410(8)	0.387(10)	0.375(13)	0.421(7)	0.433(4)	0.404(9)	0.375(12)	0.428(5)	0.442(1)	0.422(6)	0.441(2)	0.378(11)
	U41Q1	0.748(3)	0.646(12)	0.653(9)	0.672(8)	0.752(1)	0.623(14)	0.630(13)	0.648(11)	0.752(2)	0.652(10)	0.725(6)	0.702(7)	0.739(4)	0.726(5)
	U41Q2	0.730(5)	0.711(7)	0.708(10)	0.700(12)	0.731(4)	0.710(8)	0.708(9)	0.676(14)	0.731(3)	0.684(13)	0.718(6)	0.734(2)	0.749(1)	0.704(11)
	U42Q1	0.599(2)	0.569(5)	0.554(6)	0.545(7)	0.584(3)	0.475(14)	0.484(13)	0.489(11)	0.583(4)	0.487(12)	0.514(10)	0.537(8)	0.528(9)	0.625(1)
	U42Q2	0.705(2)	0.647(9)	0.644(10)	0.601(12)	0.681(4)	0.605(11)	0.591(13)	0.580(14)	0.689(3)	0.656(6)	0.654(7)	0.667(5)	0.648(8)	0.719(1)
shot=0, cot=1	ASAP	0.524(4)	0.438(12)	0.432(13)	0.402(14)	0.524(3)	0.445(11)	0.445(10)	0.446(9)	0.524(2)	0.462(8)	0.527(1)	0.466(7)	0.479(6)	0.522(5)
	Semeval	0.793(1)	0.749(7)	0.748(8)	0.759(6)	0.791(3)	0.747(9)	0.739(10)	0.768(5)	0.792(2)	0.675(14)	0.718(11)	0.703(12)	0.692(13)	0.782(4)
	T1-CK	0.665(2)	0.605(10)	0.609(9)	0.594(12)	0.664(3)	0.619(6)	0.611(7)	0.609(8)	0.667(1)	0.583(13)	0.657(5)	0.604(11)	0.574(14)	0.661(4)
	T1-DCI	0.647(8)	0.571(14)	0.574(13)	0.587(12)	0.647(10)	0.652(7)	0.658(5)	0.683(3)	0.647(8)	0.637(11)	0.718(1)	0.678(4)	0.711(2)	0.656(6)
	T1-PCK	0.538(5)	0.495(12)	0.513(10)	0.484(13)	0.535(7)	0.584(3)	0.585(2)	0.624(1)	0.536(6)	0.483(14)	0.524(9)	0.529(8)	0.506(11)	0.544(4)
	T1-SEP	0.786(1)	0.691(7)	0.696(6)	0.707(5)	0.781(4)	0.679(9)	0.666(11)	0.673(10)	0.781(3)	0.620(12)	0.681(8)	0.617(13)	0.607(14)	0.785(2)
	T2-CK	0.619(3)	0.588(11)	0.583(12)	0.590(9)	0.620(2)	0.592(8)	0.599(7)	0.617(5)	0.619(4)	0.565(13)	0.608(6)	0.589(10)	0.517(14)	0.640(1)
	T2-DCI	0.530(10)	0.492(12)	0.481(13)	0.480(14)	0.531(9)	0.534(7)	0.549(6)	0.562(5)	0.531(8)	0.565(4)	0.651(1)	0.620(3)	0.629(2)	0.513(11)
	T2-PCK	0.558(12)	0.627(5)	0.633(3)	0.619(6)	0.560(10)	0.628(4)	0.652(2)	0.654(1)	0.559(11)	0.487(14)	0.593(7)	0.556(13)	0.583(8)	0.563(9)
	T2-SEP	0.452(5)	0.467(3)	0.476(1)	0.473(2)	0.453(4)	0.432(10)	0.441(7)	0.438(8)	0.451(6)	0.414(11)	0.397(14)	0.410(13)	0.412(12)	0.432(9)
	U41Q1	0.761(3)	0.678(12)	0.673(13)	0.686(10)	0.762(2)	0.693(9)	0.697(8)	0.681(11)	0.766(1)	0.644(14)	0.719(6)	0.715(7)	0.739(4)	0.735(5)
	U41Q2	0.715(5)	0.598(14)	0.611(13)	0.622(11)	0.715(2)	0.635(8)	0.629(9)	0.620(12)	0.715(3)	0.628(10)	0.695(6)	0.715(4)	0.730(1)	0.686(7)
	U42Q1	0.555(5)	0.488(14)	0.503(12)	0.546(8)	0.560(2)	0.561(1)	0.557(3)	0.547(7)	0.554(6)	0.494(13)	0.527(10)	0.515(11)	0.540(9)	0.557(4)
	U42Q2	0.640(4)	0.543(14)	0.563(13)	0.568(12)	0.647(2)	0.601(7)	0.600(8)	0.626(5)	0.651(1)	0.645(3)	0.607(6)	0.588(11)	0.592(10)	0.597(9)
shot=2, cot=1	ASAP	0.516(5)	0.454(13)	0.459(12)	0.435(14)	0.513(6)	0.476(11)	0.488(9)	0.496(8)	0.513(7)	0.486(10)	0.538(2)	0.547(1)	0.537(3)	0.522(4)
	Semeval	0.800(3)	0.726(10)	0.733(9)	0.739(8)	0.800(2)	0.743(7)	0.749(6)	0.767(5)	0.801(1)	0.645(14)	0.690(11)	0.685(12)	0.671(13)	0.781(4)
	T1-CK	0.872(2)	0.811(13)	0.812(12)	0.818(11)	0.872(3)	0.858(5)	0.866(4)	0.858(6)	0.873(1)	0.699(14)	0.840(7)	0.838(8)	0.833(9)	0.827(10)
	T1-DCI	0.693(4)	0.494(12)	0.489(13)	0.485(14)	0.694(2)	0.541(11)	0.553(10)	0.591(9)	0.693(3)	0.632(8)	0.650(7)	0.663(6)	0.714(1)	0.686(5)
	T1-PCK	0.586(8)	0.607(5)	0.606(6)	0.603(7)	0.579(10)	0.647(2)	0.645(3)	0.668(1)	0.584(9)	0.465(14)	0.567(11)	0.526(12)	0.522(13)	0.626(4)
	T1-SEP	0.768(2)	0.743(10)	0.746(8)	0.762(5)	0.768(3)	0.757(6)	0.755(7)	0.743(9)	0.767(4)	0.615(14)	0.726(11)	0.694(13)	0.703(12)	0.772(1)
	T2-CK	0.838(1)	0.793(12)	0.796(11)	0.814(8)	0.836(3)	0.805(9)	0.817(5)	0.830(4)	0.838(2)	0.736(14)	0.816(7)	0.757(13)	0.798(10)	0.816(6)
	T2-DCI	0.585(6)	0.457(13)	0.451(14)	0.463(12)	0.582(9)	0.583(7)	0.592(5)	0.607(4)	0.583(8)	0.521(11)	0.652(1)	0.636(2)	0.615(3)	0.582(10)
	T2-PCK	0.550(7)	0.571(4)	0.582(2)	0.576(3)	0.549(9)	0.545(13)	0.546(11)	0.596(1)	0.550(8)	0.554(6)	0.546(12)	0.548(10)	0.564(5)	0.538(14)
	T2-SEP	0.513(9)	0.535(4)	0.536(3)	0.510(10)	0.517(7)	0.556(1)	0.539(2)	0.532(5)	0.516(8)	0.527(6)	0.472(14)	0.506(11)	0.483(13)	0.505(12)
	U41Q1	0.805(2)	0.706(13)	0.713(12)	0.735(9)	0.804(3)	0.718(11)	0.728(10)	0.747(8)	0.808(1)	0.677(14)	0.770(6)	0.773(5)	0.790(4)	0.763(7)
	U41Q2	0.668(6)	0.683(3)	0.688(1)	0.688(2)	0.671(5)	0.626(11)	0.632(10)	0.632(9)	0.673(4)	0.543(14)	0.607(12)	0.602(13)	0.663(7)	0.635(8)
	U42Q1	0.632(4)	0.577(10)	0.578(9)	0.605(8)	0.612(6)	0.551(12)	0.531(13)	0.569(11)	0.617(5)	0.519(14)	0.607(7)	0.636(3)	0.646(2)	0.659(1)
	U42Q2	0.627(2)	0.423(14)	0.429(13)	0.431(12)	0.626(3)	0.459(10)	0.436(11)	0.502(8)	0.636(1)	0.472(9)	0.556(6)	0.567(5)	0.583(4)	0.529(7)

AUERC

Control	Question	CE	Embed_Eigen	Embed_GE	Embed_NAD	FSD	JS_Eigen	JS_GE	JS_NAD	MAR	NLI_DSE	NLI_Eigen	NLI_GE	NLI_NAD	Numset
shot=0, cot=0	ASAP	0.229(5)	0.249(11)	0.256(12)	0.290(14)	0.230(7)	0.237(9)	0.232(8)	0.247(10)	0.230(6)	0.272(13)	0.195(1)	0.201(2)	0.204(3)	0.226(4)
	Semeval	0.209(1)	0.243(6)	0.246(7)	0.242(5)	0.210(3)	0.250(9)	0.250(8)	0.251(10)	0.209(2)	0.307(11)	0.309(12)	0.318(13)	0.352(14)	0.219(4)
	T1-CK	0.194(4)	0.207(7)	0.215(8)	0.231(12)	0.196(6)	0.243(13)	0.253(14)	0.225(11)	0.195(5)	0.223(10)	0.190(1)	0.193(3)	0.219(9)	0.191(2)
	T1-DCI	0.318(3)	0.425(12)	0.435(14)	0.431(13)	0.318(2)	0.370(10)	0.362(9)	0.340(8)	0.318(3)	0.377(11)	0.324(6)	0.334(7)	0.292(1)	0.318(5)
	T1-PCK	0.210(6)	0.213(9)	0.218(10)	0.205(5)	0.212(8)	0.179(3)	0.174(2)	0.155(1)	0.211(7)	0.225(11)	0.234(13)	0.229(12)	0.238(14)	0.200(4)
	T1-SEP	0.164(5)	0.167(9)	0.163(4)	0.153(1)	0.166(8)	0.178(10)	0.156(2)	0.157(3)	0.165(6)	0.252(12)	0.242(11)	0.258(13)	0.283(14)	0.165(7)
	T2-CK	0.197(1)	0.213(10)	0.211(9)	0.207(7)	0.202(4)	0.204(5)	0.210(8)	0.200(2)	0.200(3)	0.232(11)	0.234(12)	0.243(13)	0.261(14)	0.206(6)
	T2-DCI	0.430(7)	0.484(12)	0.496(13)	0.520(14)	0.429(5)	0.457(11)	0.450(10)	0.438(9)	0.430(6)	0.353(3)	0.373(4)	0.319(1)	0.328(2)	0.435(8)
	T2-PCK	0.223(7)	0.210(4)	0.214(5)	0.219(6)	0.224(9)	0.183(3)	0.178(2)	0.169(1)	0.223(8)	0.260(14)	0.231(11)	0.224(10)	0.231(12)	0.233(13)
	T2-SEP	0.308(14)	0.280(4)	0.292(8)	0.303(10)	0.308(13)	0.285(6)	0.279(3)	0.293(9)	0.308(12)	0.281(5)	0.276(2)	0.285(7)	0.275(1)	0.307(11)
	U41Q1	0.066(4)	0.087(10)	0.085(9)	0.080(8)	0.065(1)	0.102(14)	0.099(13)	0.093(11)	0.065(2)	0.095(12)	0.068(5)	0.075(7)	0.065(3)	0.072(6)
	U41Q2	0.068(6)	0.067(3)	0.067(4)	0.070(9)	0.068(8)	0.073(10)	0.073(11)	0.081(14)	0.068(7)	0.076(13)	0.067(5)	0.063(2)	0.062(1)	0.074(12)
	U42Q1	0.115(2)	0.121(5)	0.125(6)	0.127(7)	0.116(3)	0.140(13)	0.140(12)	0.142(14)	0.117(4)	0.133(10)	0.132(9)	0.134(11)	0.131(8)	0.106(1)
	U42Q2	0.072(2)	0.093(9)	0.094(10)	0.101(13)	0.078(4)	0.097(11)	0.100(12)	0.103(14)	0.076(3)	0.091(8)	0.085(6)	0.081(5)	0.086(7)	0.071(1)
shot=0, cot=1	ASAP	0.224(2)	0.256(13)	0.255(12)	0.275(14)	0.225(4)	0.238(7)	0.240(9)	0.243(10)	0.224(3)	0.247(11)	0.214(1)	0.238(8)	0.230(6)	0.225(5)
	Semeval	0.197(1)	0.241(7)	0.242(8)	0.231(6)	0.199(3)	0.243(9)	0.251(10)	0.222(5)	0.198(2)	0.315(14)	0.272(11)	0.287(12)	0.298(13)	0.208(4)
	T1-CK	0.219(3)	0.244(8)	0.242(7)	0.260(12)	0.219(4)	0.237(6)	0.247(9)	0.257(11)	0.218(2)	0.275(14)	0.217(1)	0.252(10)	0.274(13)	0.220(5)
	T1-DCI	0.342(8)	0.418(14)	0.414(13)	0.401(12)	0.342(8)	0.337(7)	0.331(5)	0.307(3)	0.342(8)	0.351(11)	0.271(1)	0.311(4)	0.279(2)	0.334(6)
	T1-PCK	0.227(5)	0.250(12)	0.241(10)	0.257(13)	0.228(7)	0.204(3)	0.203(2)	0.183(1)	0.228(6)	0.258(14)	0.234(9)	0.231(8)	0.242(11)	0.225(4)
	T1-SEP	0.125(1)	0.178(7)	0.177(6)	0.165(5)	0.129(3)	0.203(10)	0.209(11)	0.198(8)	0.128(2)	0.277(14)	0.202(9)	0.256(13)	0.248(12)	0.130(4)
	T2-CK	0.207(6)	0.217(11)	0.219(12)	0.214(10)	0.207(4)	0.211(8)	0.208(7)	0.203(2)	0.207(5)	0.243(13)	0.206(3)	0.214(9)	0.250(14)	0.198(1)
	T2-DCI	0.404(10)	0.439(12)	0.451(14)	0.451(13)	0.403(8)	0.380(7)	0.367(6)	0.352(4)	0.403(9)	0.360(5)	0.296(1)	0.319(3)	0.303(2)	0.416(11)
	T2-PCK	0.229(12)	0.199(5)	0.195(4)	0.206(6)	0.228(9)	0.191(3)	0.178(2)	0.175(1)	0.229(10)	0.262(14)	0.210(7)	0.229(11)	0.211(8)	0.231(13)
	T2-SEP	0.270(5)	0.262(3)	0.258(1)	0.259(2)	0.269(4)	0.279(9)	0.275(7)	0.276(8)	0.270(6)	0.288(11)	0.297(14)	0.290(13)	0.289(12)	0.280(10)
	U41Q1	0.059(2)	0.080(9)	0.081(11)	0.078(8)	0.061(3)	0.082(12)	0.081(10)	0.088(13)	0.059(1)	0.106(14)	0.071(7)	0.071(6)	0.064(4)	0.066(5)
	U41Q2	0.071(5)	0.095(12)	0.092(9)	0.089(8)	0.071(3)	0.094(11)	0.095(13)	0.100(14)	0.071(4)	0.094(10)	0.075(6)	0.070(2)	0.066(1)	0.078(7)
	U42Q1	0.120(5)	0.141(14)	0.136(12)	0.122(8)	0.118(3)	0.117(2)	0.119(4)	0.122(7)	0.120(6)	0.125(9)	0.129(11)	0.139(13)	0.128(10)	0.114(1)
	U42Q2	0.089(5)	0.113(13)	0.113(12)	0.116(14)	0.086(4)	0.089(6)	0.091(7)	0.084(1)	0.085(3)	0.085(2)	0.091(8)	0.099(10)	0.096(9)	0.103(11)
shot=2, cot=1	ASAP	0.236(8)	0.247(13)	0.244(12)	0.256(14)	0.238(11)	0.237(9)	0.228(5)	0.224(4)	0.238(10)	0.233(7)	0.211(2)	0.207(1)	0.215(3)	0.229(6)
	Semeval	0.190(3)	0.264(10)	0.257(9)	0.251(8)	0.190(2)	0.247(7)	0.241(6)	0.223(5)	0.189(1)	0.345(14)	0.300(11)	0.305(12)	0.319(13)	0.209(4)
	T1-CK	0.062(3)	0.100(13)	0.098(12)	0.092(11)	0.062(2)	0.071(6)	0.066(4)	0.071(5)	0.061(1)	0.164(14)	0.081(7)	0.086(8)	0.089(10)	0.088(9)
	T1-DCI	0.297(4)	0.496(12)	0.501(13)	0.505(14)	0.296(2)	0.449(11)	0.437(10)	0.399(9)	0.297(3)	0.358(8)	0.340(7)	0.327(6)	0.276(1)	0.304(5)
	T1-PCK	0.203(8)	0.194(5)	0.195(6)	0.196(7)	0.206(10)	0.176(2)	0.176(3)	0.164(1)	0.203(9)	0.265(14)	0.214(11)	0.236(13)	0.235(12)	0.184(4)
	T1-SEP	0.131(2)	0.157(10)	0.156(9)	0.144(6)	0.131(1)	0.147(7)	0.144(5)	0.150(8)	0.132(3)	0.228(14)	0.170(11)	0.185(13)	0.178(12)	0.132(4)
	T2-CK	0.096(5)	0.106(11)	0.106(10)	0.096(4)	0.097(7)	0.102(8)	0.096(3)	0.089(1)	0.096(6)	0.132(14)	0.092(2)	0.121(13)	0.104(9)	0.108(12)
	T2-DCI	0.334(8)	0.418(12)	0.428(13)	0.435(14)	0.335(10)	0.309(6)	0.303(5)	0.292(4)	0.335(9)	0.343(11)	0.266(2)	0.268(3)	0.266(1)	0.330(7)
	T2-PCK	0.232(5)	0.237(12)	0.230(3)	0.232(4)	0.234(10)	0.233(8)	0.232(6)	0.205(1)	0.233(9)	0.247(13)	0.236(11)	0.233(7)	0.223(2)	0.249(14)
	T2-SEP	0.239(9)	0.229(4)	0.228(3)	0.241(10)	0.237(7)	0.217(1)	0.225(2)	0.229(5)	0.237(8)	0.232(6)	0.259(14)	0.242(11)	0.253(13)	0.244(12)
	U41Q1	0.049(2)	0.071(12)	0.068(10)	0.063(8)	0.051(3)	0.073(13)	0.070(11)	0.067(9)	0.049(1)	0.091(14)	0.058(6)	0.056(5)	0.052(4)	0.061(7)
	U41Q2	0.089(4)	0.081(3)	0.079(1)	0.080(2)	0.090(7)	0.098(10)	0.098(11)	0.097(9)	0.089(5)	0.125(14)	0.104(12)	0.106(13)	0.089(6)	0.096(8)
	U42Q1	0.095(4)	0.111(9)	0.112(10)	0.108(8)	0.100(7)	0.126(13)	0.133(14)	0.116(11)	0.099(6)	0.120(12)	0.095(5)	0.091(3)	0.089(2)	0.087(1)
	U42Q2	0.092(3)	0.146(14)	0.145(12)	0.145(13)	0.091(2)	0.132(10)	0.137(11)	0.121(8)	0.090(1)	0.128(9)	0.108(6)	0.105(5)	0.101(4)	0.117(7)

Table 4: Metric value and ranking results of each uncertainty quantification methods with Gemini-3-flash over 14 questions.

Control	Question	AUROC													
		CE	Embed_Eigen	Embed_GE	Embed_NAD	FSD	JS_Eigen	JS_GE	JS_NAD	MAR	NLI_DSE	NLI_Eigen	NLI_GE	NLI_NAD	Numset
shot=0, cot=0	ASAP	0.677(4)	0.609(13)	0.623(10)	0.648(7)	0.675(6)	0.615(11)	0.610(12)	0.695(1)	0.676(5)	0.549(14)	0.630(9)	0.636(8)	0.683(2)	0.677(3)
	Semeval	0.605(1)	0.588(7)	0.583(8)	0.597(5)	0.604(3)	0.572(9)	0.561(10)	0.593(6)	0.604(2)	0.443(12)	0.433(13)	0.484(11)	0.412(14)	0.599(4)
	T1-CK	0.617(2)	0.547(10)	0.552(8)	0.531(13)	0.615(4)	0.563(6)	0.551(9)	0.521(14)	0.616(3)	0.538(12)	0.557(7)	0.543(11)	0.582(5)	0.619(1)
	T1-DCI	0.544(9)	0.471(14)	0.479(13)	0.482(12)	0.544(11)	0.689(2)	0.698(1)	0.648(4)	0.544(10)	0.590(7)	0.624(6)	0.655(3)	0.634(5)	0.547(8)
	T1-PCK	0.676(5)	0.609(12)	0.606(13)	0.627(11)	0.677(3)	0.660(7)	0.664(6)	0.680(2)	0.676(4)	0.642(10)	0.657(8)	0.646(9)	0.599(14)	0.682(1)
	T1-SEP	0.615(5)	0.635(2)	0.623(3)	0.592(10)	0.614(7)	0.638(1)	0.594(9)	0.610(8)	0.614(6)	0.550(14)	0.592(11)	0.558(13)	0.579(12)	0.618(4)
	T2-CK	0.718(4)	0.660(11)	0.659(12)	0.661(10)	0.718(4)	0.747(3)	0.772(2)	0.838(1)	0.718(6)	0.637(14)	0.653(13)	0.662(9)	0.691(8)	0.693(7)
	T2-DCI	0.743(3)	0.630(13)	0.633(12)	0.626(14)	0.749(1)	0.720(7)	0.711(8)	0.682(11)	0.745(2)	0.690(10)	0.729(5)	0.726(6)	0.699(9)	0.737(4)
	T2-PCK	0.654(7)	0.487(12)	0.471(14)	0.479(13)	0.655(5)	0.687(2)	0.685(3)	0.689(1)	0.655(6)	0.538(11)	0.633(8)	0.593(9)	0.573(10)	0.660(4)
	T2-SEP	0.590(11)	0.582(14)	0.585(12)	0.595(8)	0.591(10)	0.752(3)	0.766(1)	0.755(2)	0.591(9)	0.605(7)	0.630(6)	0.672(4)	0.631(5)	0.583(13)
	U41Q1	0.692(6)	0.684(10)	0.690(7)	0.693(5)	0.689(9)	0.778(1)	0.767(2)	0.749(3)	0.690(7)	0.629(14)	0.680(12)	0.666(13)	0.705(4)	0.681(11)
	U41Q2	0.671(10)	0.700(5)	0.697(6)	0.661(13)	0.670(12)	0.687(7)	0.681(8)	0.598(14)	0.671(11)	0.735(3)	0.735(2)	0.750(1)	0.732(4)	0.672(9)
	U42Q1	0.675(3)	0.636(8)	0.638(7)	0.664(5)	0.679(1)	0.610(13)	0.616(12)	0.626(11)	0.678(2)	0.521(14)	0.634(9)	0.634(9)	0.670(4)	0.654(6)
	U42Q2	0.579(4)	0.510(11)	0.512(10)	0.543(8)	0.589(2)	0.576(5)	0.553(7)	0.645(1)	0.586(3)	0.469(14)	0.537(9)	0.485(12)	0.474(13)	0.575(6)
shot=0, cot=1	ASAP	0.682(3)	0.655(7)	0.654(8)	0.662(6)	0.683(1)	0.652(9)	0.652(10)	0.673(5)	0.683(1)	0.606(14)	0.646(11)	0.636(13)	0.637(12)	0.677(4)
	Semeval	0.602(3)	0.580(6)	0.574(9)	0.577(7)	0.603(2)	0.577(8)	0.567(10)	0.598(4)	0.604(1)	0.489(14)	0.529(12)	0.527(13)	0.537(11)	0.594(5)
	T1-CK	0.616(4)	0.563(10)	0.562(11)	0.556(12)	0.617(3)	0.604(6)	0.600(7)	0.625(1)	0.617(2)	0.492(14)	0.574(9)	0.527(13)	0.597(8)	0.607(5)
	T1-DCI	0.584(10)	0.622(9)	0.624(8)	0.626(7)	0.583(13)	0.650(5)	0.661(4)	0.680(1)	0.583(11)	0.539(14)	0.673(2)	0.665(3)	0.636(6)	0.583(12)
	T1-PCK	0.755(4)	0.638(11)	0.629(12)	0.607(13)	0.761(2)	0.659(9)	0.642(10)	0.671(8)	0.761(3)	0.576(14)	0.744(6)	0.766(1)	0.741(7)	0.747(5)
	T1-SEP	0.602(5)	0.494(9)	0.492(10)	0.487(11)	0.602(4)	0.452(12)	0.429(13)	0.369(14)	0.602(6)	0.534(8)	0.556(7)	0.605(3)	0.623(1)	0.607(2)
	T2-CK	0.709(4)	0.636(11)	0.642(10)	0.643(9)	0.717(1)	0.684(7)	0.685(6)	0.715(3)	0.716(2)	0.567(14)	0.620(13)	0.629(12)	0.653(8)	0.688(5)
	T2-DCI	0.767(5)	0.733(12)	0.746(10)	0.752(9)	0.774(3)	0.729(13)	0.745(11)	0.754(7)	0.773(4)	0.661(14)	0.756(6)	0.795(2)	0.802(1)	0.754(8)
	T2-PCK	0.725(5)	0.643(11)	0.626(13)	0.631(12)	0.717(7)	0.693(8)	0.685(10)	0.687(9)	0.720(6)	0.495(14)	0.768(3)	0.780(1)	0.773(2)	0.736(4)
	T2-SEP	0.692(6)	0.668(10)	0.661(11)	0.655(13)	0.693(5)	0.784(3)	0.788(2)	0.811(1)	0.694(4)	0.620(14)	0.684(7)	0.680(9)	0.660(12)	0.684(7)
	U41Q1	0.706(1)	0.663(10)	0.678(7)	0.673(9)	0.705(3)	0.696(4)	0.679(6)	0.677(8)	0.705(2)	0.552(14)	0.590(13)	0.621(11)	0.620(12)	0.694(5)
	U41Q2	0.711(5)	0.653(9)	0.623(12)	0.558(14)	0.709(7)	0.664(8)	0.630(10)	0.620(13)	0.710(6)	0.625(11)	0.761(3)	0.773(2)	0.794(1)	0.725(4)
	U42Q1	0.654(5)	0.693(2)	0.694(1)	0.644(8)	0.659(3)	0.639(10)	0.626(11)	0.619(12)	0.658(4)	0.512(14)	0.650(7)	0.642(9)	0.615(13)	0.651(6)
	U42Q2	0.631(5)	0.581(12)	0.585(11)	0.580(13)	0.634(2)	0.610(8)	0.603(9)	0.665(1)	0.633(3)	0.478(14)	0.593(10)	0.633(4)	0.628(6)	0.627(7)
shot=2, cot=1	ASAP	0.686(3)	0.681(6)	0.682(5)	0.655(10)	0.680(7)	0.625(14)	0.627(13)	0.629(12)	0.683(4)	0.642(11)	0.705(1)	0.662(9)	0.664(8)	0.689(2)
	Semeval	0.733(2)	0.607(12)	0.597(13)	0.631(10)	0.732(3)	0.648(7)	0.647(9)	0.679(5)	0.733(1)	0.551(14)	0.648(7)	0.664(6)	0.619(11)	0.719(4)
	T1-CK	0.820(5)	0.788(9)	0.764(12)	0.718(14)	0.819(6)	0.820(4)	0.788(9)	0.801(8)	0.821(3)	0.746(13)	0.833(1)	0.831(2)	0.785(11)	0.817(7)
	T1-DCI	0.593(9)	0.539(14)	0.541(13)	0.556(12)	0.592(10)	0.605(5)	0.621(2)	0.651(1)	0.594(8)	0.571(11)	0.606(4)	0.599(6)	0.614(3)	0.596(7)
	T1-PCK	0.802(1)	0.716(12)	0.715(13)	0.725(11)	0.793(4)	0.762(8)	0.771(6)	0.771(6)	0.800(2)	0.639(14)	0.760(9)	0.736(10)	0.772(5)	0.799(3)
	T1-SEP	0.686(4)	0.579(11)	0.583(10)	0.585(9)	0.688(2)	0.551(14)	0.563(13)	0.574(12)	0.688(1)	0.659(7)	0.665(6)	0.649(8)	0.671(5)	0.686(3)
	T2-CK	0.727(5)	0.696(10)	0.714(9)	0.738(4)	0.723(8)	0.792(3)	0.807(2)	0.832(1)	0.726(6)	0.633(11)	0.581(13)	0.527(14)	0.596(12)	0.723(7)
	T2-DCI	0.811(3)	0.692(11)	0.689(12)	0.669(14)	0.802(7)	0.761(10)	0.783(9)	0.825(1)	0.809(5)	0.681(13)	0.787(8)	0.810(4)	0.818(2)	0.809(6)
	T2-PCK	0.706(3)	0.641(8)	0.617(9)	0.609(12)	0.716(1)	0.593(13)	0.566(14)	0.611(11)	0.714(2)	0.613(10)	0.697(6)	0.701(5)	0.681(7)	0.705(4)
	T2-SEP	0.748(6)	0.654(12)	0.649(13)	0.624(14)	0.754(3)	0.753(4)	0.756(2)	0.783(1)	0.752(5)	0.696(11)	0.714(9)	0.740(8)	0.708(10)	0.743(7)
	U41Q1	0.654(5)	0.647(8)	0.659(4)	0.645(9)	0.647(7)	0.673(2)	0.665(3)	0.742(1)	0.651(6)	0.624(11)	0.612(13)	0.620(12)	0.609(14)	0.644(10)
	U41Q2	0.646(4)	0.494(14)	0.496(13)	0.537(10)	0.643(6)	0.534(12)	0.536(11)	0.543(9)	0.645(5)	0.617(8)	0.650(3)	0.652(2)	0.706(1)	0.634(7)
	U42Q1	0.702(5)	0.673(9)	0.673(8)	0.649(11)	0.701(6)	0.647(12)	0.641(13)	0.662(10)	0.703(4)	0.620(14)	0.707(2)	0.704(3)	0.687(7)	0.723(1)
	U42Q2	0.634(7)	0.585(12)	0.581(13)	0.631(8)	0.636(5)	0.622(9)	0.609(11)	0.696(1)	0.636(6)	0.539(14)	0.691(2)	0.657(3)	0.648(4)	0.620(10)

Control	Question	C-index													
		CE	Embed_Eigen	Embed_GE	Embed_NAD	FSD	JS_Eigen	JS_GE	JS_NAD	MAR	NLI_DSE	NLI_Eigen	NLI_GE	NLI_NAD	Numset
shot=0, cot=0	ASAP	0.653(3)	0.607(11)	0.620(8)	0.633(7)	0.649(6)	0.587(12)	0.585(13)	0.652(4)	0.650(5)	0.542(14)	0.611(10)	0.619(9)	0.655(1)	0.654(2)
	Semeval	0.605(1)	0.588(7)	0.583(8)	0.597(5)	0.604(3)	0.572(9)	0.561(10)	0.593(6)	0.604(2)	0.443(12)	0.433(13)	0.484(11)	0.412(14)	0.599(4)
	T1-CK	0.613(2)	0.558(10)	0.563(8)	0.547(12)	0.608(4)	0.570(7)	0.560(9)	0.531(14)	0.609(3)	0.544(13)	0.572(6)	0.557(11)	0.596(5)	0.614(1)
	T1-DCI	0.544(9)	0.471(14)	0.479(13)	0.482(12)	0.544(11)	0.689(2)	0.698(1)	0.648(4)	0.544(10)	0.590(7)	0.624(6)	0.655(3)	0.634(5)	0.547(8)
	T1-PCK	0.671(4)	0.603(12)	0.601(13)	0.624(11)	0.672(2)	0.652(8)	0.652(7)	0.668(5)	0.672(3)	0.637(10)	0.656(6)	0.640(9)	0.596(14)	0.676(1)
	T1-SEP	0.595(5)	0.613(2)	0.602(3)	0.576(9)	0.594(7)	0.616(1)	0.571(11)	0.590(8)	0.594(6)	0.527(14)	0.573(10)	0.535(13)	0.556(12)	0.599(4)
	T2-CK	0.715(6)	0.661(9)	0.660(11)	0.660(10)	0.716(4)	0.747(3)	0.771(2)	0.835(1)	0.716(5)	0.638(14)	0.651(13)	0.659(12)	0.689(8)	0.690(7)
	T2-DCI	0.699(5)	0.607(12)	0.606(13)	0.587(14)	0.696(6)	0.699(4)	0.688(8)	0.645(11)	0.694(7)	0.673(10)	0.721(1)	0.716(2)	0.677(9)	0.707(3)
	T2-PCK	0.654(7)	0.487(12)	0.471(14)	0.479(13)	0.655(5)	0.687(2)	0.685(3)	0.689(1)	0.655(6)	0.538(11)	0.633(8)	0.593(9)	0.573(10)	0.660(4)
	T2-SEP	0.551(11)	0.544(14)	0.549(12)	0.560(8)	0.551(10)	0.725(3)	0.743(1)	0.742(2)	0.551(9)	0.580(7)	0.611(5)	0.647(4)	0.598(6)	0.545(13)
	U41Q1	0.691(8)	0.692(7)	0.698(6)	0.702(4)	0.686(10)	0.777(1)	0.766(2)	0.752(3)	0.687(9)	0.623(14)	0.672(12)	0.661(13)	0.701(5)	0.679(11)
	U41Q2	0.668(12)	0.701(5)	0.698(6)	0.660(13)	0.668(10)	0.690(7)	0.681(8)	0.596(14)	0.668(10)	0.734(3)	0.738(2)	0.753(1)	0.729(4)	0.670(9)
	U42Q1	0.644(3)	0.598(11)	0.601(9)	0.629(5)	0.653(1)	0.591(13)	0.596(12)	0.601(10)	0.648(2)	0.508(14)	0.616(7)	0.604(8)	0.637(4)	0.620(6)
	U42Q2	0.579(4)	0.510(11)	0.512(10)	0.543(8)	0.589(2)	0.576(5)	0.553(7)	0.645(1)	0.586(3)	0.469(14)	0.537(9)	0.485(12)	0.474(13)	0.575(6)
shot=0, cot=1	ASAP	0.664(1)	0.648(7)	0.646(8)	0.648(6)	0.662(3)	0.641(9)	0.639(10)	0.649(5)	0.663(2)	0.617(14)	0.631(12)	0.634(11)	0.630(13)	0.659(4)
	Semeval	0.602(3)	0.580(6)	0.574(9)	0.577(7)	0.603(2)	0.577(8)	0.567(10)	0.598(4)	0.604(1)	0.489(14)	0.529(12)	0.527(13)	0.537(11)	0.594(5)
	T1-CK	0.566(6)	0.537(10)	0.535(11)	0.522(12)	0.567(4)	0.589(2)	0.585(3)	0.596(1)	0.567(5)	0.492(14)	0.552(9)	0.508(13)	0.562(7)	0.560(8)
	T1-DCI	0.567(11)	0.609(9)	0.611(7)	0.610(8)	0.567(12)	0.638(5)	0.647(4)	0.665(1)	0.568(10)	0.524(14)	0.658(2)	0.650(3)	0.621(6)	0.567(13)
	T1-PCK	0.708(2)	0.585(11)	0.573(12)	0.553(14)	0.706(3)	0.613(9)	0.593(10)	0.623(8)	0.706(4)	0.563(13)	0.698(6)	0.729(1)	0.696(7)	0.704(5)
	T1-SEP	0.599(4)	0.494(9)	0.494(10)	0.483(11)	0.598(5)	0.452(12)	0.431(13)	0.373(14)	0.598(6)	0.535(8)	0.557(7)	0.600(3)	0.610(1)	0.603(2)
	T2-CK	0.706(4)	0.637(11)	0.643(10)	0.644(9)	0.713(2)	0.687(5)	0.686(7)	0.715(1)	0.713(3)	0.565(14)	0.618(13)	0.627(12)	0.653(8)	0.687(6)
	T2-DCI	0.666(11)	0.663(12)	0.672(9)	0.661(13)	0.675(5)	0.673(7)	0.680(4)	0.673(6)	0.671(10)	0.612(14)	0.692(3)	0.715(1)	0.709(2)	0.672(8)
	T2-PCK	0.692(5)	0.612(11)	0.592(13)	0.597(12)	0.686(7)	0.667(8)	0.655(10)	0.657(9)	0.687(6)	0.496(14)	0.742(3)	0.756(1)	0.743(2)	0.705(4)
	T2-SEP	0.676(7)	0.661(10)	0.653(11)	0.646(13)	0.678(6)	0.778(3)	0.781(2)	0.804(1)	0.679(4)	0.630(14)	0.679(5)	0.670(8)	0.650(12)	0.669(9)
	U41Q1	0.712(1)	0.668(10)	0.684(6)	0.680(9)	0.706(3)	0.699(5)	0.683(7)	0.680(8)	0.707(2)	0.549(14)	0.593(13)	0.623(11)	0.622(12)	0.700(4)
	U41Q2	0.676(5)	0.639(9)	0.608(11)	0.535(14)	0.670(7)	0.645(8)	0.608(11)	0.585(13)	0.672(6)	0.613(10)	0.746(3)	0.754(2)	0.765(1)	0.692(4)
	U42Q1	0.639(6)	0.687(2)	0.688(1)	0.635(7)	0.643(4)	0.631(9)	0.620(11)	0.609(12)	0.642(5)	0.517(14)	0.635(7)	0.628(10)	0.604(13)	0.644(3)
	U42Q2	0.635(5)	0.586(12)	0.589(11)	0.585(13)	0.637(3)	0.614(8)	0.607(9)	0.668(1)	0.637(3)	0.478(14)	0.597(10)	0.637(2)	0.632(6)	0.631(7)
shot=2, cot=1	ASAP	0.666(5)	0.675(3)	0.678(2)	0.654(8)	0.661(7)	0.614(13)	0.619(12)	0.611(14)	0.665(6)	0.630(11)	0.681(1)	0.641(10)	0.646(9)	0.669(4)
	Semeval	0.701(2)	0.572(12)	0.561(13)	0.594(10)	0.702(1)	0.613(9)	0.615(8)	0.654(5)	0.701(3)	0.539(14)	0.620(7)	0.633(6)	0.585(11)	0.691(4)
	T1-CK	0.776(5)	0.743(10)	0.715(12)	0.663(14)	0.775(6)	0.783(3)	0.747(9)	0.755(8)	0.777(4)	0.700(13)	0.795(1)	0.791(2)	0.739(11)	0.774(7)
	T1-DCI	0.576(9)	0.525(14)	0.525(13)	0.539(12)	0.576(10)	0.589(5)	0.606(2)	0.635(1)	0.578(8)	0.560(11)	0.597(4)	0.585(6)	0.600(3)	0.579(7)
	T1-PCK	0.782(1)	0.694(12)	0.693(13)	0.704(11)	0.775(4)	0.744(9)	0.751(6)	0.750(7)	0.781(2)	0.623(14)	0.745(8)	0.720(10)	0.755(5)	0.779(3)
	T1-SEP	0.638(3)	0.539(11)	0.545(9)	0.543(10)	0.637(4)	0.509(14)	0.519(13)	0.527(12)	0.639(2)	0.610(5)	0.610(6)	0.597(8)	0.608(7)	0.643(1)
	T2-CK	0.623(5)	0.597(10)	0.612(9)	0.641(4)	0.621(8)	0.718(3)	0.733(2)	0.759(1)	0.621(6)	0.539(11)	0.459(13)	0.394(14)	0.467(12)	0.621(6)
	T2-DCI	0.752(3)	0.658(11)	0.653(12)	0.636(13)	0.748(4)	0.714(10)	0.728(8)	0.759(1)	0.752(2)	0.624(14)	0.721(9)	0.742(6)	0.747(5)	0.742(7)
	T2-PCK	0.686(4)	0.640(8)	0.615(9)	0.606(10)	0.695(1)	0.587(13)	0.554(14)	0.598(11)	0.694(2)	0.592(12)	0.679(6)	0.683(5)	0.662(7)	0.688(3)
	T2-SEP	0.734(5)	0.635(12)	0.631(13)	0.610(14)	0.740(2)	0.730(6)	0.735(4)	0.762(1)	0.738(3)	0.706(10)	0.707(9)	0.730(7)	0.699(11)	0.729(8)
	U41Q1	0.637(5)	0.631(8)	0.644(4)	0.630(9)	0.628(10)	0.657(2)	0.648(3)	0.727(1)	0.633(7)	0.633(6)	0.600(13)	0.612(12)	0.596(14)	0.628(11)
	U41Q2	0.620(4)	0.472(14)	0.473(13)	0.512(9)	0.615(6)	0.509(12)	0.511(11)	0.512(9)	0.617(5)	0.598(8)	0.631(2)	0.630(3)	0.684(1)	0.608(7)
	U42Q1	0.706(4)	0.673(8)	0.673(9)	0.650(11)	0.706(3)	0.643(12)	0.639(13)	0.659(10)	0.707(2)	0.623(14)	0.704(5)	0.703(6)	0.687(7)	0.721(1)
	U42Q2	0.636(7)	0.587(12)	0.584(13)	0.634(8)	0.637(5)	0.622(10)	0.610(11)	0.695(1)	0.637(6)	0.538(14)	0.694(2)	0.660(3)	0.651(4)	0.624(9)

Control	Question	AUARC													
		CE	Embed_Eigen	Embed_GE	Embed_NAD	FSD	JS_Eigen	JS_GE	JS_NAD	MAR	NLI_DSE	NLI_Eigen	NLI_GE	NLI_NAD	Numset
shot=0, cot=0	ASAP	0.654(4)	0.600(13)	0.618(12)	0.659(3)	0.653(6)	0.627(11)	0.629(10)	0.687(1)	0.653(5)	0.518(14)	0.636(8)	0.634(9)	0.669(2)	0.652(7)
	Semeval	0.777(1)	0.739(7)	0.735(8)	0.741(6)	0.777(2)	0.735(9)	0.728(11)	0.746(5)	0.777(3)	0.730(10)	0.681(14)	0.724(12)	0.684(13)	0.776(4)
	T1-CK	0.639(2)	0.625(9)	0.628(7)	0.620(10)	0.638(4)	0.595(13)	0.600(12)	0.576(14)	0.638(3)	0.634(6)	0.619(11)	0.626(8)	0.635(5)	0.639(1)
	T1-DCI	0.704(7)	0.573(14)	0.576(13)	0.595(12)	0.704(9)	0.739(2)	0.743(1)	0.723(3)	0.704(8)	0.695(11)	0.705(6)	0.714(4)	0.699(10)	0.705(5)
	T1-PCK	0.803(5)	0.779(13)	0.780(12)	0.791(8)	0.803(4)	0.836(3)	0.840(2)	0.851(1)	0.803(7)	0.786(11)	0.787(10)	0.791(9)	0.778(14)	0.803(6)
	T1-SEP	0.715(2)	0.702(7)	0.686(10)	0.691(9)	0.715(3)	0.718(1)	0.700(8)	0.709(6)	0.715(3)	0.665(11)	0.652(13)	0.645(14)	0.657(12)	0.712(5)
	T2-CK	0.867(10)	0.884(4)	0.884(5)	0.884(6)	0.867(8)	0.908(3)	0.914(2)	0.934(1)	0.867(8)	0.840(14)	0.866(11)	0.858(13)	0.873(7)	0.863(12)
	T2-DCI	0.700(2)	0.600(12)	0.599(13)	0.585(14)	0.702(1)	0.680(6)	0.669(7)	0.651(11)	0.700(3)	0.652(10)	0.663(8)	0.686(5)	0.652(9)	0.696(4)
	T2-PCK	0.816(6)	0.705(12)	0.692(14)	0.704(13)	0.816(4)	0.847(2)	0.844(3)	0.847(1)	0.816(5)	0.792(11)	0.809(9)	0.813(8)	0.805(10)	0.815(7)
	T2-SEP	0.732(13)	0.799(8)	0.796(9)	0.805(7)	0.732(12)	0.893(3)	0.899(1)	0.896(2)	0.732(11)	0.746(10)	0.846(4)	0.841(5)	0.831(6)	0.729(14)
	U41Q1	0.771(10)	0.806(8)	0.812(6)	0.818(5)	0.770(12)	0.878(1)	0.873(2)	0.857(3)	0.771(11)	0.737(14)	0.807(7)	0.805(9)	0.832(4)	0.767(13)
	U41Q2	0.761(10)	0.793(6)	0.796(5)	0.781(7)	0.761(9)	0.768(8)	0.760(12)	0.723(14)	0.761(11)	0.823(2)	0.821(3)	0.828(1)	0.808(4)	0.759(13)
	U42Q1	0.815(12)	0.842(4)	0.840(5)	0.846(2)	0.816(11)	0.824(8)	0.824(9)	0.834(6)	0.816(10)	0.795(14)	0.844(3)	0.832(7)	0.858(1)	0.813(13)
	U42Q2	0.795(7)	0.755(13)	0.766(12)	0.782(9)	0.797(4)	0.806(2)	0.802(3)	0.840(1)	0.795(6)	0.748(14)	0.793(8)	0.781(11)	0.782(10)	0.796(5)
shot=0, cot=1	ASAP	0.666(3)	0.604(6)	0.597(9)	0.604(7)	0.666(1)	0.589(13)	0.595(10)	0.627(5)	0.666(2)	0.594(11)	0.601(8)	0.590(12)	0.582(14)	0.664(4)
	Semeval	0.802(3)	0.756(6)	0.750(7)	0.732(8)	0.802(2)	0.725(9)	0.718(12)	0.764(5)	0.802(1)	0.724(10)	0.712(14)	0.719(11)	0.714(13)	0.801(4)
	T1-CK	0.624(8)	0.614(10)	0.610(11)	0.605(13)	0.625(6)	0.640(4)	0.648(3)	0.650(2)	0.625(7)	0.578(14)	0.636(5)	0.609(12)	0.654(1)	0.622(9)
	T1-DCI	0.685(8)	0.663(12)	0.658(13)	0.655(14)	0.685(9)	0.699(5)	0.693(6)	0.717(2)	0.685(9)	0.665(11)	0.731(1)	0.711(3)	0.709(4)	0.685(7)
	T1-PCK	0.848(6)	0.781(11)	0.771(14)	0.772(13)	0.849(5)	0.789(9)	0.781(10)	0.803(8)	0.850(4)	0.773(12)	0.865(2)	0.878(1)	0.861(3)	0.845(7)
	T1-SEP	0.743(1)	0.686(8)	0.685(9)	0.682(10)	0.743(2)	0.643(12)	0.632(13)	0.613(14)	0.743(3)	0.655(11)	0.705(7)	0.716(6)	0.731(5)	0.742(4)
	T2-CK	0.876(4)	0.842(12)	0.841(13)	0.859(7)	0.877(2)	0.867(6)	0.857(8)	0.878(1)	0.877(3)	0.811(14)	0.843(11)	0.849(10)	0.857(9)	0.871(5)
	T2-DCI	0.669(6)	0.648(11)	0.658(10)	0.669(7)	0.671(5)	0.632(13)	0.644(12)	0.661(9)	0.671(4)	0.625(14)	0.673(3)	0.707(1)	0.702(2)	0.667(8)
	T2-PCK	0.829(8)	0.798(11)	0.791(12)	0.779(13)	0.828(10)	0.841(4)	0.831(7)	0.832(5)	0.828(9)	0.726(14)	0.883(3)	0.887(1)	0.883(2)	0.831(6)
	T2-SEP	0.757(12)	0.836(4)	0.824(5)	0.816(7)	0.757(11)	0.887(3)	0.890(2)	0.899(1)	0.758(10)	0.713(14)	0.821(6)	0.800(9)	0.801(8)	0.755(13)
	U41Q1	0.805(8)	0.814(6)	0.830(5)	0.838(4)	0.804(9)	0.850(1)	0.838(3)	0.842(2)	0.805(7)	0.729(14)	0.771(13)	0.788(11)	0.784(12)	0.801(10)
	U41Q2	0.816(5)	0.750(9)	0.739(11)	0.715(14)	0.815(7)	0.755(8)	0.736(12)	0.740(10)	0.815(6)	0.725(13)	0.850(3)	0.862(2)	0.872(1)	0.817(4)
	U42Q1	0.754(13)	0.818(2)	0.823(1)	0.805(3)	0.755(11)	0.796(4)	0.795(5)	0.786(7)	0.755(12)	0.709(14)	0.783(8)	0.789(6)	0.760(9)	0.755(10)
	U42Q2	0.760(13)	0.762(10)	0.771(4)	0.769(6)	0.761(11)	0.765(8)	0.770(5)	0.814(1)	0.761(11)	0.683(14)	0.767(7)	0.797(2)	0.781(3)	0.762(9)
shot=2, cot=1	ASAP	0.727(3)	0.662(10)	0.665(9)	0.640(12)	0.726(5)	0.633(13)	0.629(14)	0.642(11)	0.727(4)	0.690(7)	0.741(1)	0.699(6)	0.678(8)	0.729(2)
	Semeval	0.837(2)	0.815(9)	0.809(12)	0.814(11)	0.837(4)	0.820(7)	0.816(8)	0.846(1)	0.837(3)	0.738(14)	0.815(10)	0.826(6)	0.805(13)	0.835(5)
	T1-CK	0.910(9)	0.920(7)	0.915(8)	0.898(13)	0.910(11)	0.941(3)	0.932(5)	0.931(6)	0.910(9)	0.885(14)	0.945(2)	0.947(1)	0.940(4)	0.910(12)
	T1-DCI	0.720(6)	0.679(12)	0.672(13)	0.659(14)	0.720(8)	0.708(10)	0.713(9)	0.734(3)	0.720(5)	0.688(11)	0.736(1)	0.734(4)	0.734(2)	0.720(7)
	T1-PCK	0.820(12)	0.825(8)	0.826(7)	0.833(6)	0.818(13)	0.841(3)	0.842(2)	0.840(5)	0.820(11)	0.770(14)	0.843(1)	0.821(9)	0.841(4)	0.820(10)
	T1-SEP	0.758(3)	0.628(14)	0.631(13)	0.643(11)	0.758(2)	0.642(12)	0.643(10)	0.654(9)	0.758(1)	0.752(5)	0.708(8)	0.713(7)	0.735(6)	0.757(4)
	T2-CK	0.915(7)	0.934(6)	0.938(5)	0.943(4)	0.914(9)	0.954(3)	0.957(2)	0.962(1)	0.915(8)	0.902(11)	0.875(13)	0.851(14)	0.883(12)	0.914(10)
	T2-DCI	0.809(8)	0.764(11)	0.764(12)	0.748(13)	0.808(9)	0.803(10)	0.816(5)	0.840(1)	0.810(7)	0.719(14)	0.826(4)	0.840(2)	0.829(3)	0.810(6)
	T2-PCK	0.884(4)	0.851(8)	0.830(13)	0.835(12)	0.885(1)	0.843(10)	0.824(14)	0.849(9)	0.885(2)	0.843(11)	0.880(5)	0.876(6)	0.857(7)	0.884(3)
	T2-SEP	0.788(10)	0.786(11)	0.778(12)	0.765(14)	0.789(7)	0.853(2)	0.848(3)	0.860(1)	0.789(8)	0.772(13)	0.819(6)	0.826(4)	0.824(5)	0.788(9)
	U41Q1	0.793(11)	0.815(6)	0.825(5)	0.833(4)	0.792(13)	0.851(2)	0.851(3)	0.881(1)	0.792(12)	0.794(10)	0.802(9)	0.807(7)	0.804(8)	0.791(14)
	U41Q2	0.734(4)	0.644(14)	0.646(13)	0.656(11)	0.734(6)	0.653(12)	0.671(10)	0.675(9)	0.734(5)	0.711(8)	0.757(2)	0.750(3)	0.794(1)	0.731(7)
	U42Q1	0.838(4)	0.815(7)	0.812(9)	0.807(10)	0.839(2)	0.814(8)	0.799(12)	0.804(11)	0.838(3)	0.784(14)	0.818(6)	0.823(5)	0.787(13)	0.843(1)
	U42Q2	0.763(10)	0.781(6)	0.776(7)	0.799(4)	0.763(8)	0.758(13)	0.758(12)	0.807(1)	0.763(9)	0.718(14)	0.803(2)	0.802(3)	0.786(5)	0.760(11)

Control	Question	AUERC													
		CE	Embed_Eigen	Embed_GE	Embed_NAD	FSD	JS_Eigen	JS_GE	JS_NAD	MAR	NLI_DSE	NLI_Eigen	NLI_GE	NLI_NAD	Numset
shot=0, cot=0	ASAP	0.151(4)	0.159(12)	0.152(7)	0.137(2)	0.152(8)	0.162(13)	0.159(11)	0.135(1)	0.152(6)	0.204(14)	0.155(10)	0.155(9)	0.141(3)	0.151(5)
	Semeval	0.213(1)	0.251(7)	0.255(8)	0.249(6)	0.213(2)	0.255(9)	0.262(11)	0.244(5)	0.213(3)	0.260(10)	0.309(14)	0.266(12)	0.306(13)	0.214(4)
	T1-CK	0.214(8)	0.200(5)	0.198(4)	0.201(7)	0.215(11)	0.220(13)	0.218(12)	0.232(14)	0.215(10)	0.201(6)	0.198(3)	0.196(2)	0.187(1)	0.214(9)
	T1-DCI	0.286(7)	0.417(14)	0.414(13)	0.395(12)	0.286(9)	0.251(2)	0.247(1)	0.267(3)	0.286(8)	0.295(11)	0.285(6)	0.276(4)	0.291(10)	0.285(5)
	T1-PCK	0.099(5)	0.116(14)	0.115(11)	0.106(8)	0.099(4)	0.086(2)	0.086(3)	0.079(1)	0.099(6)	0.116(13)	0.110(10)	0.109(9)	0.115(12)	0.099(7)
	T1-SEP	0.175(3)	0.180(7)	0.188(10)	0.181(8)	0.175(4)	0.168(1)	0.184(9)	0.172(2)	0.175(4)	0.210(11)	0.223(14)	0.221(13)	0.215(12)	0.175(6)
	T2-CK	0.079(12)	0.058(6)	0.058(5)	0.057(4)	0.079(10)	0.043(3)	0.040(2)	0.029(1)	0.079(11)	0.093(14)	0.069(8)	0.072(9)	0.066(7)	0.082(13)
	T2-DCI	0.216(2)	0.305(12)	0.308(13)	0.329(14)	0.217(3)	0.233(6)	0.244(8)	0.270(11)	0.218(4)	0.245(9)	0.237(7)	0.222(5)	0.252(10)	0.216(1)
	T2-PCK	0.087(6)	0.142(12)	0.149(14)	0.143(13)	0.087(4)	0.071(2)	0.073(3)	0.071(1)	0.087(5)	0.099(11)	0.090(9)	0.089(8)	0.092(10)	0.088(7)
	T2-SEP	0.129(11)	0.096(8)	0.097(9)	0.092(7)	0.130(13)	0.048(3)	0.045(1)	0.045(2)	0.129(12)	0.121(10)	0.070(4)	0.073(5)	0.079(6)	0.130(14)
	U41Q1	0.058(10)	0.046(7)	0.045(6)	0.043(5)	0.058(12)	0.029(1)	0.030(2)	0.034(3)	0.058(11)	0.068(14)	0.048(8)	0.048(9)	0.041(4)	0.059(13)
	U41Q2	0.059(12)	0.050(6)	0.049(5)	0.053(7)	0.058(10)	0.056(8)	0.058(9)	0.069(14)	0.058(11)	0.042(2)	0.042(3)	0.041(1)	0.047(4)	0.059(13)
	U42Q1	0.044(12)	0.040(5)	0.040(6)	0.038(3)	0.044(11)	0.043(9)	0.043(8)	0.040(4)	0.044(10)	0.050(14)	0.038(2)	0.041(7)	0.034(1)	0.046(13)
	U42Q2	0.049(7)	0.059(13)	0.056(12)	0.052(9)	0.048(4)	0.046(2)	0.047(3)	0.038(1)	0.049(6)	0.060(14)	0.049(8)	0.052(11)	0.052(10)	0.049(5)
shot=0, cot=1	ASAP	0.146(1)	0.154(6)	0.158(8)	0.157(7)	0.146(3)	0.165(12)	0.165(11)	0.153(5)	0.146(2)	0.163(9)	0.169(13)	0.165(10)	0.172(14)	0.146(4)
	Semeval	0.188(3)	0.234(6)	0.240(7)	0.258(8)	0.188(2)	0.265(9)	0.272(12)	0.226(5)	0.188(1)	0.266(10)	0.278(14)	0.271(11)	0.276(13)	0.189(4)
	T1-CK	0.226(12)	0.208(6)	0.211(7)	0.216(9)	0.225(10)	0.187(3)	0.182(2)	0.181(1)	0.226(11)	0.230(14)	0.200(5)	0.215(8)	0.191(4)	0.226(13)
	T1-DCI	0.305(10)	0.326(12)	0.332(13)	0.335(14)	0.305(8)	0.290(5)	0.297(6)	0.273(2)	0.305(9)	0.325(11)	0.259(1)	0.278(3)	0.281(4)	0.305(7)
	T1-PCK	0.087(4)	0.124(12)	0.127(14)	0.124(11)	0.088(6)	0.119(10)	0.125(13)	0.106(8)	0.088(5)	0.118(9)	0.077(2)	0.064(1)	0.078(3)	0.089(7)
	T1-SEP	0.151(1)	0.194(8)	0.196(9)	0.203(10)	0.151(2)	0.219(12)	0.227(13)	0.239(14)	0.151(3)	0.206(11)	0.169(7)	0.163(6)	0.155(5)	0.151(4)
	T2-CK	0.075(8)	0.078(11)	0.079(12)	0.071(5)	0.074(6)	0.063(2)	0.068(3)	0.058(1)	0.075(7)	0.110(14)	0.080(13)	0.078(10)	0.071(4)	0.077(9)
	T2-DCI	0.248(8)	0.252(12)	0.245(5)	0.248(9)	0.245(6)	0.256(13)	0.250(11)	0.250(10)	0.246(7)	0.267(14)	0.232(3)	0.211(1)	0.219(2)	0.244(4)
	T2-PCK	0.080(8)	0.095(11)	0.099(12)	0.105(13)	0.081(10)	0.073(4)	0.079(6)	0.079(5)	0.081(9)	0.126(14)	0.053(3)	0.051(1)	0.053(2)	0.079(7)
	T2-SEP	0.117(12)	0.077(4)	0.083(5)	0.087(7)	0.117(11)	0.051(3)	0.050(2)	0.045(1)	0.117(10)	0.127(14)	0.084(6)	0.095(8)	0.095(9)	0.118(13)
	U41Q1	0.046(7)	0.044(6)	0.040(5)	0.038(3)	0.047(9)	0.035(1)	0.038(4)	0.038(2)	0.047(8)	0.069(14)	0.055(13)	0.051(11)	0.052(12)	0.047(10)
	U41Q2	0.049(5)	0.065(9)	0.068(12)	0.073(14)	0.050(7)	0.060(8)	0.067(10)	0.068(11)	0.050(6)	0.068(13)	0.035(3)	0.033(2)	0.031(1)	0.049(4)
	U42Q1	0.061(12)	0.044(2)	0.043(1)	0.048(3)	0.061(11)	0.051(4)	0.051(5)	0.054(7)	0.061(10)	0.073(14)	0.055(8)	0.053(6)	0.061(13)	0.060(9)
	U42Q2	0.058(13)	0.057(10)	0.055(4)	0.055(6)	0.057(11)	0.056(8)	0.055(5)	0.044(1)	0.057(12)	0.080(14)	0.056(7)	0.048(2)	0.052(3)	0.057(9)
shot=2, cot=1	ASAP	0.112(2)	0.125(8)	0.123(6)	0.133(11)	0.112(5)	0.147(14)	0.145(12)	0.146(13)	0.112(3)	0.131(10)	0.104(1)	0.124(7)	0.130(9)	0.112(4)
	Semeval	0.153(3)	0.174(10)	0.181(12)	0.175(11)	0.153(2)	0.169(7)	0.173(9)	0.142(1)	0.153(4)	0.245(14)	0.173(8)	0.163(6)	0.184(13)	0.155(5)
	T1-CK	0.057(10)	0.039(7)	0.043(8)	0.052(9)	0.058(13)	0.027(3)	0.033(6)	0.032(5)	0.058(11)	0.070(14)	0.027(2)	0.025(1)	0.029(4)	0.058(12)
	T1-DCI	0.269(7)	0.310(12)	0.318(13)	0.331(14)	0.270(8)	0.282(10)	0.276(9)	0.256(4)	0.269(5)	0.300(11)	0.252(1)	0.255(3)	0.255(2)	0.269(6)
	T1-PCK	0.086(11)	0.084(8)	0.084(7)	0.080(6)	0.086(13)	0.075(3)	0.075(4)	0.076(5)	0.085(10)	0.119(14)	0.074(1)	0.085(9)	0.074(2)	0.086(12)
	T1-SEP	0.139(2)	0.227(14)	0.221(13)	0.208(9)	0.140(4)	0.218(12)	0.216(11)	0.210(10)	0.139(3)	0.154(5)	0.200(8)	0.190(7)	0.180(6)	0.138(1)
	T2-CK	0.056(7)	0.034(6)	0.032(5)	0.031(4)	0.056(9)	0.025(3)	0.022(2)	0.019(1)	0.056(8)	0.062(12)	0.067(13)	0.081(14)	0.062(11)	0.056(10)
	T2-DCI	0.130(7)	0.150(11)	0.150(12)	0.159(13)	0.130(8)	0.133(10)	0.126(5)	0.110(1)	0.129(6)	0.199(14)	0.125(4)	0.112(2)	0.124(3)	0.132(9)
	T2-PCK	0.054(4)	0.068(8)	0.078(13)	0.076(12)	0.053(2)	0.072(10)	0.082(14)	0.070(9)	0.053(3)	0.074(11)	0.055(5)	0.057(6)	0.067(7)	0.053(1)
	T2-SEP	0.103(10)	0.109(12)	0.114(13)	0.120(14)	0.102(7)	0.077(2)	0.079(3)	0.072(1)	0.102(8)	0.109(11)	0.087(6)	0.084(4)	0.086(5)	0.103(9)
	U41Q1	0.049(11)	0.044(6)	0.041(5)	0.039(4)	0.050(13)	0.035(2)	0.035(3)	0.027(1)	0.049(12)	0.048(10)	0.047(9)	0.045(7)	0.046(8)	0.050(14)
	U41Q2	0.064(4)	0.086(14)	0.086(13)	0.084(11)	0.065(5)	0.085(12)	0.080(9)	0.080(10)	0.065(6)	0.070(8)	0.058(2)	0.060(3)	0.049(1)	0.065(7)
	U42Q1	0.039(4)	0.045(7)	0.046(8)	0.047(10)	0.039(3)	0.047(9)	0.050(12)	0.049(11)	0.038(2)	0.052(14)	0.045(6)	0.043(5)	0.052(13)	0.038(1)
	U42Q2	0.057(8)	0.052(6)	0.054(7)	0.048(4)	0.057(9)	0.059(13)	0.058(12)	0.046(1)	0.057(10)	0.071(14)	0.047(2)	0.047(3)	0.051(5)	0.058(11)

Table 5: Metric value and ranking results of each uncertainty quantification methods with Claude-Haiku-4.5 over 14 questions.

Control	Question	AUROC													
		CE	Embed_Eigen	Embed_GE	Embed_NAD	FSD	JS_Eigen	JS_GE	JS_NAD	MAR	NLI_DSE	NLI_Eigen	NLI_GE	NLI_NAD	Numset
shot=0, cot=0	ASAP	0.693(2)	0.594(7)	0.589(8)	0.603(6)	0.689(3)	0.567(9)	0.546(10)	0.610(5)	0.693(1)	0.495(11)	0.396(14)	0.465(12)	0.453(13)	0.667(4)
	Semeval	0.583(6)	0.558(10)	0.566(9)	0.590(2)	0.587(4)	0.602(1)	0.573(8)	0.582(7)	0.586(5)	0.531(11)	0.506(13)	0.527(12)	0.503(14)	0.589(3)
	T1-CK	0.701(2)	0.540(13)	0.553(12)	0.604(8)	0.697(3)	0.611(7)	0.619(6)	0.684(4)	0.702(1)	0.532(14)	0.592(11)	0.595(10)	0.603(9)	0.665(5)
	T1-DCI	0.517(9)	0.532(8)	0.506(14)	0.517(12)	0.517(9)	0.763(1)	0.748(2)	0.722(3)	0.517(9)	0.537(6)	0.534(7)	0.597(4)	0.577(5)	0.515(13)
	T1-PCK	0.654(4)	0.416(12)	0.409(13)	0.392(14)	0.655(1)	0.529(7)	0.486(10)	0.494(9)	0.655(1)	0.494(8)	0.463(11)	0.639(5)	0.654(3)	0.630(6)
	T1-SEP	0.698(1)	0.483(9)	0.481(10)	0.533(8)	0.693(2)	0.579(7)	0.594(6)	0.671(4)	0.692(3)	0.469(11)	0.393(14)	0.434(12)	0.396(13)	0.656(5)
	T2-CK	0.790(3)	0.574(10)	0.589(9)	0.593(8)	0.815(1)	0.635(7)	0.637(6)	0.700(5)	0.809(2)	0.488(14)	0.534(12)	0.553(11)	0.534(13)	0.705(4)
	T2-DCI	0.664(4)	0.493(13)	0.486(14)	0.500(12)	0.659(5)	0.620(8)	0.616(9)	0.589(10)	0.659(6)	0.508(11)	0.653(7)	0.692(1)	0.687(2)	0.686(3)
	T2-PCK	0.690(1)	0.399(13)	0.392(14)	0.405(12)	0.689(2)	0.458(10)	0.447(11)	0.461(9)	0.689(3)	0.512(8)	0.588(7)	0.651(6)	0.657(5)	0.679(4)
	T2-SEP	0.788(2)	0.572(10)	0.573(9)	0.603(8)	0.783(4)	0.780(5)	0.768(6)	0.797(1)	0.785(3)	0.467(14)	0.511(13)	0.537(12)	0.544(11)	0.742(7)
	U41Q1	0.403(14)	0.595(4)	0.586(5)	0.566(6)	0.424(11)	0.644(3)	0.665(2)	0.708(1)	0.421(13)	0.465(8)	0.454(9)	0.452(10)	0.422(12)	0.478(7)
	U41Q2	0.597(2)	0.557(7)	0.574(6)	0.590(5)	0.592(4)	0.533(10)	0.556(8)	0.616(1)	0.595(3)	0.508(12)	0.462(14)	0.466(13)	0.515(11)	0.537(9)
	U42Q1	0.641(4)	0.482(14)	0.496(12)	0.491(13)	0.660(1)	0.611(6)	0.568(9)	0.579(8)	0.657(2)	0.517(11)	0.561(10)	0.656(3)	0.608(7)	0.616(5)
	U42Q2	0.671(1)	0.541(7)	0.522(10)	0.472(14)	0.661(3)	0.524(9)	0.518(11)	0.510(12)	0.665(2)	0.507(13)	0.572(6)	0.535(8)	0.616(4)	0.575(5)
shot=0, cot=1	ASAP	0.638(3)	0.550(12)	0.557(11)	0.598(5)	0.642(1)	0.564(10)	0.568(7)	0.618(4)	0.641(2)	0.500(14)	0.541(13)	0.571(6)	0.567(9)	0.568(8)
	Semeval	0.594(6)	0.576(10)	0.590(7)	0.607(3)	0.587(9)	0.621(1)	0.612(2)	0.598(4)	0.589(8)	0.500(13)	0.495(14)	0.546(11)	0.526(12)	0.596(5)
	T1-CK	0.723(1)	0.571(13)	0.579(11)	0.590(10)	0.709(3)	0.635(9)	0.636(8)	0.641(7)	0.713(2)	0.508(14)	0.579(12)	0.657(5)	0.678(4)	0.644(6)
	T1-DCI	0.584(4)	0.564(8)	0.547(9)	0.538(10)	0.584(4)	0.726(2)	0.734(1)	0.722(3)	0.584(4)	0.487(14)	0.505(12)	0.506(11)	0.497(13)	0.579(7)
	T1-PCK	0.751(2)	0.482(12)	0.481(13)	0.456(14)	0.736(4)	0.532(8)	0.494(10)	0.483(11)	0.740(3)	0.500(9)	0.554(7)	0.764(1)	0.727(5)	0.680(6)
	T1-SEP	0.758(1)	0.524(11)	0.529(10)	0.516(12)	0.750(3)	0.691(5)	0.669(8)	0.686(6)	0.754(2)	0.500(14)	0.505(13)	0.671(7)	0.612(9)	0.705(4)
	T2-CK	0.661(1)	0.539(8)	0.535(10)	0.556(7)	0.656(3)	0.486(13)	0.483(14)	0.539(9)	0.660(2)	0.500(11)	0.498(12)	0.632(5)	0.629(6)	0.636(4)
	T2-DCI	0.772(1)	0.637(13)	0.650(11)	0.669(9)	0.766(3)	0.749(6)	0.756(5)	0.744(7)	0.768(2)	0.500(14)	0.639(12)	0.682(8)	0.667(10)	0.764(4)
	T2-PCK	0.662(3)	0.459(9)	0.451(10)	0.332(14)	0.664(1)	0.425(11)	0.371(12)	0.332(13)	0.664(2)	0.500(8)	0.553(7)	0.604(6)	0.623(5)	0.650(4)
	T2-SEP	0.796(3)	0.532(11)	0.520(12)	0.544(9)	0.802(2)	0.698(7)	0.712(6)	0.731(4)	0.804(1)	0.500(13)	0.487(14)	0.593(8)	0.539(10)	0.719(5)
	U41Q1	0.417(13)	0.431(10)	0.436(9)	0.471(5)	0.457(6)	0.560(2)	0.558(3)	0.605(1)	0.448(7)	0.500(4)	0.414(14)	0.447(8)	0.430(11)	0.419(12)
	U41Q2	0.596(2)	0.558(6)	0.537(9)	0.534(10)	0.569(5)	0.539(8)	0.522(12)	0.552(7)	0.572(4)	0.517(13)	0.474(14)	0.582(3)	0.526(11)	0.644(1)
	U42Q1	0.727(1)	0.474(14)	0.487(11)	0.526(8)	0.720(3)	0.481(13)	0.485(12)	0.508(9)	0.722(2)	0.500(10)	0.553(7)	0.610(6)	0.624(5)	0.684(4)
	U42Q2	0.507(4)	0.504(5)	0.501(7)	0.487(9)	0.557(1)	0.518(3)	0.484(10)	0.466(11)	0.549(2)	0.500(8)	0.428(14)	0.501(6)	0.451(13)	0.455(12)
shot=2, cot=1	ASAP	0.644(5)	0.624(8)	0.632(7)	0.607(10)	0.659(3)	0.641(6)	0.662(2)	0.670(1)	0.656(4)	0.500(13)	0.461(14)	0.518(12)	0.533(11)	0.613(9)
	Semeval	0.596(6)	0.691(3)	0.701(1)	0.696(2)	0.591(9)	0.603(5)	0.612(4)	0.588(10)	0.592(8)	0.492(12)	0.487(13)	0.438(14)	0.508(11)	0.595(7)
	T1-CK	0.685(5)	0.603(13)	0.612(11)	0.604(12)	0.679(6)	0.665(8)	0.665(7)	0.693(3)	0.688(4)	0.500(14)	0.613(10)	0.699(2)	0.710(1)	0.647(9)
	T1-DCI	0.616(6)	0.564(9)	0.575(8)	0.526(10)	0.619(4)	0.644(3)	0.653(2)	0.664(1)	0.618(5)	0.494(11)	0.417(13)	0.414(14)	0.419(12)	0.609(7)
	T1-PCK	0.722(3)	0.432(13)	0.410(14)	0.449(12)	0.722(5)	0.526(10)	0.527(9)	0.566(7)	0.722(4)	0.500(11)	0.535(8)	0.692(6)	0.728(2)	0.730(1)
	T1-SEP	0.693(3)	0.485(11)	0.482(12)	0.469(13)	0.697(1)	0.561(7)	0.553(9)	0.568(6)	0.695(2)	0.496(10)	0.468(14)	0.599(5)	0.553(8)	0.668(4)
	T2-CK	0.614(7)	0.663(4)	0.668(3)	0.718(1)	0.550(12)	0.595(8)	0.571(11)	0.586(9)	0.575(10)	0.500(14)	0.544(13)	0.676(2)	0.615(6)	0.653(5)
	T2-DCI	0.628(5)	0.463(14)	0.464(13)	0.470(12)	0.619(8)	0.629(4)	0.614(9)	0.627(6)	0.622(7)	0.500(11)	0.583(10)	0.673(1)	0.660(3)	0.671(2)
	T2-PCK	0.650(3)	0.532(8)	0.523(9)	0.460(11)	0.648(5)	0.397(13)	0.389(14)	0.443(12)	0.649(4)	0.500(10)	0.573(7)	0.617(6)	0.665(1)	0.656(2)
	T2-SEP	0.612(2)	0.406(11)	0.408(10)	0.484(9)	0.610(4)	0.604(6)	0.605(5)	0.699(1)	0.612(3)	0.500(8)	0.373(13)	0.391(12)	0.359(14)	0.580(7)
	U41Q1	0.707(1)	0.509(12)	0.530(10)	0.520(11)	0.697(3)	0.610(7)	0.612(5)	0.649(4)	0.700(2)	0.500(14)	0.503(13)	0.543(9)	0.559(8)	0.612(6)
	U41Q2	0.650(5)	0.679(1)	0.669(2)	0.643(7)	0.667(3)	0.633(8)	0.615(9)	0.644(6)	0.662(4)	0.500(12)	0.435(14)	0.499(13)	0.506(11)	0.592(10)
	U42Q1	0.733(3)	0.677(4)	0.673(5)	0.636(7)	0.751(1)	0.557(11)	0.518(12)	0.512(13)	0.750(2)	0.502(14)	0.617(8)	0.612(9)	0.597(10)	0.658(6)
	U42Q2	0.663(1)	0.466(14)	0.478(12)	0.480(11)	0.646(4)	0.574(7)	0.546(9)	0.554(8)	0.648(2)	0.500(10)	0.475(13)	0.648(3)	0.629(5)	0.616(6)

Control	Question	C-index													
		CE	Embed_Eigen	Embed_GE	Embed_NAD	FSD	JS_Eigen	JS_GE	JS_NAD	MAR	NLI_DSE	NLI_Eigen	NLI_GE	NLI_NAD	Numset
shot=0, cot=0	ASAP	0.651(2)	0.611(5)	0.606(7)	0.607(6)	0.649(3)	0.565(9)	0.552(10)	0.599(8)	0.652(1)	0.494(11)	0.426(14)	0.468(12)	0.465(13)	0.637(4)
	Semeval	0.583(6)	0.558(10)	0.566(9)	0.590(2)	0.587(4)	0.602(1)	0.573(8)	0.582(7)	0.586(5)	0.531(11)	0.506(13)	0.527(12)	0.503(14)	0.589(3)
	T1-CK	0.592(4)	0.456(14)	0.470(13)	0.512(12)	0.594(3)	0.552(8)	0.555(7)	0.630(1)	0.596(2)	0.530(11)	0.545(9)	0.556(6)	0.533(10)	0.567(5)
	T1-DCI	0.517(9)	0.532(8)	0.506(14)	0.517(12)	0.517(9)	0.763(1)	0.748(2)	0.722(3)	0.517(9)	0.537(6)	0.534(7)	0.597(4)	0.577(5)	0.515(13)
	T1-PCK	0.658(3)	0.414(12)	0.408(13)	0.396(14)	0.659(1)	0.527(7)	0.488(10)	0.496(8)	0.659(1)	0.495(9)	0.470(11)	0.630(6)	0.645(4)	0.632(5)
	T1-SEP	0.635(2)	0.445(10)	0.444(11)	0.491(8)	0.635(3)	0.557(7)	0.571(6)	0.640(1)	0.633(4)	0.460(9)	0.367(14)	0.396(12)	0.373(13)	0.602(5)
	T2-CK	0.746(3)	0.539(10)	0.553(9)	0.559(8)	0.780(1)	0.601(6)	0.601(7)	0.671(4)	0.770(2)	0.486(14)	0.502(12)	0.521(11)	0.502(13)	0.665(5)
	T2-DCI	0.616(6)	0.495(12)	0.484(14)	0.493(13)	0.616(5)	0.609(8)	0.602(9)	0.562(10)	0.615(7)	0.511(11)	0.649(3)	0.672(1)	0.665(2)	0.640(4)
	T2-PCK	0.692(1)	0.399(13)	0.392(14)	0.404(12)	0.692(2)	0.458(10)	0.449(11)	0.463(9)	0.692(3)	0.511(8)	0.590(7)	0.652(6)	0.659(5)	0.679(4)
	T2-SEP	0.783(2)	0.573(9)	0.573(9)	0.597(8)	0.779(4)	0.773(5)	0.760(6)	0.790(1)	0.780(3)	0.467(14)	0.511(13)	0.536(12)	0.543(11)	0.740(7)
	U41Q1	0.524(9)	0.602(4)	0.586(5)	0.558(6)	0.545(7)	0.607(3)	0.614(2)	0.647(1)	0.543(8)	0.477(13)	0.509(10)	0.484(12)	0.466(14)	0.486(11)
	U41Q2	0.581(4)	0.574(7)	0.592(3)	0.608(2)	0.577(6)	0.555(9)	0.573(8)	0.620(1)	0.581(5)	0.515(11)	0.450(14)	0.454(13)	0.485(12)	0.526(10)
	U42Q1	0.647(4)	0.492(13)	0.501(12)	0.490(14)	0.666(1)	0.594(7)	0.555(10)	0.565(9)	0.662(2)	0.528(11)	0.570(8)	0.661(3)	0.602(6)	0.614(5)
	U42Q2	0.667(1)	0.528(8)	0.512(11)	0.468(14)	0.659(3)	0.518(9)	0.514(10)	0.508(12)	0.662(2)	0.506(13)	0.570(6)	0.531(7)	0.617(4)	0.573(5)
shot=0, cot=1	ASAP	0.605(3)	0.547(9)	0.554(6)	0.588(5)	0.615(1)	0.541(12)	0.547(10)	0.588(4)	0.612(2)	0.500(14)	0.546(11)	0.548(8)	0.553(7)	0.537(13)
	Semeval	0.594(6)	0.576(10)	0.590(7)	0.607(3)	0.587(9)	0.621(1)	0.612(2)	0.598(4)	0.589(8)	0.500(13)	0.495(14)	0.546(11)	0.526(12)	0.596(5)
	T1-CK	0.718(1)	0.584(13)	0.595(11)	0.602(10)	0.714(3)	0.645(7)	0.653(6)	0.659(5)	0.715(2)	0.505(14)	0.589(12)	0.637(8)	0.660(4)	0.627(9)
	T1-DCI	0.584(4)	0.564(8)	0.547(9)	0.538(10)	0.584(4)	0.726(2)	0.734(1)	0.722(3)	0.584(4)	0.487(14)	0.505(12)	0.506(11)	0.497(13)	0.579(7)
	T1-PCK	0.755(1)	0.479(13)	0.483(11)	0.460(14)	0.742(4)	0.533(8)	0.498(10)	0.479(12)	0.746(3)	0.500(9)	0.550(7)	0.749(2)	0.714(5)	0.673(6)
	T1-SEP	0.688(3)	0.482(13)	0.487(12)	0.478(14)	0.694(2)	0.631(6)	0.613(8)	0.638(5)	0.695(1)	0.500(10)	0.488(11)	0.631(7)	0.571(9)	0.642(4)
	T2-CK	0.676(1)	0.538(8)	0.537(10)	0.562(7)	0.670(3)	0.484(13)	0.482(14)	0.537(9)	0.676(2)	0.500(11)	0.497(12)	0.623(5)	0.620(6)	0.642(4)
	T2-DCI	0.701(6)	0.597(13)	0.613(11)	0.640(8)	0.710(4)	0.719(2)	0.724(1)	0.710(3)	0.710(5)	0.500(14)	0.603(12)	0.618(10)	0.620(9)	0.684(7)
	T2-PCK	0.668(3)	0.462(9)	0.454(10)	0.343(13)	0.670(1)	0.422(11)	0.371(12)	0.335(14)	0.670(2)	0.500(8)	0.551(7)	0.601(6)	0.621(5)	0.649(4)
	T2-SEP	0.774(3)	0.516(11)	0.506(12)	0.532(9)	0.782(2)	0.682(7)	0.696(6)	0.712(4)	0.782(1)	0.500(13)	0.486(14)	0.583(8)	0.529(10)	0.702(5)
	U41Q1	0.453(9)	0.446(12)	0.442(13)	0.463(8)	0.491(5)	0.568(2)	0.564(3)	0.587(1)	0.480(6)	0.500(4)	0.451(11)	0.472(7)	0.452(10)	0.421(14)
	U41Q2	0.594(2)	0.563(6)	0.542(9)	0.523(11)	0.565(4)	0.549(8)	0.531(10)	0.562(7)	0.569(3)	0.514(12)	0.462(14)	0.564(5)	0.511(13)	0.635(1)
	U42Q1	0.697(2)	0.456(14)	0.467(11)	0.511(8)	0.698(1)	0.464(13)	0.466(12)	0.489(10)	0.696(3)	0.500(9)	0.555(7)	0.608(6)	0.624(5)	0.653(4)
	U42Q2	0.510(7)	0.517(4)	0.513(5)	0.499(9)	0.557(1)	0.523(3)	0.492(10)	0.472(11)	0.550(2)	0.500(8)	0.441(14)	0.512(6)	0.466(12)	0.460(13)
shot=2, cot=1	ASAP	0.627(6)	0.624(7)	0.628(5)	0.594(10)	0.643(1)	0.622(8)	0.641(3)	0.642(2)	0.639(4)	0.500(13)	0.464(14)	0.506(12)	0.520(11)	0.594(9)
	Semeval	0.596(6)	0.691(3)	0.701(1)	0.696(2)	0.591(9)	0.603(5)	0.612(4)	0.588(10)	0.592(8)	0.492(12)	0.487(13)	0.438(14)	0.508(11)	0.595(7)
	T1-CK	0.641(4)	0.570(13)	0.580(11)	0.575(12)	0.635(6)	0.628(7)	0.626(8)	0.669(2)	0.640(5)	0.500(14)	0.583(10)	0.665(3)	0.673(1)	0.608(9)
	T1-DCI	0.616(6)	0.564(9)	0.575(8)	0.526(10)	0.619(4)	0.644(3)	0.653(2)	0.664(1)	0.618(5)	0.494(11)	0.417(13)	0.414(14)	0.419(12)	0.609(7)
	T1-PCK	0.726(1)	0.430(13)	0.410(14)	0.452(12)	0.725(2)	0.522(10)	0.524(9)	0.562(7)	0.725(2)	0.500(11)	0.528(8)	0.686(6)	0.718(5)	0.724(4)
	T1-SEP	0.645(3)	0.480(11)	0.472(12)	0.458(14)	0.648(1)	0.531(8)	0.525(9)	0.539(6)	0.646(2)	0.495(10)	0.469(13)	0.575(5)	0.534(7)	0.630(4)
	T2-CK	0.643(6)	0.674(4)	0.679(2)	0.727(1)	0.592(9)	0.584(10)	0.560(12)	0.575(11)	0.611(8)	0.500(14)	0.553(13)	0.677(3)	0.637(7)	0.663(5)
	T2-DCI	0.626(5)	0.469(14)	0.471(13)	0.479(12)	0.623(7)	0.628(4)	0.611(9)	0.619(8)	0.624(6)	0.500(11)	0.570(10)	0.648(2)	0.639(3)	0.658(1)
	T2-PCK	0.657(2)	0.536(8)	0.528(9)	0.469(11)	0.655(5)	0.398(13)	0.388(14)	0.444(12)	0.656(3)	0.500(10)	0.567(7)	0.618(6)	0.666(1)	0.656(4)
	T2-SEP	0.611(2)	0.405(11)	0.409(10)	0.484(9)	0.609(4)	0.604(6)	0.605(5)	0.696(1)	0.611(3)	0.500(8)	0.373(13)	0.391(12)	0.359(14)	0.579(7)
	U41Q1	0.694(1)	0.510(12)	0.530(10)	0.523(11)	0.680(3)	0.606(6)	0.608(5)	0.649(4)	0.684(2)	0.500(13)	0.487(14)	0.535(9)	0.546(8)	0.604(7)
	U41Q2	0.585(9)	0.644(1)	0.631(2)	0.603(5)	0.601(6)	0.612(4)	0.598(7)	0.625(3)	0.595(8)	0.500(11)	0.418(14)	0.466(13)	0.479(12)	0.541(10)
	U42Q1	0.715(3)	0.687(4)	0.684(5)	0.641(7)	0.732(1)	0.550(11)	0.512(12)	0.507(13)	0.731(2)	0.502(14)	0.598(8)	0.593(9)	0.578(10)	0.643(6)
	U42Q2	0.664(1)	0.471(14)	0.481(12)	0.484(11)	0.648(3)	0.576(7)	0.547(9)	0.557(8)	0.650(2)	0.500(10)	0.477(13)	0.644(4)	0.627(5)	0.614(6)

Control	Question	AUARC													
		CE	Embed_Eigen	Embed_GE	Embed_NAD	FSD	JS_Eigen	JS_GE	JS_NAD	MAR	NLI_DSE	NLI_Eigen	NLI_GE	NLI_NAD	Numset
shot=0, cot=0	ASAP	0.677(1)	0.612(7)	0.600(10)	0.601(9)	0.675(3)	0.622(6)	0.609(8)	0.641(5)	0.676(2)	0.579(11)	0.459(14)	0.495(12)	0.478(13)	0.669(4)
	Semeval	0.571(5)	0.557(8)	0.559(7)	0.581(1)	0.572(2)	0.560(6)	0.536(11)	0.543(9)	0.572(3)	0.526(12)	0.502(14)	0.537(10)	0.519(13)	0.571(4)
	T1-CK	0.520(3)	0.407(13)	0.422(10)	0.460(8)	0.523(1)	0.479(7)	0.482(6)	0.507(5)	0.520(2)	0.364(14)	0.414(12)	0.418(11)	0.438(9)	0.518(4)
	T1-DCI	0.722(5)	0.651(11)	0.625(14)	0.627(13)	0.722(5)	0.815(1)	0.804(2)	0.790(3)	0.722(5)	0.738(4)	0.639(12)	0.711(9)	0.677(10)	0.721(8)
	T1-PCK	0.765(5)	0.636(13)	0.631(14)	0.637(12)	0.766(4)	0.755(7)	0.716(9)	0.724(8)	0.766(3)	0.683(11)	0.698(10)	0.803(2)	0.810(1)	0.761(6)
	T1-SEP	0.656(1)	0.510(10)	0.508(11)	0.534(8)	0.653(2)	0.569(7)	0.574(6)	0.634(5)	0.651(3)	0.526(9)	0.449(14)	0.482(12)	0.457(13)	0.644(4)
	T2-CK	0.775(3)	0.651(10)	0.658(9)	0.672(8)	0.783(2)	0.693(6)	0.687(7)	0.725(5)	0.784(1)	0.549(14)	0.591(12)	0.615(11)	0.584(13)	0.726(4)
	T2-DCI	0.648(4)	0.551(13)	0.546(14)	0.571(12)	0.648(5)	0.617(9)	0.624(8)	0.608(10)	0.647(6)	0.573(11)	0.644(7)	0.678(1)	0.650(3)	0.662(2)
	T2-PCK	0.775(2)	0.572(12)	0.567(14)	0.568(13)	0.775(3)	0.613(10)	0.612(11)	0.644(9)	0.775(3)	0.723(8)	0.730(7)	0.758(5)	0.758(6)	0.779(1)
	T2-SEP	0.858(4)	0.754(10)	0.760(9)	0.785(8)	0.857(6)	0.870(2)	0.865(3)	0.875(1)	0.857(5)	0.587(14)	0.746(13)	0.754(11)	0.752(12)	0.821(7)
	U41Q1	0.168(14)	0.237(6)	0.247(5)	0.250(4)	0.172(12)	0.283(3)	0.292(2)	0.330(1)	0.172(11)	0.208(7)	0.186(8)	0.181(10)	0.168(13)	0.186(9)
	U41Q2	0.532(5)	0.511(8)	0.530(6)	0.536(3)	0.536(4)	0.487(10)	0.517(7)	0.552(1)	0.538(2)	0.409(14)	0.429(13)	0.440(12)	0.482(11)	0.505(9)
	U42Q1	0.652(3)	0.568(14)	0.575(13)	0.575(12)	0.660(1)	0.618(6)	0.594(11)	0.596(10)	0.659(2)	0.605(8)	0.603(9)	0.635(5)	0.617(7)	0.649(4)
	U42Q2	0.594(2)	0.505(12)	0.490(13)	0.446(14)	0.591(3)	0.518(8)	0.513(10)	0.508(11)	0.594(1)	0.517(9)	0.541(6)	0.523(7)	0.574(4)	0.552(5)
shot=0, cot=1	ASAP	0.671(4)	0.624(12)	0.624(13)	0.653(10)	0.671(5)	0.664(6)	0.674(2)	0.697(1)	0.672(3)	0.613(14)	0.641(11)	0.662(7)	0.659(8)	0.654(9)
	Semeval	0.562(2)	0.533(9)	0.551(6)	0.570(1)	0.560(4)	0.537(8)	0.523(10)	0.538(7)	0.560(3)	0.490(11)	0.462(14)	0.474(12)	0.473(13)	0.559(5)
	T1-CK	0.613(2)	0.517(11)	0.516(12)	0.494(13)	0.609(4)	0.529(9)	0.533(8)	0.522(10)	0.610(3)	0.465(14)	0.567(7)	0.606(5)	0.620(1)	0.578(6)
	T1-DCI	0.743(4)	0.629(12)	0.618(13)	0.600(14)	0.743(4)	0.744(3)	0.750(1)	0.747(2)	0.743(4)	0.675(8)	0.656(9)	0.644(10)	0.640(11)	0.739(7)
	T1-PCK	0.795(3)	0.678(10)	0.682(9)	0.657(13)	0.790(5)	0.683(8)	0.674(11)	0.665(12)	0.792(4)	0.630(14)	0.742(7)	0.840(1)	0.824(2)	0.763(6)
	T1-SEP	0.704(1)	0.564(11)	0.549(13)	0.541(14)	0.702(2)	0.675(6)	0.665(8)	0.669(7)	0.700(3)	0.585(10)	0.557(12)	0.679(5)	0.634(9)	0.683(4)
	T2-CK	0.666(3)	0.630(7)	0.628(8)	0.626(9)	0.664(5)	0.583(12)	0.583(13)	0.614(10)	0.664(4)	0.519(14)	0.610(11)	0.712(1)	0.703(2)	0.658(6)
	T2-DCI	0.656(1)	0.513(13)	0.529(12)	0.546(10)	0.653(3)	0.638(7)	0.641(5)	0.641(6)	0.655(2)	0.498(14)	0.553(9)	0.538(11)	0.556(8)	0.651(4)
	T2-PCK	0.751(3)	0.601(10)	0.612(9)	0.546(12)	0.751(2)	0.562(11)	0.524(13)	0.496(14)	0.751(1)	0.680(8)	0.724(7)	0.745(5)	0.747(4)	0.744(6)
	T2-SEP	0.871(3)	0.743(11)	0.736(12)	0.755(9)	0.872(1)	0.845(6)	0.846(5)	0.866(4)	0.871(2)	0.601(14)	0.700(13)	0.779(8)	0.746(10)	0.833(7)
	U41Q1	0.238(10)	0.232(11)	0.227(13)	0.249(7)	0.252(6)	0.337(2)	0.332(3)	0.364(1)	0.249(8)	0.281(4)	0.228(12)	0.245(9)	0.226(14)	0.266(5)
	U41Q2	0.636(2)	0.566(8)	0.553(12)	0.557(11)	0.628(4)	0.568(7)	0.563(9)	0.594(6)	0.633(3)	0.443(14)	0.526(13)	0.610(5)	0.560(10)	0.650(1)
	U42Q1	0.732(1)	0.594(10)	0.601(9)	0.627(7)	0.731(3)	0.565(14)	0.569(13)	0.580(12)	0.731(2)	0.614(8)	0.594(11)	0.641(6)	0.645(5)	0.714(4)
	U42Q2	0.611(4)	0.604(5)	0.603(6)	0.568(11)	0.625(1)	0.611(3)	0.600(7)	0.580(8)	0.625(2)	0.571(10)	0.540(14)	0.574(9)	0.553(13)	0.564(12)
shot=2, cot=1	ASAP	0.734(6)	0.692(10)	0.695(8)	0.692(9)	0.741(4)	0.748(3)	0.756(2)	0.765(1)	0.739(5)	0.663(13)	0.629(14)	0.669(12)	0.671(11)	0.725(7)
	Semeval	0.535(7)	0.630(3)	0.641(2)	0.643(1)	0.534(9)	0.548(6)	0.553(4)	0.551(5)	0.534(8)	0.491(11)	0.459(13)	0.416(14)	0.472(12)	0.522(10)
	T1-CK	0.804(5)	0.760(13)	0.764(12)	0.764(11)	0.804(4)	0.794(8)	0.792(9)	0.805(3)	0.803(6)	0.675(14)	0.794(7)	0.834(2)	0.834(1)	0.786(10)
	T1-DCI	0.726(4)	0.663(9)	0.674(8)	0.651(10)	0.727(2)	0.721(7)	0.725(6)	0.749(1)	0.727(3)	0.645(11)	0.590(14)	0.591(13)	0.617(12)	0.726(5)
	T1-PCK	0.810(3)	0.656(13)	0.643(14)	0.663(12)	0.810(4)	0.722(10)	0.727(9)	0.754(7)	0.809(5)	0.686(11)	0.748(8)	0.825(2)	0.835(1)	0.800(6)
	T1-SEP	0.721(2)	0.543(14)	0.558(13)	0.568(12)	0.721(1)	0.630(8)	0.627(9)	0.653(6)	0.721(3)	0.581(10)	0.574(11)	0.662(5)	0.645(7)	0.710(4)
	T2-CK	0.821(8)	0.869(2)	0.868(3)	0.885(1)	0.802(12)	0.822(7)	0.818(9)	0.824(6)	0.806(11)	0.761(14)	0.778(13)	0.844(4)	0.816(10)	0.834(5)
	T2-DCI	0.610(4)	0.498(13)	0.497(14)	0.502(12)	0.609(5)	0.605(7)	0.594(10)	0.599(8)	0.607(6)	0.558(11)	0.594(9)	0.652(1)	0.643(2)	0.632(3)
	T2-PCK	0.774(2)	0.666(9)	0.659(10)	0.625(11)	0.773(4)	0.573(13)	0.567(14)	0.599(12)	0.774(3)	0.727(7)	0.713(8)	0.750(6)	0.770(5)	0.784(1)
	T2-SEP	0.767(5)	0.705(10)	0.710(9)	0.739(7)	0.767(6)	0.803(3)	0.805(2)	0.855(1)	0.769(4)	0.660(13)	0.666(12)	0.677(11)	0.622(14)	0.736(8)
	U41Q1	0.767(1)	0.589(14)	0.603(13)	0.604(12)	0.765(3)	0.711(6)	0.708(7)	0.742(5)	0.766(2)	0.618(11)	0.631(10)	0.677(9)	0.683(8)	0.742(4)
	U41Q2	0.752(4)	0.714(9)	0.722(8)	0.709(11)	0.757(1)	0.743(5)	0.736(6)	0.753(3)	0.757(2)	0.666(13)	0.650(14)	0.703(12)	0.710(10)	0.723(7)
	U42Q1	0.824(3)	0.804(6)	0.806(4)	0.787(7)	0.828(2)	0.740(10)	0.712(12)	0.704(13)	0.828(1)	0.703(14)	0.756(9)	0.729(11)	0.768(8)	0.805(5)
	U42Q2	0.810(1)	0.687(13)	0.688(12)	0.697(10)	0.804(2)	0.768(7)	0.747(8)	0.742(9)	0.803(3)	0.635(14)	0.695(11)	0.800(4)	0.788(5)	0.770(6)

Control	Question	AUERC													
		CE	Embed_Eigen	Embed_GE	Embed_NAD	FSD	JS_Eigen	JS_GE	JS_NAD	MAR	NLI_DSE	NLI_Eigen	NLI_GE	NLI_NAD	Numset
shot=0, cot=0	ASAP	0.144(5)	0.137(1)	0.140(3)	0.146(9)	0.145(8)	0.145(7)	0.147(10)	0.138(2)	0.144(6)	0.175(11)	0.204(14)	0.203(13)	0.202(12)	0.143(4)
	Semeval	0.419(5)	0.433(8)	0.431(7)	0.409(1)	0.418(2)	0.430(6)	0.454(11)	0.447(9)	0.418(3)	0.464(12)	0.488(14)	0.453(10)	0.471(13)	0.419(4)
	T1-CK	0.292(5)	0.371(14)	0.357(12)	0.325(8)	0.289(2)	0.298(6)	0.299(7)	0.271(1)	0.290(3)	0.357(13)	0.352(11)	0.337(9)	0.341(10)	0.291(4)
	T1-DCI	0.268(5)	0.339(11)	0.365(14)	0.363(13)	0.268(5)	0.175(1)	0.186(2)	0.200(3)	0.268(5)	0.252(4)	0.351(12)	0.279(9)	0.313(10)	0.269(8)
	T1-PCK	0.119(5)	0.208(13)	0.211(14)	0.205(12)	0.118(4)	0.138(7)	0.158(9)	0.156(8)	0.118(3)	0.178(11)	0.165(10)	0.112(2)	0.110(1)	0.123(6)
	T1-SEP	0.228(2)	0.317(10)	0.319(11)	0.297(9)	0.229(3)	0.256(7)	0.249(6)	0.216(1)	0.230(4)	0.295(8)	0.350(14)	0.328(12)	0.334(13)	0.231(5)
	T2-CK	0.130(3)	0.229(10)	0.224(9)	0.209(8)	0.120(1)	0.196(6)	0.200(7)	0.166(5)	0.122(2)	0.265(12)	0.273(13)	0.251(11)	0.284(14)	0.165(4)
	T2-DCI	0.209(7)	0.266(13)	0.268(14)	0.251(12)	0.208(5)	0.215(9)	0.213(8)	0.229(10)	0.208(6)	0.241(11)	0.194(2)	0.182(1)	0.194(3)	0.196(4)
	T2-PCK	0.108(2)	0.213(12)	0.217(14)	0.216(13)	0.108(3)	0.193(11)	0.191(10)	0.175(9)	0.108(3)	0.136(8)	0.131(7)	0.117(6)	0.117(5)	0.106(1)
	T2-SEP	0.068(4)	0.119(10)	0.117(9)	0.108(8)	0.068(6)	0.064(2)	0.067(3)	0.061(1)	0.068(5)	0.204(14)	0.125(13)	0.121(11)	0.122(12)	0.087(7)
	U41Q1	0.229(9)	0.210(5)	0.209(4)	0.212(6)	0.225(7)	0.203(3)	0.201(2)	0.189(1)	0.225(8)	0.236(12)	0.231(10)	0.234(11)	0.240(14)	0.239(13)
	U41Q2	0.136(8)	0.131(5)	0.125(3)	0.123(2)	0.134(6)	0.137(9)	0.129(4)	0.120(1)	0.135(7)	0.170(13)	0.171(14)	0.165(12)	0.158(11)	0.145(10)
	U42Q1	0.086(3)	0.110(9)	0.111(11)	0.112(12)	0.084(1)	0.111(10)	0.114(14)	0.113(13)	0.084(2)	0.097(6)	0.099(8)	0.091(5)	0.098(7)	0.087(4)
	U42Q2	0.102(2)	0.137(12)	0.138(13)	0.147(14)	0.103(3)	0.131(9)	0.132(11)	0.132(10)	0.102(1)	0.127(8)	0.117(6)	0.125(7)	0.108(4)	0.114(5)
shot=0, cot=1	ASAP	0.137(7)	0.139(9)	0.138(8)	0.129(2)	0.135(5)	0.135(6)	0.130(3)	0.121(1)	0.135(4)	0.154(14)	0.140(10)	0.150(13)	0.145(11)	0.146(12)
	Semeval	0.428(2)	0.457(9)	0.439(6)	0.420(1)	0.430(4)	0.453(8)	0.467(10)	0.452(7)	0.430(3)	0.500(11)	0.528(14)	0.516(12)	0.517(13)	0.431(5)
	T1-CK	0.212(1)	0.264(12)	0.260(11)	0.272(13)	0.213(3)	0.252(10)	0.245(8)	0.250(9)	0.212(2)	0.329(14)	0.238(6)	0.225(5)	0.216(4)	0.242(7)
	T1-DCI	0.247(4)	0.361(12)	0.372(13)	0.390(14)	0.247(4)	0.246(3)	0.240(1)	0.243(2)	0.247(4)	0.315(8)	0.334(9)	0.346(10)	0.350(11)	0.251(7)
	T1-PCK	0.100(3)	0.177(11)	0.172(9)	0.184(12)	0.102(5)	0.166(8)	0.173(10)	0.189(13)	0.101(4)	0.200(14)	0.138(7)	0.084(1)	0.092(2)	0.126(6)
	T1-SEP	0.200(2)	0.297(12)	0.307(13)	0.307(14)	0.197(1)	0.205(4)	0.211(7)	0.205(5)	0.200(3)	0.285(10)	0.293(11)	0.206(6)	0.243(9)	0.220(8)
	T2-CK	0.165(3)	0.198(9)	0.196(8)	0.194(7)	0.167(5)	0.235(13)	0.230(12)	0.213(10)	0.166(4)	0.275(14)	0.213(11)	0.153(1)	0.158(2)	0.176(6)
	T2-DCI	0.202(6)	0.282(12)	0.268(10)	0.252(8)	0.200(5)	0.200(3)	0.198(1)	0.200(4)	0.199(2)	0.285(13)	0.260(9)	0.297(14)	0.270(11)	0.208(7)
	T2-PCK	0.121(3)	0.204(10)	0.198(9)	0.231(12)	0.121(2)	0.231(11)	0.254(13)	0.267(14)	0.121(1)	0.162(8)	0.141(7)	0.129(6)	0.127(4)	0.127(5)
	T2-SEP	0.063(3)	0.126(11)	0.129(12)	0.119(9)	0.062(1)	0.074(6)	0.073(5)	0.064(4)	0.063(2)	0.188(14)	0.144(13)	0.106(8)	0.124(10)	0.080(7)
	U41Q1	0.206(10)	0.213(13)	0.219(14)	0.212(12)	0.201(5)	0.175(2)	0.179(3)	0.173(1)	0.202(6)	0.205(9)	0.204(7)	0.199(4)	0.207(11)	0.205(8)
	U41Q2	0.098(2)	0.112(8)	0.115(10)	0.120(11)	0.101(4)	0.111(6)	0.113(9)	0.104(5)	0.100(3)	0.162(14)	0.137(13)	0.112(7)	0.128(12)	0.096(1)
	U42Q1	0.066(3)	0.101(11)	0.099(10)	0.092(8)	0.065(1)	0.113(14)	0.111(13)	0.108(12)	0.066(2)	0.091(7)	0.097(9)	0.086(6)	0.085(5)	0.071(4)
	U42Q2	0.101(6)	0.100(4)	0.100(5)	0.109(9)	0.097(1)	0.099(3)	0.103(7)	0.109(10)	0.097(2)	0.114(13)	0.117(14)	0.107(8)	0.113(12)	0.111(11)
shot=2, cot=1	ASAP	0.114(6)	0.124(8)	0.126(9)	0.131(10)	0.110(4)	0.103(3)	0.099(2)	0.098(1)	0.112(5)	0.142(11)	0.157(14)	0.144(12)	0.145(13)	0.119(7)
	Semeval	0.455(7)	0.360(3)	0.349(2)	0.347(1)	0.456(9)	0.442(6)	0.437(4)	0.439(5)	0.456(8)	0.499(11)	0.531(13)	0.574(14)	0.518(12)	0.468(10)
	T1-CK	0.092(5)	0.112(13)	0.110(12)	0.109(11)	0.092(4)	0.096(8)	0.098(9)	0.089(3)	0.093(6)	0.143(14)	0.095(7)	0.076(1)	0.076(2)	0.100(10)
	T1-DCI	0.264(4)	0.327(9)	0.316(8)	0.339(10)	0.263(2)	0.269(7)	0.265(6)	0.241(1)	0.263(3)	0.345(11)	0.400(14)	0.399(13)	0.373(12)	0.264(5)
	T1-PCK	0.092(4)	0.183(13)	0.189(14)	0.175(12)	0.091(3)	0.145(10)	0.144(9)	0.129(7)	0.092(5)	0.166(11)	0.134(8)	0.089(2)	0.086(1)	0.099(6)
	T1-SEP	0.201(1)	0.294(14)	0.289(12)	0.285(11)	0.201(3)	0.256(8)	0.258(9)	0.237(7)	0.201(2)	0.291(13)	0.275(10)	0.223(5)	0.233(6)	0.207(4)
	T2-CK	0.087(6)	0.065(2)	0.066(3)	0.058(1)	0.095(11)	0.094(8)	0.097(12)	0.093(7)	0.094(10)	0.146(14)	0.116(13)	0.080(4)	0.094(9)	0.084(5)
	T2-DCI	0.198(5)	0.257(14)	0.257(13)	0.254(12)	0.197(4)	0.200(7)	0.208(9)	0.207(8)	0.198(6)	0.226(11)	0.220(10)	0.189(2)	0.195(3)	0.186(1)
	T2-PCK	0.109(2)	0.168(9)	0.170(10)	0.188(11)	0.110(4)	0.224(13)	0.228(14)	0.208(12)	0.109(3)	0.139(7)	0.149(8)	0.124(6)	0.113(5)	0.106(1)
	T2-SEP	0.113(5)	0.147(10)	0.145(9)	0.129(7)	0.114(6)	0.095(3)	0.094(2)	0.069(1)	0.112(4)	0.168(13)	0.167(12)	0.162(11)	0.190(14)	0.131(8)
	U41Q1	0.058(1)	0.101(13)	0.098(11)	0.097(10)	0.059(3)	0.071(6)	0.072(7)	0.063(4)	0.059(2)	0.105(14)	0.099(12)	0.083(9)	0.082(8)	0.064(5)
	U41Q2	0.067(6)	0.072(8)	0.070(7)	0.073(9)	0.066(4)	0.061(2)	0.062(3)	0.058(1)	0.066(5)	0.088(13)	0.093(14)	0.079(12)	0.075(11)	0.073(10)
	U42Q1	0.045(5)	0.045(4)	0.044(1)	0.051(7)	0.044(3)	0.067(10)	0.074(13)	0.076(14)	0.044(2)	0.071(12)	0.061(9)	0.068(11)	0.058(8)	0.050(6)
	U42Q2	0.046(1)	0.076(13)	0.076(12)	0.074(10)	0.047(2)	0.056(6)	0.061(8)	0.062(9)	0.047(3)	0.095(14)	0.075(11)	0.049(4)	0.052(5)	0.056(7)

Table 6: Metric value and ranking results of each uncertainty quantification methods with Qwen3-4B over 14 questions.

Control	Question	AUROC CE	Embed_Eigen	Embed_GE	Embed_NAD	FSD	JS_Eigen	JS_GE	JS_NAD	MAR	NLI_DSE	NLI_Eigen	NLI_GE	NLI_NAD	Numset
shot=0, cot=0	ASAP	0.630(1)	0.525(9)	0.516(10)	0.527(8)	0.629(3)	0.572(7)	0.587(6)	0.592(5)	0.629(2)	0.481(13)	0.492(12)	0.446(14)	0.501(11)	0.610(4)
	Semeval	0.579(3)	0.530(10)	0.542(8)	0.569(6)	0.577(5)	0.544(7)	0.537(9)	0.602(1)	0.578(4)	0.491(11)	0.428(12)	0.424(13)	0.413(14)	0.586(2)
	T1-CK	0.608(5)	0.561(10)	0.572(9)	0.616(4)	0.618(2)	0.597(7)	0.601(6)	0.665(1)	0.617(3)	0.440(11)	0.379(12)	0.345(14)	0.369(13)	0.588(8)
	T1-DCI	0.586(5)	0.530(10)	0.537(9)	0.553(8)	0.586(5)	0.663(2)	0.660(3)	0.664(1)	0.586(5)	0.460(11)	0.434(12)	0.394(14)	0.433(13)	0.587(4)
	T1-PCK	0.673(3)	0.596(7)	0.601(6)	0.610(5)	0.674(1)	0.537(10)	0.524(11)	0.554(8)	0.674(2)	0.473(13)	0.548(9)	0.468(14)	0.503(12)	0.666(4)
	T1-SEP	0.684(1)	0.588(10)	0.603(9)	0.657(6)	0.670(4)	0.630(7)	0.626(8)	0.675(3)	0.676(2)	0.441(14)	0.462(11)	0.442(13)	0.458(12)	0.664(5)
	T2-CK	0.713(2)	0.551(7)	0.557(5)	0.551(8)	0.712(3)	0.552(6)	0.537(10)	0.545(9)	0.714(1)	0.485(11)	0.410(12)	0.354(13)	0.302(14)	0.673(4)
	T2-DCI	0.641(6)	0.584(11)	0.577(12)	0.544(13)	0.637(8)	0.666(4)	0.667(3)	0.645(5)	0.639(7)	0.493(14)	0.675(2)	0.602(10)	0.683(1)	0.632(9)
	T2-PCK	0.669(3)	0.505(11)	0.503(12)	0.539(9)	0.669(3)	0.647(6)	0.628(7)	0.683(1)	0.669(3)	0.458(14)	0.578(8)	0.462(13)	0.508(10)	0.673(2)
	T2-SEP	0.647(8)	0.653(7)	0.644(9)	0.626(10)	0.653(5)	0.709(2)	0.708(3)	0.727(1)	0.653(5)	0.516(11)	0.503(12)	0.461(14)	0.490(13)	0.669(4)
	U41Q1	0.371(13)	0.537(3)	0.526(5)	0.505(9)	0.407(11)	0.531(4)	0.558(1)	0.495(10)	0.397(12)	0.514(7)	0.511(8)	0.554(2)	0.520(6)	0.351(14)
	U41Q2	0.499(11)	0.560(2)	0.559(4)	0.552(5)	0.497(13)	0.534(8)	0.539(6)	0.559(3)	0.499(12)	0.524(9)	0.535(7)	0.519(10)	0.579(1)	0.473(14)
	U42Q1	0.570(8)	0.664(1)	0.644(2)	0.618(4)	0.568(10)	0.627(3)	0.617(5)	0.605(6)	0.570(9)	0.515(13)	0.486(14)	0.537(11)	0.517(12)	0.591(7)
	U42Q2	0.414(12)	0.462(9)	0.450(10)	0.434(11)	0.414(14)	0.567(2)	0.574(1)	0.526(3)	0.414(13)	0.500(7)	0.505(6)	0.522(4)	0.516(5)	0.464(8)
shot=0, cot=1	ASAP	0.525(1)	0.456(9)	0.433(12)	0.438(11)	0.517(4)	0.513(6)	0.491(7)	0.522(2)	0.518(3)	0.487(8)	0.419(14)	0.451(10)	0.419(13)	0.517(5)
	Semeval	0.534(9)	0.551(2)	0.557(1)	0.536(7)	0.539(4)	0.540(3)	0.539(5)	0.536(8)	0.537(6)	0.492(12)	0.417(14)	0.493(11)	0.432(13)	0.526(10)
	T1-CK	0.625(2)	0.445(11)	0.439(12)	0.469(9)	0.625(3)	0.534(6)	0.513(7)	0.556(5)	0.627(1)	0.504(8)	0.455(10)	0.427(13)	0.390(14)	0.604(4)
	T1-DCI	0.535(2)	0.517(8)	0.508(9)	0.491(10)	0.535(2)	0.532(6)	0.534(5)	0.531(7)	0.535(2)	0.464(11)	0.420(14)	0.455(12)	0.448(13)	0.536(1)
	T1-PCK	0.654(3)	0.561(7)	0.568(6)	0.577(5)	0.657(1)	0.491(13)	0.490(14)	0.529(10)	0.656(2)	0.495(12)	0.518(11)	0.547(9)	0.556(8)	0.617(4)
	T1-SEP	0.717(2)	0.637(8)	0.652(7)	0.689(5)	0.718(1)	0.538(10)	0.542(9)	0.654(6)	0.716(3)	0.364(12)	0.346(13)	0.337(14)	0.379(11)	0.695(4)
	T2-CK	0.735(1)	0.579(6)	0.574(10)	0.575(8)	0.730(3)	0.574(9)	0.541(11)	0.579(5)	0.732(2)	0.479(12)	0.464(13)	0.576(7)	0.460(14)	0.723(4)
	T2-DCI	0.701(4)	0.650(12)	0.639(13)	0.651(11)	0.694(7)	0.723(3)	0.738(2)	0.746(1)	0.698(6)	0.511(14)	0.688(9)	0.677(10)	0.692(8)	0.700(5)
	T2-PCK	0.688(4)	0.555(13)	0.569(12)	0.578(11)	0.687(6)	0.667(7)	0.640(9)	0.697(1)	0.687(5)	0.500(14)	0.618(10)	0.650(8)	0.696(2)	0.696(3)
	T2-SEP	0.773(3)	0.621(11)	0.624(10)	0.616(12)	0.779(1)	0.722(6)	0.714(7)	0.737(5)	0.777(2)	0.494(14)	0.595(13)	0.673(9)	0.678(8)	0.760(4)
	U41Q1	0.466(9)	0.416(13)	0.420(11)	0.404(14)	0.488(5)	0.485(6)	0.475(8)	0.418(12)	0.482(7)	0.499(3)	0.582(2)	0.492(4)	0.582(1)	0.434(10)
	U41Q2	0.498(11)	0.602(4)	0.608(2)	0.584(5)	0.483(13)	0.509(9)	0.517(8)	0.517(7)	0.490(12)	0.499(10)	0.606(3)	0.555(6)	0.683(1)	0.438(14)
	U42Q1	0.562(10)	0.625(2)	0.622(4)	0.636(1)	0.551(13)	0.608(5)	0.607(6)	0.594(7)	0.555(12)	0.505(14)	0.562(9)	0.580(8)	0.625(3)	0.557(11)
	U42Q2	0.400(12)	0.400(11)	0.402(9)	0.422(6)	0.398(14)	0.401(10)	0.418(7)	0.416(8)	0.400(12)	0.528(4)	0.530(3)	0.576(1)	0.546(2)	0.426(5)
shot=2, cot=1	ASAP	0.675(1)	0.571(5)	0.554(7)	0.551(9)	0.673(2)	0.561(6)	0.539(10)	0.538(11)	0.672(3)	0.478(13)	0.478(14)	0.553(8)	0.484(12)	0.653(4)
	Semeval	0.567(4)	0.612(2)	0.617(1)	0.590(3)	0.566(6)	0.534(8)	0.533(9)	0.517(10)	0.566(5)	0.500(11)	0.381(14)	0.429(12)	0.413(13)	0.555(7)
	T1-CK	0.826(1)	0.549(11)	0.539(13)	0.555(10)	0.802(3)	0.729(6)	0.721(7)	0.735(5)	0.819(2)	0.489(14)	0.546(12)	0.580(9)	0.662(8)	0.788(4)
	T1-DCI	0.558(4)	0.522(10)	0.505(11)	0.461(13)	0.558(4)	0.624(2)	0.635(1)	0.572(3)	0.558(4)	0.452(14)	0.498(12)	0.538(8)	0.532(9)	0.556(7)
	T1-PCK	0.664(4)	0.586(11)	0.583(12)	0.571(13)	0.672(2)	0.623(8)	0.621(9)	0.609(10)	0.670(3)	0.494(14)	0.660(5)	0.639(7)	0.674(1)	0.659(6)
	T1-SEP	0.696(4)	0.626(7)	0.632(5)	0.624(8)	0.697(2)	0.558(10)	0.578(9)	0.630(6)	0.701(1)	0.476(11)	0.392(14)	0.420(12)	0.393(13)	0.697(3)
	T2-CK	0.826(3)	0.495(11)	0.487(13)	0.514(10)	0.880(1)	0.518(9)	0.487(12)	0.534(7)	0.866(2)	0.483(14)	0.518(8)	0.729(5)	0.648(6)	0.779(4)
	T2-DCI	0.617(4)	0.468(13)	0.475(12)	0.467(14)	0.609(6)	0.516(8)	0.505(9)	0.490(11)	0.611(5)	0.500(10)	0.602(7)	0.684(1)	0.655(2)	0.651(3)
	T2-PCK	0.731(2)	0.646(13)	0.652(12)	0.670(8)	0.712(4)	0.666(10)	0.669(9)	0.698(5)	0.717(3)	0.495(14)	0.659(11)	0.693(6)	0.688(7)	0.739(1)
	T2-SEP	0.623(4)	0.522(11)	0.521(12)	0.535(9)	0.629(3)	0.527(10)	0.520(13)	0.555(7)	0.631(2)	0.481(14)	0.542(8)	0.613(5)	0.645(1)	0.605(6)
	U41Q1	0.552(6)	0.495(11)	0.503(9)	0.530(7)	0.565(4)	0.480(12)	0.478(13)	0.461(14)	0.562(5)	0.508(8)	0.610(2)	0.599(3)	0.628(1)	0.499(10)
	U41Q2	0.668(1)	0.611(6)	0.616(5)	0.607(7)	0.652(4)	0.510(12)	0.517(11)	0.508(13)	0.658(2)	0.486(14)	0.595(9)	0.599(8)	0.655(3)	0.588(10)
	U42Q1	0.635(3)	0.618(7)	0.612(8)	0.604(9)	0.628(5)	0.641(1)	0.634(4)	0.600(10)	0.626(6)	0.505(13)	0.559(11)	0.474(14)	0.538(12)	0.637(2)
	U42Q2	0.481(10)	0.471(14)	0.481(9)	0.477(11)	0.476(13)	0.495(6)	0.519(3)	0.514(4)	0.477(12)	0.497(5)	0.487(7)	0.572(1)	0.524(2)	0.484(8)

Control	Question	C-index CE	Embed_Eigen	Embed_GE	Embed_NAD	FSD	JS_Eigen	JS_GE	JS_NAD	MAR	NLI_DSE	NLI_Eigen	NLI_GE	NLI_NAD	Numset
shot=0, cot=0	ASAP	0.616(1)	0.535(8)	0.526(9)	0.521(10)	0.616(3)	0.570(6)	0.581(5)	0.563(7)	0.616(2)	0.479(13)	0.490(12)	0.450(14)	0.504(11)	0.593(4)
	Semeval	0.579(3)	0.530(10)	0.542(8)	0.569(6)	0.577(5)	0.544(7)	0.537(9)	0.602(1)	0.578(4)	0.491(11)	0.428(12)	0.424(13)	0.413(14)	0.586(2)
	T1-CK	0.584(7)	0.560(10)	0.563(8)	0.590(6)	0.595(4)	0.596(3)	0.597(2)	0.651(1)	0.595(5)	0.490(11)	0.439(12)	0.425(14)	0.439(13)	0.562(9)
	T1-DCI	0.586(5)	0.530(10)	0.537(9)	0.553(8)	0.586(5)	0.663(2)	0.660(3)	0.664(1)	0.586(5)	0.460(11)	0.434(12)	0.394(14)	0.433(13)	0.587(4)
	T1-PCK	0.673(3)	0.599(7)	0.604(6)	0.616(5)	0.674(1)	0.544(10)	0.532(11)	0.559(8)	0.673(2)	0.474(13)	0.547(9)	0.470(14)	0.507(12)	0.660(4)
	T1-SEP	0.641(2)	0.559(10)	0.572(9)	0.624(6)	0.629(4)	0.615(7)	0.608(8)	0.657(1)	0.633(3)	0.444(11)	0.441(12)	0.426(14)	0.441(13)	0.627(5)
	T2-CK	0.648(3)	0.502(6)	0.504(5)	0.497(8)	0.651(1)	0.497(7)	0.484(10)	0.490(9)	0.650(2)	0.479(11)	0.405(12)	0.352(13)	0.317(14)	0.611(4)
	T2-DCI	0.577(7)	0.544(11)	0.538(12)	0.509(13)	0.570(10)	0.628(4)	0.631(3)	0.609(5)	0.573(8)	0.490(14)	0.644(2)	0.597(6)	0.669(1)	0.572(9)
	T2-PCK	0.673(3)	0.509(11)	0.508(12)	0.543(9)	0.673(3)	0.646(6)	0.627(7)	0.684(1)	0.673(3)	0.458(14)	0.578(8)	0.465(13)	0.509(10)	0.674(2)
	T2-SEP	0.647(8)	0.653(7)	0.644(9)	0.626(10)	0.653(5)	0.709(2)	0.708(3)	0.727(1)	0.653(5)	0.516(11)	0.503(12)	0.461(14)	0.490(13)	0.669(4)
	U41Q1	0.418(13)	0.564(2)	0.559(4)	0.543(6)	0.444(11)	0.563(3)	0.590(1)	0.527(7)	0.437(12)	0.497(10)	0.512(9)	0.554(5)	0.514(8)	0.409(14)
	U41Q2	0.484(11)	0.534(2)	0.530(4)	0.523(5)	0.479(13)	0.507(10)	0.511(9)	0.532(3)	0.482(12)	0.522(6)	0.519(7)	0.513(8)	0.566(1)	0.460(14)
	U42Q1	0.577(8)	0.639(1)	0.619(2)	0.592(5)	0.575(10)	0.608(3)	0.600(4)	0.591(6)	0.577(9)	0.510(13)	0.486(14)	0.546(11)	0.516(12)	0.587(7)
	U42Q2	0.451(12)	0.471(9)	0.467(10)	0.454(11)	0.450(14)	0.550(2)	0.551(1)	0.526(3)	0.451(13)	0.495(7)	0.503(5)	0.500(6)	0.511(4)	0.489(8)
shot=0, cot=1	ASAP	0.528(1)	0.455(11)	0.436(14)	0.438(13)	0.525(3)	0.491(7)	0.474(9)	0.497(5)	0.525(2)	0.493(6)	0.450(12)	0.485(8)	0.461(10)	0.516(4)
	Semeval	0.534(9)	0.551(2)	0.557(1)	0.536(7)	0.539(4)	0.540(3)	0.539(5)	0.536(8)	0.537(6)	0.492(12)	0.417(14)	0.493(11)	0.432(13)	0.526(10)
	T1-CK	0.557(3)	0.448(9)	0.438(11)	0.441(10)	0.565(1)	0.515(6)	0.501(7)	0.524(5)	0.560(2)	0.497(8)	0.421(12)	0.413(13)	0.368(14)	0.547(4)
	T1-DCI	0.535(2)	0.517(8)	0.508(9)	0.491(10)	0.535(2)	0.532(6)	0.534(5)	0.531(7)	0.535(2)	0.464(11)	0.420(14)	0.455(12)	0.448(13)	0.536(1)
	T1-PCK	0.655(3)	0.563(7)	0.568(6)	0.577(5)	0.659(1)	0.491(14)	0.491(13)	0.528(10)	0.658(2)	0.495(12)	0.514(11)	0.549(9)	0.552(8)	0.619(4)
	T1-SEP	0.660(3)	0.595(8)	0.607(7)	0.637(5)	0.666(1)	0.514(10)	0.515(9)	0.615(6)	0.663(2)	0.379(11)	0.341(14)	0.342(13)	0.362(12)	0.645(4)
	T2-CK	0.632(2)	0.488(8)	0.484(10)	0.496(7)	0.626(4)	0.497(6)	0.471(12)	0.506(5)	0.628(3)	0.484(11)	0.421(13)	0.486(9)	0.406(14)	0.649(1)
	T2-DCI	0.633(7)	0.599(12)	0.590(13)	0.600(11)	0.627(10)	0.690(3)	0.704(2)	0.714(1)	0.631(8)	0.509(14)	0.660(5)	0.647(6)	0.665(4)	0.630(9)
	T2-PCK	0.689(4)	0.558(13)	0.571(12)	0.576(11)	0.688(6)	0.666(7)	0.636(9)	0.691(3)	0.689(5)	0.500(14)	0.618(10)	0.651(8)	0.696(1)	0.695(2)
	T2-SEP	0.719(3)	0.570(11)	0.572(10)	0.560(12)	0.724(1)	0.669(6)	0.660(7)	0.682(5)	0.720(2)	0.494(14)	0.550(13)	0.621(9)	0.628(8)	0.706(4)
	U41Q1	0.483(9)	0.413(14)	0.421(13)	0.428(12)	0.502(6)	0.513(4)	0.507(5)	0.468(10)	0.497(7)	0.491(8)	0.580(2)	0.516(3)	0.605(1)	0.466(11)
	U41Q2	0.509(10)	0.602(3)	0.610(2)	0.596(5)	0.493(13)	0.521(9)	0.531(8)	0.534(7)	0.501(11)	0.498(12)	0.599(4)	0.549(6)	0.669(1)	0.452(14)
	U42Q1	0.579(8)	0.607(3)	0.601(4)	0.610(2)	0.568(12)	0.600(5)	0.597(6)	0.578(9)	0.571(10)	0.501(14)	0.562(13)	0.583(7)	0.618(1)	0.570(11)
	U42Q2	0.442(12)	0.458(11)	0.460(10)	0.478(7)	0.441(14)	0.474(8)	0.486(5)	0.484(6)	0.442(12)	0.514(3)	0.497(4)	0.530(1)	0.520(2)	0.465(9)
shot=2, cot=1	ASAP	0.636(1)	0.545(5)	0.532(8)	0.527(9)	0.632(2)	0.535(7)	0.516(10)	0.515(11)	0.630(3)	0.481(12)	0.472(13)	0.535(6)	0.470(14)	0.615(4)
	Semeval	0.551(4)	0.600(2)	0.604(1)	0.576(3)	0.551(5)	0.537(8)	0.532(9)	0.516(10)	0.550(6)	0.500(11)	0.371(14)	0.415(12)	0.399(13)	0.541(7)
	T1-CK	0.786(1)	0.550(11)	0.538(13)	0.553(10)	0.749(4)	0.708(6)	0.699(7)	0.716(5)	0.770(2)	0.492(14)	0.543(12)	0.584(9)	0.652(8)	0.762(3)
	T1-DCI	0.558(4)	0.522(10)	0.505(11)	0.461(13)	0.558(4)	0.624(2)	0.635(1)	0.572(3)	0.558(4)	0.452(14)	0.498(12)	0.538(8)	0.532(9)	0.556(7)
	T1-PCK	0.641(6)	0.588(11)	0.587(12)	0.576(13)	0.653(4)	0.637(8)	0.636(9)	0.631(10)	0.648(5)	0.494(14)	0.671(2)	0.660(3)	0.688(1)	0.638(7)
	T1-SEP	0.682(3)	0.625(6)	0.628(5)	0.616(7)	0.685(2)	0.534(10)	0.552(9)	0.601(8)	0.687(1)	0.477(11)	0.386(14)	0.415(12)	0.387(13)	0.678(4)
	T2-CK	0.757(3)	0.469(13)	0.460(14)	0.497(9)	0.800(1)	0.513(8)	0.483(11)	0.544(7)	0.788(2)	0.484(10)	0.472(12)	0.660(5)	0.586(6)	0.728(4)
	T2-DCI	0.575(5)	0.465(13)	0.471(12)	0.459(14)	0.569(7)	0.516(8)	0.509(9)	0.485(11)	0.570(6)	0.499(10)	0.621(3)	0.673(1)	0.647(2)	0.607(4)
	T2-PCK	0.744(2)	0.644(13)	0.652(12)	0.674(8)	0.727(4)	0.664(10)	0.666(9)	0.694(5)	0.732(3)	0.495(14)	0.653(11)	0.682(7)	0.689(6)	0.749(1)
	T2-SEP	0.627(4)	0.526(11)	0.525(12)	0.540(9)	0.634(3)	0.532(10)	0.525(12)	0.560(7)	0.636(2)	0.481(14)	0.544(8)	0.613(5)	0.646(1)	0.609(6)
	U41Q1	0.554(6)	0.483(11)	0.488(10)	0.510(7)	0.565(4)	0.477(12)	0.473(13)	0.458(14)	0.562(5)	0.506(8)	0.599(3)	0.607(2)	0.627(1)	0.496(9)
	U41Q2	0.648(2)	0.592(7)	0.600(5)	0.589(8)	0.637(4)	0.493(12)	0.500(11)	0.493(13)	0.642(3)	0.487(14)	0.597(6)	0.582(9)	0.655(1)	0.565(10)
	U42Q1	0.630(3)	0.615(7)	0.610(8)	0.607(9)	0.622(5)	0.634(2)	0.625(4)	0.595(10)	0.622(6)	0.504(13)	0.559(11)	0.476(14)	0.539(12)	0.635(1)
	U42Q2	0.482(9)	0.470(14)	0.478(11)	0.479(10)	0.476(13)	0.510(5)	0.528(2)	0.515(4)	0.477(12)	0.496(6)	0.493(7)	0.578(1)	0.520(3)	0.484(8)

Control	Question	AUARC CE	Embed_Eigen	Embed_GE	Embed_NAD	FSD	JS_Eigen	JS_GE	JS_NAD	MAR	NLI_DSE	NLI_Eigen	NLI_GE	NLI_NAD	Numset
shot=0, cot=0	ASAP	0.598(4)	0.559(9)	0.555(10)	0.565(8)	0.597(6)	0.614(2)	0.620(1)	0.612(3)	0.598(5)	0.554(11)	0.524(13)	0.481(14)	0.530(12)	0.596(7)
	Semeval	0.663(3)	0.634(8)	0.635(7)	0.654(6)	0.662(5)	0.619(10)	0.610(11)	0.669(1)	0.663(4)	0.630(9)	0.511(12)	0.500(13)	0.493(14)	0.667(2)
	T1-CK	0.386(5)	0.338(10)	0.351(9)	0.388(3)	0.389(2)	0.382(6)	0.379(7)	0.421(1)	0.388(4)	0.275(11)	0.231(12)	0.197(14)	0.227(13)	0.377(8)
	T1-DCI	0.714(2)	0.567(11)	0.574(10)	0.602(9)	0.714(2)	0.645(7)	0.651(6)	0.662(5)	0.714(2)	0.639(8)	0.517(14)	0.534(13)	0.536(12)	0.715(1)
	T1-PCK	0.802(5)	0.798(6)	0.803(2)	0.804(1)	0.802(3)	0.725(12)	0.726(11)	0.750(8)	0.802(4)	0.655(14)	0.742(9)	0.696(13)	0.727(10)	0.792(7)
	T1-SEP	0.746(2)	0.678(10)	0.688(9)	0.730(6)	0.743(4)	0.714(8)	0.714(7)	0.754(1)	0.743(3)	0.600(11)	0.578(13)	0.578(12)	0.572(14)	0.737(5)
	T2-CK	0.585(1)	0.444(10)	0.450(9)	0.463(7)	0.585(2)	0.469(6)	0.456(8)	0.477(5)	0.584(3)	0.434(11)	0.335(12)	0.308(13)	0.291(14)	0.559(4)
	T2-DCI	0.592(7)	0.566(12)	0.567(11)	0.546(13)	0.591(9)	0.645(4)	0.649(2)	0.623(5)	0.592(8)	0.507(14)	0.646(3)	0.605(6)	0.650(1)	0.587(10)
	T2-PCK	0.809(3)	0.708(13)	0.710(12)	0.733(9)	0.809(3)	0.804(6)	0.796(7)	0.817(1)	0.809(3)	0.724(11)	0.747(8)	0.704(14)	0.728(10)	0.810(2)
	T2-SEP	0.814(9)	0.854(4)	0.852(5)	0.847(6)	0.814(8)	0.874(3)	0.878(2)	0.884(1)	0.815(7)	0.716(14)	0.768(12)	0.774(11)	0.765(13)	0.813(10)
	U41Q1	0.212(13)	0.291(8)	0.285(9)	0.278(10)	0.222(11)	0.309(4)	0.318(1)	0.292(7)	0.219(12)	0.311(3)	0.296(6)	0.312(2)	0.305(5)	0.200(14)
	U41Q2	0.542(11)	0.590(4)	0.588(5)	0.581(6)	0.542(10)	0.538(13)	0.551(9)	0.560(8)	0.542(12)	0.566(7)	0.628(2)	0.605(3)	0.649(1)	0.516(14)
	U42Q1	0.567(5)	0.572(4)	0.563(8)	0.542(10)	0.566(7)	0.577(1)	0.576(2)	0.553(9)	0.567(6)	0.517(12)	0.487(14)	0.529(11)	0.501(13)	0.573(3)
	U42Q2	0.320(7)	0.313(13)	0.318(10)	0.308(14)	0.320(8)	0.380(2)	0.383(1)	0.347(3)	0.320(6)	0.320(9)	0.315(12)	0.344(4)	0.318(11)	0.333(5)
shot=0, cot=1	ASAP	0.549(4)	0.492(12)	0.485(13)	0.500(10)	0.547(5)	0.543(7)	0.531(8)	0.565(1)	0.547(6)	0.550(3)	0.492(11)	0.515(9)	0.476(14)	0.557(2)
	Semeval	0.591(4)	0.567(7)	0.573(6)	0.564(8)	0.593(2)	0.540(10)	0.537(11)	0.545(9)	0.593(3)	0.632(1)	0.451(14)	0.493(12)	0.462(13)	0.589(5)
	T1-CK	0.375(3)	0.308(10)	0.300(11)	0.310(9)	0.376(1)	0.338(7)	0.323(8)	0.354(5)	0.376(2)	0.353(6)	0.267(12)	0.253(13)	0.222(14)	0.371(4)
	T1-DCI	0.734(2)	0.601(9)	0.599(11)	0.582(13)	0.734(2)	0.595(12)	0.601(10)	0.606(8)	0.734(2)	0.651(5)	0.579(14)	0.616(6)	0.611(7)	0.735(1)
	T1-PCK	0.789(2)	0.771(6)	0.775(4)	0.775(5)	0.790(1)	0.713(12)	0.710(13)	0.726(11)	0.789(3)	0.674(14)	0.733(10)	0.767(7)	0.763(9)	0.767(8)
	T1-SEP	0.681(1)	0.636(8)	0.637(7)	0.660(6)	0.680(2)	0.588(9)	0.586(10)	0.668(5)	0.679(3)	0.501(11)	0.388(14)	0.429(12)	0.425(13)	0.671(4)
	T2-CK	0.565(2)	0.453(5)	0.447(7)	0.449(6)	0.565(3)	0.434(10)	0.418(11)	0.438(9)	0.565(1)	0.358(12)	0.337(13)	0.439(8)	0.325(14)	0.557(4)
	T2-DCI	0.692(8)	0.645(12)	0.642(13)	0.659(11)	0.689(10)	0.716(3)	0.725(2)	0.726(1)	0.689(9)	0.543(14)	0.708(5)	0.701(6)	0.715(4)	0.693(7)
	T2-PCK	0.780(4)	0.697(13)	0.701(12)	0.717(11)	0.780(5)	0.772(8)	0.763(9)	0.795(1)	0.780(5)	0.685(14)	0.736(10)	0.776(7)	0.791(2)	0.780(3)
	T2-SEP	0.784(5)	0.736(10)	0.735(11)	0.718(12)	0.785(4)	0.791(2)	0.780(6)	0.799(1)	0.785(3)	0.544(14)	0.712(13)	0.756(8)	0.745(9)	0.779(7)
	U41Q1	0.206(10)	0.190(14)	0.194(12)	0.196(11)	0.214(7)	0.222(4)	0.217(6)	0.192(13)	0.212(9)	0.257(2)	0.265(1)	0.219(5)	0.241(3)	0.213(8)
	U41Q2	0.433(11)	0.541(6)	0.550(4)	0.550(5)	0.427(13)	0.478(9)	0.482(8)	0.466(10)	0.431(12)	0.498(7)	0.563(3)	0.563(2)	0.617(1)	0.407(14)
	U42Q1	0.485(10)	0.542(1)	0.542(2)	0.541(3)	0.482(12)	0.519(6)	0.513(7)	0.487(9)	0.482(11)	0.421(14)	0.495(8)	0.531(5)	0.534(4)	0.482(13)
	U42Q2	0.247(7)	0.227(13)	0.226(14)	0.237(11)	0.247(9)	0.231(12)	0.239(10)	0.247(6)	0.247(7)	0.369(1)	0.291(4)	0.335(2)	0.309(3)	0.259(5)
shot=2, cot=1	ASAP	0.656(2)	0.626(6)	0.618(7)	0.610(10)	0.656(1)	0.630(5)	0.617(8)	0.616(9)	0.655(3)	0.562(12)	0.548(13)	0.582(11)	0.534(14)	0.649(4)
	Semeval	0.578(7)	0.600(2)	0.596(3)	0.590(4)	0.579(6)	0.573(9)	0.573(10)	0.563(11)	0.577(8)	0.648(1)	0.430(14)	0.443(12)	0.432(13)	0.583(5)
	T1-CK	0.852(1)	0.683(11)	0.673(14)	0.678(13)	0.845(4)	0.812(6)	0.805(7)	0.821(5)	0.850(2)	0.683(12)	0.714(10)	0.734(9)	0.760(8)	0.845(3)
	T1-DCI	0.755(1)	0.682(12)	0.668(13)	0.627(14)	0.755(1)	0.733(6)	0.750(5)	0.711(9)	0.755(1)	0.700(11)	0.702(10)	0.728(7)	0.727(8)	0.753(4)
	T1-PCK	0.766(4)	0.726(10)	0.730(9)	0.732(8)	0.766(6)	0.721(12)	0.713(13)	0.721(11)	0.766(5)	0.640(14)	0.794(2)	0.771(3)	0.798(1)	0.756(7)
	T1-SEP	0.745(4)	0.694(6)	0.695(5)	0.675(8)	0.746(3)	0.635(10)	0.644(9)	0.681(7)	0.748(1)	0.587(11)	0.550(13)	0.558(12)	0.545(14)	0.748(2)
	T2-CK	0.872(3)	0.685(12)	0.675(13)	0.709(8)	0.887(1)	0.704(9)	0.688(11)	0.719(7)	0.883(2)	0.594(14)	0.704(10)	0.813(5)	0.768(6)	0.848(4)
	T2-DCI	0.621(4)	0.491(12)	0.492(10)	0.481(14)	0.619(6)	0.497(9)	0.492(11)	0.487(13)	0.619(5)	0.552(8)	0.611(7)	0.661(1)	0.633(3)	0.637(2)
	T2-PCK	0.851(2)	0.808(11)	0.803(12)	0.799(13)	0.846(4)	0.827(9)	0.825(10)	0.840(6)	0.847(3)	0.777(14)	0.837(7)	0.844(5)	0.834(8)	0.856(1)
	T2-SEP	0.711(4)	0.639(10)	0.634(12)	0.624(13)	0.710(5)	0.649(9)	0.636(11)	0.666(8)	0.712(3)	0.583(14)	0.706(6)	0.732(2)	0.755(1)	0.667(7)
	U41Q1	0.524(6)	0.455(14)	0.467(13)	0.478(11)	0.525(5)	0.486(9)	0.489(8)	0.475(12)	0.526(4)	0.483(10)	0.564(2)	0.540(3)	0.579(1)	0.515(7)
	U41Q2	0.647(4)	0.639(8)	0.650(2)	0.650(3)	0.644(6)	0.556(12)	0.559(11)	0.549(13)	0.644(5)	0.521(14)	0.640(7)	0.636(9)	0.679(1)	0.598(10)
	U42Q1	0.711(3)	0.702(9)	0.703(8)	0.704(7)	0.710(4)	0.707(6)	0.717(2)	0.695(10)	0.709(5)	0.653(13)	0.679(11)	0.613(14)	0.669(12)	0.720(1)
	U42Q2	0.368(14)	0.392(9)	0.394(8)	0.390(10)	0.368(12)	0.405(7)	0.418(2)	0.416(4)	0.368(13)	0.418(3)	0.405(6)	0.475(1)	0.410(5)	0.386(11)

Control	Question	AUERC CE	Embed_Eigen	Embed_GE	Embed_NAD	FSD	JS_Eigen	JS_GE	JS_NAD	MAR	NLI_DSE	NLI_Eigen	NLI_GE	NLI_NAD	Numset
shot=0, cot=0	ASAP	0.160(3)	0.168(8)	0.170(9)	0.177(10)	0.161(5)	0.153(2)	0.150(1)	0.166(7)	0.160(4)	0.186(11)	0.198(13)	0.208(14)	0.190(12)	0.162(6)
	Semeval	0.327(3)	0.356(8)	0.355(7)	0.336(6)	0.328(5)	0.371(10)	0.380(11)	0.321(1)	0.327(4)	0.360(9)	0.479(12)	0.490(13)	0.497(14)	0.323(2)
	T1-CK	0.390(7)	0.407(10)	0.401(9)	0.384(4)	0.388(5)	0.375(2)	0.376(3)	0.352(1)	0.389(6)	0.446(11)	0.468(12)	0.497(14)	0.468(13)	0.399(8)
	T1-DCI	0.276(2)	0.423(11)	0.416(10)	0.388(9)	0.276(2)	0.345(7)	0.339(6)	0.328(5)	0.276(2)	0.351(8)	0.473(14)	0.456(13)	0.454(12)	0.275(1)
	T1-PCK	0.099(4)	0.102(6)	0.100(5)	0.097(1)	0.099(2)	0.136(10)	0.136(9)	0.125(8)	0.099(3)	0.196(14)	0.142(11)	0.162(13)	0.147(12)	0.107(7)
	T1-SEP	0.175(4)	0.215(10)	0.210(9)	0.177(5)	0.177(6)	0.173(2)	0.174(3)	0.148(1)	0.177(7)	0.258(11)	0.285(13)	0.273(12)	0.285(14)	0.179(8)
	T2-CK	0.247(1)	0.330(11)	0.326(9)	0.317(7)	0.247(2)	0.314(6)	0.323(8)	0.314(5)	0.248(3)	0.327(10)	0.374(12)	0.396(13)	0.399(14)	0.265(4)
	T2-DCI	0.341(7)	0.358(12)	0.355(11)	0.373(13)	0.342(9)	0.272(4)	0.267(3)	0.291(6)	0.341(8)	0.374(14)	0.257(2)	0.283(5)	0.248(1)	0.342(10)
	T2-PCK	0.091(3)	0.142(13)	0.140(12)	0.129(9)	0.091(3)	0.095(6)	0.099(7)	0.087(1)	0.091(3)	0.135(11)	0.123(8)	0.145(14)	0.134(10)	0.091(2)
	T2-SEP	0.088(9)	0.068(4)	0.069(5)	0.072(6)	0.088(8)	0.058(3)	0.056(2)	0.053(1)	0.088(7)	0.137(14)	0.111(12)	0.108(11)	0.113(13)	0.088(10)
	U41Q1	0.241(13)	0.212(6)	0.213(7)	0.215(10)	0.238(11)	0.205(3)	0.199(1)	0.211(4)	0.239(12)	0.215(9)	0.213(8)	0.205(2)	0.212(5)	0.243(14)
	U41Q2	0.131(11)	0.117(4)	0.117(5)	0.121(6)	0.132(12)	0.130(10)	0.128(9)	0.125(7)	0.132(13)	0.126(8)	0.107(2)	0.108(3)	0.095(1)	0.139(14)
	U42Q1	0.115(2)	0.123(7)	0.125(8)	0.133(12)	0.116(4)	0.119(6)	0.118(5)	0.126(9)	0.116(3)	0.128(10)	0.145(14)	0.129(11)	0.141(13)	0.113(1)
	U42Q2	0.190(8)	0.191(11)	0.189(6)	0.189(7)	0.190(10)	0.171(1)	0.171(2)	0.178(3)	0.190(9)	0.201(14)	0.194(13)	0.188(5)	0.192(12)	0.187(4)
shot=0, cot=1	ASAP	0.180(3)	0.208(12)	0.212(14)	0.205(11)	0.179(1)	0.204(10)	0.208(13)	0.198(9)	0.180(4)	0.193(7)	0.196(8)	0.179(2)	0.193(6)	0.181(5)
	Semeval	0.399(4)	0.423(7)	0.417(6)	0.426(8)	0.397(2)	0.450(10)	0.453(11)	0.445(9)	0.397(3)	0.358(1)	0.539(14)	0.497(12)	0.528(13)	0.401(5)
	T1-CK	0.377(3)	0.427(9)	0.435(10)	0.439(11)	0.372(1)	0.390(7)	0.397(8)	0.378(4)	0.374(2)	0.386(5)	0.471(13)	0.466(12)	0.501(14)	0.387(6)
	T1-DCI	0.256(2)	0.389(9)	0.391(11)	0.408(13)	0.256(2)	0.395(12)	0.389(10)	0.384(8)	0.256(2)	0.339(5)	0.411(14)	0.374(6)	0.379(7)	0.255(1)
	T1-PCK	0.106(3)	0.122(7)	0.120(5)	0.119(4)	0.105(1)	0.158(12)	0.159(13)	0.149(10)	0.106(2)	0.185(14)	0.155(11)	0.122(8)	0.135(9)	0.121(6)
	T1-SEP	0.216(4)	0.231(7)	0.231(8)	0.216(2)	0.216(3)	0.258(9)	0.260(10)	0.206(1)	0.217(5)	0.316(11)	0.390(14)	0.353(12)	0.368(13)	0.225(6)
	T2-CK	0.271(2)	0.372(7)	0.374(8)	0.358(5)	0.271(1)	0.378(10)	0.386(11)	0.365(6)	0.271(4)	0.390(12)	0.428(13)	0.375(9)	0.439(14)	0.271(3)
	T2-DCI	0.254(8)	0.281(12)	0.284(13)	0.268(11)	0.256(9)	0.197(3)	0.190(1)	0.190(2)	0.256(10)	0.351(14)	0.211(5)	0.215(6)	0.208(4)	0.253(7)
	T2-PCK	0.105(3)	0.147(13)	0.146(12)	0.139(11)	0.105(5)	0.111(8)	0.116(9)	0.102(2)	0.105(5)	0.155(14)	0.128(10)	0.108(7)	0.100(1)	0.105(4)
	T2-SEP	0.104(4)	0.128(10)	0.129(11)	0.138(12)	0.104(3)	0.100(2)	0.106(6)	0.097(1)	0.104(5)	0.210(14)	0.139(13)	0.118(8)	0.123(9)	0.107(7)
	U41Q1	0.249(11)	0.265(14)	0.261(13)	0.255(12)	0.247(6)	0.236(3)	0.239(4)	0.248(10)	0.247(7)	0.247(8)	0.224(1)	0.239(5)	0.231(2)	0.248(9)
	U41Q2	0.164(11)	0.128(6)	0.123(4)	0.122(3)	0.166(13)	0.146(8)	0.143(7)	0.147(9)	0.165(12)	0.153(10)	0.119(2)	0.125(5)	0.103(1)	0.172(14)
	U42Q1	0.138(6)	0.133(3)	0.134(4)	0.136(5)	0.139(9)	0.139(8)	0.141(11)	0.151(13)	0.139(10)	0.158(14)	0.141(12)	0.130(2)	0.129(1)	0.138(7)
	U42Q2	0.211(12)	0.207(10)	0.207(9)	0.204(7)	0.211(14)	0.203(6)	0.201(4)	0.198(3)	0.211(12)	0.194(2)	0.207(8)	0.191(1)	0.201(5)	0.210(11)
shot=2, cot=1	ASAP	0.142(1)	0.151(6)	0.154(7)	0.157(10)	0.142(2)	0.149(5)	0.154(8)	0.157(9)	0.143(3)	0.178(12)	0.179(13)	0.162(11)	0.182(14)	0.144(4)
	Semeval	0.412(9)	0.389(2)	0.393(3)	0.399(4)	0.411(7)	0.409(6)	0.412(8)	0.422(11)	0.413(10)	0.339(1)	0.558(14)	0.546(12)	0.558(13)	0.407(5)
	T1-CK	0.070(1)	0.143(11)	0.149(13)	0.147(12)	0.075(4)	0.085(6)	0.089(7)	0.081(5)	0.072(2)	0.157(14)	0.136(10)	0.123(9)	0.112(8)	0.072(3)
	T1-DCI	0.235(1)	0.308(12)	0.322(13)	0.363(14)	0.235(1)	0.257(6)	0.240(5)	0.279(9)	0.235(1)	0.290(11)	0.288(10)	0.262(7)	0.263(8)	0.237(4)
	T1-PCK	0.131(6)	0.137(12)	0.135(9)	0.136(11)	0.130(4)	0.136(10)	0.140(13)	0.135(8)	0.131(5)	0.189(14)	0.102(2)	0.110(3)	0.101(1)	0.134(7)
	T1-SEP	0.151(3)	0.177(6)	0.177(5)	0.194(7)	0.150(2)	0.248(10)	0.241(9)	0.211(8)	0.149(1)	0.273(11)	0.287(13)	0.279(12)	0.292(14)	0.152(4)
	T2-CK	0.066(3)	0.157(11)	0.162(12)	0.140(8)	0.060(1)	0.140(9)	0.146(10)	0.123(6)	0.062(2)	0.177(14)	0.163(13)	0.105(5)	0.131(7)	0.075(4)
	T2-DCI	0.225(5)	0.294(13)	0.292(11)	0.301(14)	0.226(7)	0.281(9)	0.281(10)	0.293(12)	0.226(6)	0.258(8)	0.199(3)	0.180(1)	0.195(2)	0.217(4)
	T2-PCK	0.070(2)	0.094(11)	0.096(12)	0.098(13)	0.073(4)	0.085(9)	0.086(10)	0.078(5)	0.072(3)	0.123(14)	0.080(7)	0.078(6)	0.081(8)	0.069(1)
	T2-SEP	0.139(4)	0.176(10)	0.178(12)	0.183(13)	0.140(5)	0.170(9)	0.177(11)	0.162(8)	0.139(3)	0.206(14)	0.143(6)	0.130(2)	0.119(1)	0.161(7)
	U41Q1	0.134(6)	0.164(14)	0.162(13)	0.160(12)	0.134(4)	0.155(10)	0.154(9)	0.156(11)	0.134(5)	0.148(8)	0.119(2)	0.121(3)	0.113(1)	0.141(7)
	U41Q2	0.092(4)	0.094(8)	0.090(2)	0.091(3)	0.092(5)	0.119(12)	0.118(11)	0.120(13)	0.093(6)	0.124(14)	0.093(7)	0.100(9)	0.082(1)	0.108(10)
	U42Q1	0.074(3)	0.077(9)	0.077(8)	0.075(5)	0.075(4)	0.076(7)	0.074(2)	0.080(10)	0.075(6)	0.086(12)	0.085(11)	0.102(14)	0.087(13)	0.071(1)
	U42Q2	0.177(12)	0.163(10)	0.163(8)	0.162(7)	0.177(13)	0.153(3)	0.151(2)	0.156(6)	0.177(14)	0.163(9)	0.155(4)	0.135(1)	0.155(5)	0.172(11)

Table 7: Metric value and ranking results of each uncertainty quantification methods with Llama-3.1-8B over 14 questions.

Control	Question	AUROC CE	Embed_Eigen	Embed_GE	Embed_NAD	FSD	JS_Eigen	JS_GE	JS_NAD	MAR	NLI_DSE	NLI_Eigen	NLI_GE	NLI_NAD	Numset
shot=0, cot=0	ASAP	0.683(1)	0.380(14)	0.383(13)	0.399(12)	0.642(5)	0.604(8)	0.594(11)	0.659(3)	0.670(2)	0.603(9)	0.611(7)	0.598(10)	0.622(6)	0.652(4)
	Semeval	0.468(8)	0.453(11)	0.438(12)	0.401(14)	0.483(4)	0.533(1)	0.532(2)	0.486(3)	0.481(5)	0.470(7)	0.480(6)	0.466(9)	0.419(13)	0.459(10)
	T1-CK	0.525(5)	0.519(7)	0.516(10)	0.503(12)	0.517(9)	0.456(14)	0.462(13)	0.513(11)	0.520(6)	0.517(8)	0.612(1)	0.603(2)	0.567(4)	0.568(3)
	T1-DCI	0.597(2)	0.396(13)	0.392(14)	0.434(12)	0.597(2)	0.545(8)	0.517(9)	0.490(10)	0.597(2)	0.486(11)	0.592(5)	0.574(6)	0.631(1)	0.567(7)
	T1-PCK	0.622(1)	0.528(9)	0.515(11)	0.513(12)	0.605(4)	0.600(6)	0.589(7)	0.601(5)	0.612(2)	0.470(13)	0.517(10)	0.466(14)	0.573(8)	0.609(3)
	T1-SEP	0.520(5)	0.507(10)	0.502(12)	0.535(4)	0.509(8)	0.574(2)	0.555(3)	0.587(1)	0.513(7)	0.477(13)	0.453(14)	0.508(9)	0.514(6)	0.505(11)
	T2-CK	0.520(9)	0.525(6)	0.501(11)	0.512(10)	0.523(8)	0.584(1)	0.547(2)	0.494(12)	0.524(7)	0.446(14)	0.543(4)	0.466(13)	0.544(3)	0.538(5)
	T2-DCI	0.766(1)	0.496(12)	0.496(12)	0.482(14)	0.736(3)	0.572(10)	0.558(11)	0.610(9)	0.749(2)	0.649(8)	0.676(6)	0.670(7)	0.695(5)	0.703(4)
	T2-PCK	0.769(1)	0.665(10)	0.663(11)	0.681(9)	0.766(3)	0.608(14)	0.609(13)	0.644(12)	0.768(2)	0.694(8)	0.759(4)	0.732(7)	0.753(5)	0.744(6)
	T2-SEP	0.677(1)	0.452(13)	0.460(12)	0.449(14)	0.662(3)	0.491(10)	0.515(8)	0.527(7)	0.666(2)	0.549(6)	0.487(11)	0.491(9)	0.578(5)	0.589(4)
	U41Q1	0.576(8)	0.663(4)	0.642(5)	0.669(2)	0.581(7)	0.636(6)	0.665(3)	0.743(1)	0.571(9)	0.544(10)	0.438(13)	0.424(14)	0.493(11)	0.486(12)
	U41Q2	0.604(5)	0.598(7)	0.625(3)	0.635(2)	0.580(12)	0.625(3)	0.636(1)	0.601(6)	0.596(8)	0.516(14)	0.546(13)	0.580(11)	0.586(10)	0.593(9)
	U42Q1	0.452(14)	0.611(6)	0.641(4)	0.637(5)	0.472(10)	0.647(3)	0.662(2)	0.682(1)	0.469(11)	0.492(9)	0.511(7)	0.495(8)	0.466(12)	0.464(13)
	U42Q2	0.621(3)	0.469(14)	0.487(11)	0.525(7)	0.638(2)	0.487(10)	0.477(13)	0.479(12)	0.641(1)	0.549(6)	0.555(5)	0.520(8)	0.568(4)	0.518(9)
shot=0, cot=1	ASAP	0.719(1)	0.362(13)	0.361(14)	0.383(12)	0.674(4)	0.435(10)	0.434(11)	0.516(9)	0.699(2)	0.558(6)	0.528(8)	0.568(5)	0.552(7)	0.678(3)
	Semeval	0.577(4)	0.501(11)	0.489(12)	0.451(14)	0.542(9)	0.591(1)	0.591(2)	0.585(3)	0.556(7)	0.553(8)	0.525(10)	0.560(6)	0.478(13)	0.565(5)
	T1-CK	0.415(12)	0.550(2)	0.527(4)	0.500(5)	0.385(14)	0.538(3)	0.496(6)	0.476(10)	0.392(13)	0.486(8)	0.496(7)	0.457(11)	0.483(9)	0.556(1)
	T1-DCI	0.535(5)	0.455(13)	0.431(14)	0.463(12)	0.532(8)	0.562(2)	0.564(1)	0.533(6)	0.532(7)	0.532(9)	0.552(4)	0.517(11)	0.556(3)	0.529(10)
	T1-PCK	0.646(1)	0.517(12)	0.505(13)	0.490(14)	0.621(5)	0.627(4)	0.597(7)	0.553(9)	0.629(3)	0.561(8)	0.550(10)	0.547(11)	0.598(6)	0.631(2)
	T1-SEP	0.478(9)	0.437(14)	0.438(12)	0.438(13)	0.473(10)	0.502(2)	0.490(5)	0.485(6)	0.472(11)	0.480(7)	0.505(1)	0.493(4)	0.479(8)	0.494(3)
	T2-CK	0.559(4)	0.451(12)	0.443(14)	0.448(13)	0.553(5)	0.469(9)	0.465(10)	0.459(11)	0.553(6)	0.527(8)	0.588(3)	0.552(7)	0.597(1)	0.597(2)
	T2-DCI	0.767(1)	0.556(14)	0.570(11)	0.568(12)	0.739(4)	0.565(13)	0.580(9)	0.576(10)	0.754(2)	0.648(8)	0.704(6)	0.676(7)	0.748(3)	0.717(5)
	T2-PCK	0.713(5)	0.529(11)	0.551(10)	0.552(9)	0.708(7)	0.445(14)	0.451(13)	0.529(12)	0.710(6)	0.667(8)	0.732(1)	0.717(2)	0.713(4)	0.714(3)
	T2-SEP	0.741(1)	0.511(12)	0.510(13)	0.482(14)	0.741(2)	0.588(8)	0.576(10)	0.568(11)	0.739(3)	0.598(6)	0.577(9)	0.614(5)	0.698(4)	0.590(7)
	U41Q1	0.468(9)	0.574(4)	0.568(5)	0.577(2)	0.518(7)	0.577(2)	0.612(1)	0.567(6)	0.516(8)	0.372(14)	0.394(13)	0.428(10)	0.411(12)	0.421(11)
	U41Q2	0.577(8)	0.552(13)	0.582(7)	0.633(1)	0.548(14)	0.554(12)	0.596(6)	0.620(4)	0.563(11)	0.576(9)	0.626(3)	0.633(2)	0.613(5)	0.566(10)
	U42Q1	0.474(13)	0.602(6)	0.602(5)	0.600(7)	0.531(10)	0.620(4)	0.636(1)	0.629(2)	0.519(11)	0.480(12)	0.599(8)	0.553(9)	0.621(3)	0.421(14)
	U42Q2	0.566(5)	0.485(10)	0.480(12)	0.480(11)	0.673(1)	0.509(8)	0.471(14)	0.487(9)	0.661(2)	0.544(6)	0.541(7)	0.573(4)	0.576(3)	0.476(13)
shot=2, cot=1	ASAP	0.737(2)	0.349(14)	0.352(13)	0.386(12)	0.720(3)	0.419(10)	0.402(11)	0.495(9)	0.747(1)	0.610(7)	0.590(8)	0.616(6)	0.636(4)	0.632(5)
	Semeval	0.599(2)	0.497(13)	0.497(12)	0.476(14)	0.550(6)	0.548(7)	0.540(8)	0.551(5)	0.571(3)	0.531(9)	0.565(4)	0.524(10)	0.502(11)	0.641(1)
	T1-CK	0.791(2)	0.624(11)	0.635(9)	0.699(6)	0.779(3)	0.620(12)	0.639(8)	0.688(7)	0.798(1)	0.616(13)	0.590(14)	0.634(10)	0.778(4)	0.746(5)
	T1-DCI	0.569(3)	0.482(13)	0.492(11)	0.501(10)	0.564(5)	0.422(14)	0.487(12)	0.525(8)	0.564(5)	0.517(9)	0.592(2)	0.568(4)	0.600(1)	0.553(7)
	T1-PCK	0.651(3)	0.524(12)	0.522(13)	0.504(14)	0.702(1)	0.553(9)	0.529(11)	0.549(10)	0.689(2)	0.588(6)	0.569(7)	0.606(5)	0.636(4)	0.564(8)
	T1-SEP	0.628(1)	0.505(10)	0.510(9)	0.522(8)	0.620(3)	0.544(7)	0.560(4)	0.558(5)	0.626(2)	0.453(11)	0.347(14)	0.379(12)	0.360(13)	0.554(6)
	T2-CK	0.657(2)	0.577(7)	0.583(6)	0.546(10)	0.636(3)	0.536(12)	0.500(14)	0.502(13)	0.663(1)	0.550(9)	0.540(11)	0.564(8)	0.610(4)	0.601(5)
	T2-DCI	0.734(1)	0.525(10)	0.526(9)	0.516(12)	0.695(4)	0.513(13)	0.522(11)	0.587(7)	0.707(3)	0.448(14)	0.625(6)	0.563(8)	0.653(5)	0.726(2)
	T2-PCK	0.733(4)	0.544(14)	0.559(13)	0.594(12)	0.731(5)	0.632(10)	0.626(11)	0.633(9)	0.731(6)	0.689(8)	0.751(1)	0.751(2)	0.745(3)	0.715(7)
	T2-SEP	0.597(2)	0.463(12)	0.450(13)	0.431(14)	0.536(10)	0.579(3)	0.560(6)	0.556(8)	0.546(9)	0.530(11)	0.577(4)	0.569(5)	0.557(7)	0.628(1)
	U41Q1	0.527(4)	0.531(3)	0.518(6)	0.517(9)	0.561(2)	0.462(13)	0.483(11)	0.525(5)	0.562(1)	0.461(14)	0.508(10)	0.482(12)	0.518(6)	0.518(6)
	U41Q2	0.562(5)	0.560(6)	0.540(13)	0.522(14)	0.558(7)	0.550(10)	0.551(9)	0.599(1)	0.557(8)	0.544(11)	0.587(2)	0.542(12)	0.582(3)	0.580(4)
	U42Q1	0.602(10)	0.645(6)	0.664(4)	0.661(5)	0.609(9)	0.584(12)	0.597(11)	0.569(13)	0.615(7)	0.610(8)	0.717(1)	0.703(2)	0.677(3)	0.559(14)
	U42Q2	0.562(3)	0.400(14)	0.407(13)	0.422(12)	0.618(2)	0.453(9)	0.443(10)	0.479(7)	0.619(1)	0.510(4)	0.469(8)	0.505(6)	0.439(11)	0.506(5)

Control	Question	C-index CE	Embed_Eigen	Embed_GE	Embed_NAD	FSD	JS_Eigen	JS_GE	JS_NAD	MAR	NLI_DSE	NLI_Eigen	NLI_GE	NLI_NAD	Numset
shot=0, cot=0	ASAP	0.591(3)	0.441(12)	0.431(13)	0.413(14)	0.559(11)	0.576(6)	0.566(10)	0.604(1)	0.576(5)	0.575(7)	0.574(8)	0.576(4)	0.574(9)	0.596(2)
	Semeval	0.468(8)	0.453(11)	0.438(12)	0.401(14)	0.483(4)	0.533(1)	0.532(2)	0.486(3)	0.481(5)	0.470(7)	0.480(6)	0.466(9)	0.419(13)	0.459(10)
	T1-CK	0.529(5)	0.529(6)	0.526(8)	0.514(11)	0.523(10)	0.467(14)	0.470(13)	0.524(9)	0.526(7)	0.505(12)	0.597(1)	0.594(2)	0.564(4)	0.572(3)
	T1-DCI	0.597(2)	0.396(13)	0.392(14)	0.434(12)	0.597(2)	0.545(8)	0.517(9)	0.490(10)	0.597(2)	0.486(11)	0.592(5)	0.574(6)	0.631(1)	0.567(7)
	T1-PCK	0.591(1)	0.523(9)	0.511(10)	0.505(12)	0.577(4)	0.577(5)	0.567(7)	0.571(6)	0.583(2)	0.461(13)	0.507(11)	0.459(14)	0.553(8)	0.578(3)
	T1-SEP	0.531(5)	0.530(6)	0.528(8)	0.563(2)	0.526(10)	0.560(3)	0.537(4)	0.577(1)	0.529(7)	0.462(13)	0.459(14)	0.493(12)	0.526(9)	0.502(11)
	T2-CK	0.541(7)	0.543(5)	0.523(11)	0.533(10)	0.545(4)	0.571(1)	0.541(8)	0.497(12)	0.546(3)	0.443(14)	0.542(6)	0.471(13)	0.550(2)	0.541(9)
	T2-DCI	0.609(2)	0.456(12)	0.446(13)	0.419(14)	0.598(3)	0.501(11)	0.502(10)	0.522(9)	0.597(4)	0.629(1)	0.597(5)	0.588(6)	0.586(8)	0.588(6)
	T2-PCK	0.748(2)	0.650(10)	0.650(11)	0.663(9)	0.746(4)	0.589(14)	0.591(13)	0.623(12)	0.748(2)	0.677(8)	0.756(1)	0.721(7)	0.733(5)	0.730(6)
	T2-SEP	0.677(1)	0.452(13)	0.460(12)	0.449(14)	0.662(3)	0.491(10)	0.515(8)	0.527(7)	0.666(2)	0.549(6)	0.487(11)	0.491(9)	0.578(5)	0.589(4)
	U41Q1	0.467(13)	0.623(5)	0.619(6)	0.642(4)	0.487(11)	0.647(3)	0.656(2)	0.700(1)	0.477(12)	0.572(7)	0.523(10)	0.548(9)	0.554(8)	0.457(14)
	U41Q2	0.468(13)	0.613(3)	0.630(1)	0.620(2)	0.474(12)	0.588(6)	0.599(4)	0.569(8)	0.475(11)	0.549(10)	0.562(9)	0.589(5)	0.578(7)	0.467(14)
	U42Q1	0.463(14)	0.619(6)	0.645(3)	0.628(5)	0.485(11)	0.634(4)	0.649(2)	0.666(1)	0.481(12)	0.498(9)	0.509(8)	0.514(7)	0.490(10)	0.465(13)
	U42Q2	0.578(3)	0.447(14)	0.461(10)	0.493(8)	0.601(1)	0.459(11)	0.450(13)	0.454(12)	0.599(2)	0.532(6)	0.539(5)	0.496(7)	0.544(4)	0.483(9)
shot=0, cot=1	ASAP	0.598(2)	0.385(14)	0.388(13)	0.402(12)	0.559(4)	0.427(11)	0.442(10)	0.508(7)	0.570(3)	0.558(5)	0.490(9)	0.532(6)	0.499(8)	0.599(1)
	Semeval	0.524(7)	0.463(11)	0.449(12)	0.409(14)	0.494(10)	0.563(2)	0.558(3)	0.553(4)	0.504(8)	0.572(1)	0.498(9)	0.541(5)	0.443(13)	0.529(6)
	T1-CK	0.453(12)	0.586(1)	0.570(2)	0.550(5)	0.433(14)	0.561(3)	0.528(6)	0.518(7)	0.438(13)	0.484(10)	0.510(8)	0.477(11)	0.509(9)	0.555(4)
	T1-DCI	0.535(5)	0.455(13)	0.431(14)	0.463(12)	0.532(8)	0.562(2)	0.564(1)	0.533(6)	0.532(7)	0.532(9)	0.552(4)	0.517(11)	0.556(3)	0.529(10)
	T1-PCK	0.620(2)	0.513(12)	0.498(13)	0.475(14)	0.600(5)	0.621(1)	0.595(6)	0.530(11)	0.605(4)	0.557(8)	0.550(10)	0.553(9)	0.583(7)	0.610(3)
	T1-SEP	0.515(2)	0.454(14)	0.463(13)	0.477(12)	0.514(3)	0.500(6)	0.491(9)	0.484(10)	0.513(4)	0.480(11)	0.516(1)	0.500(7)	0.502(5)	0.497(8)
	T2-CK	0.576(4)	0.505(10)	0.503(12)	0.508(9)	0.575(5)	0.483(14)	0.487(13)	0.505(11)	0.573(6)	0.525(8)	0.604(1)	0.562(7)	0.595(2)	0.586(3)
	T2-DCI	0.624(5)	0.502(14)	0.517(12)	0.527(11)	0.624(3)	0.530(10)	0.540(9)	0.505(13)	0.624(6)	0.582(8)	0.634(2)	0.624(4)	0.645(1)	0.615(7)
	T2-PCK	0.712(5)	0.532(11)	0.553(9)	0.553(10)	0.707(7)	0.447(14)	0.453(13)	0.531(12)	0.709(6)	0.667(8)	0.735(1)	0.720(2)	0.712(3)	0.712(4)
	T2-SEP	0.741(1)	0.511(12)	0.510(13)	0.482(14)	0.741(2)	0.588(8)	0.576(10)	0.568(11)	0.739(3)	0.598(6)	0.577(9)	0.614(5)	0.698(4)	0.590(7)
	U41Q1	0.443(9)	0.540(5)	0.530(6)	0.548(4)	0.432(11)	0.611(3)	0.628(1)	0.614(2)	0.436(10)	0.472(8)	0.420(12)	0.420(13)	0.410(14)	0.474(7)
	U41Q2	0.510(13)	0.588(9)	0.599(5)	0.639(2)	0.514(11)	0.586(10)	0.624(3)	0.657(1)	0.511(12)	0.589(8)	0.596(6)	0.621(4)	0.594(7)	0.506(14)
	U42Q1	0.480(13)	0.601(5)	0.600(7)	0.600(6)	0.535(10)	0.612(4)	0.628(1)	0.623(2)	0.524(11)	0.494(12)	0.595(8)	0.549(9)	0.617(3)	0.431(14)
	U42Q2	0.562(5)	0.489(9)	0.484(11)	0.483(12)	0.672(1)	0.511(8)	0.475(13)	0.488(10)	0.659(2)	0.539(6)	0.538(7)	0.564(4)	0.575(3)	0.472(14)
shot=2, cot=1	ASAP	0.611(3)	0.355(13)	0.355(14)	0.392(10)	0.620(2)	0.386(11)	0.379(12)	0.453(9)	0.622(1)	0.570(4)	0.536(8)	0.567(5)	0.562(6)	0.548(7)
	Semeval	0.508(5)	0.505(6)	0.503(7)	0.478(12)	0.474(13)	0.528(3)	0.518(4)	0.497(8)	0.481(11)	0.495(9)	0.544(2)	0.494(10)	0.456(14)	0.588(1)
	T1-CK	0.723(1)	0.607(10)	0.602(11)	0.649(7)	0.677(5)	0.614(9)	0.631(8)	0.665(6)	0.705(3)	0.588(13)	0.556(14)	0.591(12)	0.700(4)	0.706(2)
	T1-DCI	0.569(3)	0.482(13)	0.492(11)	0.501(10)	0.564(5)	0.422(14)	0.487(12)	0.525(8)	0.564(5)	0.517(9)	0.592(2)	0.568(4)	0.600(1)	0.553(7)
	T1-PCK	0.622(4)	0.500(12)	0.496(13)	0.485(14)	0.652(1)	0.520(10)	0.517(11)	0.558(9)	0.636(2)	0.595(6)	0.567(8)	0.626(3)	0.605(5)	0.569(7)
	T1-SEP	0.581(3)	0.505(10)	0.511(9)	0.523(5)	0.583(2)	0.518(7)	0.531(4)	0.522(6)	0.586(1)	0.458(11)	0.367(13)	0.385(12)	0.367(14)	0.517(8)
	T2-CK	0.619(1)	0.599(3)	0.597(4)	0.555(8)	0.588(5)	0.548(9)	0.512(13)	0.490(14)	0.614(2)	0.540(11)	0.513(12)	0.542(10)	0.570(7)	0.582(6)
	T2-DCI	0.659(1)	0.529(9)	0.525(10)	0.499(12)	0.636(4)	0.495(13)	0.508(11)	0.555(7)	0.642(3)	0.450(14)	0.586(6)	0.548(8)	0.609(5)	0.656(2)
	T2-PCK	0.720(1)	0.524(14)	0.542(13)	0.593(12)	0.708(6)	0.633(10)	0.620(11)	0.647(9)	0.713(3)	0.660(8)	0.717(2)	0.711(4)	0.710(5)	0.699(7)
	T2-SEP	0.583(2)	0.453(12)	0.443(13)	0.432(14)	0.527(10)	0.565(3)	0.550(8)	0.553(6)	0.537(9)	0.523(11)	0.565(3)	0.559(5)	0.551(7)	0.609(1)
	U41Q1	0.527(7)	0.540(4)	0.530(5)	0.527(8)	0.558(2)	0.504(12)	0.519(9)	0.550(3)	0.563(1)	0.465(14)	0.515(10)	0.481(13)	0.528(6)	0.511(11)
	U41Q2	0.498(12)	0.543(6)	0.523(7)	0.499(11)	0.494(13)	0.520(8)	0.519(10)	0.571(3)	0.493(14)	0.568(4)	0.607(1)	0.561(5)	0.592(2)	0.519(9)
	U42Q1	0.591(9)	0.630(6)	0.647(4)	0.644(5)	0.596(8)	0.577(13)	0.590(10)	0.583(11)	0.600(7)	0.580(12)	0.696(1)	0.676(2)	0.661(3)	0.555(14)
	U42Q2	0.530(3)	0.399(14)	0.408(13)	0.419(12)	0.583(2)	0.440(9)	0.429(11)	0.468(8)	0.585(1)	0.517(5)	0.484(7)	0.520(4)	0.437(10)	0.487(6)

Control	Question	AUARC CE	Embed_Eigen	Embed_GE	Embed_NAD	FSD	JS_Eigen	JS_GE	JS_NAD	MAR	NLI_DSE	NLI_Eigen	NLI_GE	NLI_NAD	Numset
shot=0, cot=0	ASAP	0.573(1)	0.393(14)	0.398(12)	0.394(13)	0.555(5)	0.516(8)	0.503(11)	0.561(4)	0.562(3)	0.517(7)	0.523(6)	0.507(10)	0.515(9)	0.564(2)
	Semeval	0.356(6)	0.309(12)	0.298(13)	0.269(14)	0.359(4)	0.411(2)	0.417(1)	0.352(7)	0.360(3)	0.357(5)	0.346(8)	0.328(10)	0.309(11)	0.343(9)
	T1-CK	0.435(6)	0.424(9)	0.420(10)	0.408(11)	0.429(8)	0.374(14)	0.398(12)	0.436(5)	0.434(7)	0.384(13)	0.476(1)	0.466(2)	0.461(3)	0.443(4)
	T1-DCI	0.709(1)	0.498(13)	0.487(14)	0.505(12)	0.709(1)	0.597(9)	0.581(10)	0.544(11)	0.709(1)	0.628(7)	0.636(6)	0.626(8)	0.643(5)	0.701(4)
	T1-PCK	0.602(4)	0.564(10)	0.557(11)	0.569(9)	0.597(7)	0.631(1)	0.618(3)	0.619(2)	0.597(6)	0.509(13)	0.557(12)	0.503(14)	0.570(8)	0.602(5)
	T1-SEP	0.361(9)	0.412(5)	0.416(4)	0.431(2)	0.358(11)	0.425(3)	0.412(6)	0.445(1)	0.360(10)	0.345(13)	0.352(12)	0.397(7)	0.388(8)	0.321(14)
	T2-CK	0.275(12)	0.303(6)	0.290(9)	0.306(3)	0.276(11)	0.355(1)	0.326(2)	0.303(4)	0.277(10)	0.221(14)	0.303(5)	0.252(13)	0.296(8)	0.298(7)
	T2-DCI	0.653(1)	0.446(12)	0.442(13)	0.441(14)	0.643(3)	0.525(10)	0.497(11)	0.533(8)	0.649(2)	0.531(9)	0.574(5)	0.556(7)	0.564(6)	0.601(4)
	T2-PCK	0.834(1)	0.767(9)	0.762(11)	0.781(7)	0.831(3)	0.751(13)	0.738(14)	0.759(12)	0.831(2)	0.780(8)	0.812(5)	0.766(10)	0.811(6)	0.824(4)
	T2-SEP	0.839(1)	0.716(14)	0.719(13)	0.730(12)	0.833(3)	0.768(10)	0.780(7)	0.774(8)	0.834(2)	0.758(11)	0.786(6)	0.770(9)	0.810(5)	0.823(4)
	U41Q1	0.072(10)	0.090(6)	0.087(7)	0.109(2)	0.073(9)	0.106(4)	0.109(3)	0.118(1)	0.071(11)	0.096(5)	0.051(13)	0.045(14)	0.059(12)	0.080(8)
	U41Q2	0.287(7)	0.285(8)	0.289(6)	0.309(2)	0.276(10)	0.298(4)	0.297(5)	0.262(14)	0.273(11)	0.270(12)	0.262(13)	0.304(3)	0.280(9)	0.319(1)
	U42Q1	0.363(13)	0.492(6)	0.510(3)	0.494(5)	0.367(11)	0.498(4)	0.515(1)	0.514(2)	0.365(12)	0.382(8)	0.382(7)	0.371(10)	0.352(14)	0.373(9)
	U42Q2	0.669(3)	0.541(13)	0.565(12)	0.582(6)	0.672(2)	0.565(11)	0.566(10)	0.572(8)	0.675(1)	0.569(9)	0.619(5)	0.580(7)	0.631(4)	0.530(14)
shot=0, cot=1	ASAP	0.597(1)	0.337(14)	0.342(13)	0.347(12)	0.574(4)	0.423(10)	0.406(11)	0.461(9)	0.586(3)	0.495(5)	0.463(8)	0.492(6)	0.486(7)	0.590(2)
	Semeval	0.320(7)	0.296(11)	0.287(13)	0.262(14)	0.311(10)	0.360(2)	0.364(1)	0.354(3)	0.315(8)	0.336(4)	0.323(6)	0.330(5)	0.289(12)	0.315(9)
	T1-CK	0.297(11)	0.399(1)	0.382(2)	0.380(3)	0.286(14)	0.373(4)	0.343(6)	0.326(8)	0.287(13)	0.340(7)	0.321(9)	0.288(12)	0.317(10)	0.362(5)
	T1-DCI	0.722(1)	0.602(12)	0.583(14)	0.591(13)	0.722(3)	0.657(10)	0.661(9)	0.652(11)	0.721(4)	0.708(6)	0.707(7)	0.692(8)	0.711(5)	0.722(2)
	T1-PCK	0.628(2)	0.535(12)	0.526(13)	0.523(14)	0.620(5)	0.638(1)	0.621(4)	0.602(8)	0.626(3)	0.587(10)	0.591(9)	0.563(11)	0.611(7)	0.617(6)
	T1-SEP	0.302(11)	0.286(12)	0.283(13)	0.278(14)	0.308(9)	0.324(6)	0.324(7)	0.336(5)	0.304(10)	0.353(1)	0.347(2)	0.338(4)	0.345(3)	0.311(8)
	T2-CK	0.307(4)	0.277(10)	0.275(11)	0.278(9)	0.305(6)	0.269(13)	0.262(14)	0.272(12)	0.305(5)	0.280(8)	0.352(2)	0.302(7)	0.353(1)	0.317(3)
	T2-DCI	0.708(1)	0.527(14)	0.538(13)	0.548(12)	0.697(3)	0.591(10)	0.593(9)	0.586(11)	0.704(2)	0.598(8)	0.668(6)	0.629(7)	0.678(5)	0.678(4)
	T2-PCK	0.809(1)	0.693(11)	0.700(9)	0.694(10)	0.807(2)	0.636(14)	0.640(13)	0.688(12)	0.807(3)	0.759(8)	0.787(5)	0.778(7)	0.786(6)	0.802(4)
	T2-SEP	0.821(1)	0.676(14)	0.678(13)	0.679(12)	0.819(2)	0.726(8)	0.709(10)	0.720(9)	0.818(3)	0.681(11)	0.755(6)	0.766(5)	0.810(4)	0.739(7)
	U41Q1	0.045(11)	0.071(2)	0.071(3)	0.077(1)	0.057(8)	0.059(7)	0.065(5)	0.068(4)	0.049(9)	0.045(10)	0.041(14)	0.041(13)	0.043(12)	0.060(6)
	U41Q2	0.251(12)	0.266(8)	0.287(7)	0.345(1)	0.249(14)	0.290(6)	0.308(4)	0.316(3)	0.250(13)	0.253(11)	0.318(2)	0.299(5)	0.265(9)	0.261(10)
	U42Q1	0.368(13)	0.417(8)	0.425(7)	0.430(6)	0.397(10)	0.480(2)	0.491(1)	0.469(3)	0.384(11)	0.379(12)	0.435(5)	0.415(9)	0.458(4)	0.346(14)
	U42Q2	0.731(3)	0.644(10)	0.636(12)	0.633(13)	0.783(1)	0.688(7)	0.646(9)	0.639(11)	0.775(2)	0.657(8)	0.698(6)	0.721(5)	0.727(4)	0.633(14)
shot=2, cot=1	ASAP	0.620(3)	0.355(13)	0.352(14)	0.371(12)	0.620(2)	0.420(10)	0.398(11)	0.443(9)	0.630(1)	0.554(6)	0.537(8)	0.553(7)	0.570(5)	0.595(4)
	Semeval	0.318(2)	0.255(11)	0.251(12)	0.231(14)	0.297(6)	0.302(5)	0.295(7)	0.318(3)	0.309(4)	0.257(10)	0.279(8)	0.265(9)	0.250(13)	0.323(1)
	T1-CK	0.833(1)	0.771(7)	0.771(8)	0.789(6)	0.823(4)	0.750(11)	0.738(13)	0.766(9)	0.833(2)	0.727(14)	0.745(12)	0.762(10)	0.831(3)	0.814(5)
	T1-DCI	0.720(4)	0.648(12)	0.660(10)	0.674(9)	0.719(6)	0.613(14)	0.632(13)	0.655(11)	0.720(5)	0.688(8)	0.740(1)	0.731(2)	0.727(3)	0.708(7)
	T1-PCK	0.737(3)	0.589(13)	0.582(14)	0.590(12)	0.766(1)	0.650(9)	0.634(10)	0.630(11)	0.753(2)	0.662(6)	0.661(7)	0.675(5)	0.694(4)	0.660(8)
	T1-SEP	0.457(2)	0.391(8)	0.389(9)	0.401(7)	0.450(3)	0.403(6)	0.417(4)	0.408(5)	0.458(1)	0.340(11)	0.265(14)	0.315(12)	0.297(13)	0.373(10)
	T2-CK	0.601(3)	0.577(5)	0.584(4)	0.556(8)	0.606(2)	0.494(12)	0.477(14)	0.490(13)	0.614(1)	0.503(11)	0.508(10)	0.527(9)	0.568(7)	0.569(6)
	T2-DCI	0.701(1)	0.510(12)	0.517(11)	0.521(10)	0.685(4)	0.498(14)	0.498(13)	0.571(8)	0.691(3)	0.560(9)	0.610(6)	0.587(7)	0.638(5)	0.700(2)
	T2-PCK	0.790(4)	0.632(14)	0.647(13)	0.683(12)	0.785(6)	0.707(8)	0.700(10)	0.699(11)	0.787(5)	0.701(9)	0.802(1)	0.800(3)	0.800(2)	0.770(7)
	T2-SEP	0.668(1)	0.593(11)	0.584(12)	0.580(13)	0.640(8)	0.641(7)	0.630(9)	0.620(10)	0.645(6)	0.565(14)	0.667(2)	0.649(5)	0.651(4)	0.659(3)
	U41Q1	0.449(5)	0.462(3)	0.449(6)	0.440(8)	0.463(2)	0.395(14)	0.407(13)	0.443(7)	0.461(4)	0.440(9)	0.434(11)	0.425(12)	0.438(10)	0.476(1)
	U41Q2	0.446(5)	0.427(10)	0.413(11)	0.411(12)	0.441(6)	0.434(9)	0.435(8)	0.467(2)	0.439(7)	0.393(14)	0.481(1)	0.402(13)	0.454(3)	0.448(4)
	U42Q1	0.578(9)	0.601(6)	0.609(4)	0.602(5)	0.579(8)	0.572(12)	0.578(10)	0.558(13)	0.583(7)	0.552(14)	0.659(1)	0.655(2)	0.640(3)	0.573(11)
	U42Q2	0.733(3)	0.630(14)	0.636(13)	0.652(12)	0.747(2)	0.670(8)	0.665(9)	0.694(5)	0.748(1)	0.655(11)	0.680(7)	0.696(4)	0.657(10)	0.683(6)

Control	Question	AUERC CE	Embed_Eigen	Embed_GE	Embed_NAD	FSD	JS_Eigen	JS_GE	JS_NAD	MAR	NLI_DSE	NLI_Eigen	NLI_GE	NLI_NAD	Numset
shot=0, cot=0	ASAP	0.177(2)	0.211(12)	0.214(13)	0.228(14)	0.187(6)	0.186(5)	0.192(10)	0.175(1)	0.184(4)	0.188(7)	0.191(9)	0.191(8)	0.193(11)	0.179(3)
	Semeval	0.634(6)	0.681(12)	0.692(13)	0.721(14)	0.631(4)	0.579(2)	0.573(1)	0.638(7)	0.630(3)	0.633(5)	0.644(8)	0.662(10)	0.681(11)	0.647(9)
	T1-CK	0.288(7)	0.292(9)	0.294(10)	0.301(11)	0.290(8)	0.322(13)	0.311(12)	0.287(5)	0.288(6)	0.326(14)	0.272(1)	0.276(2)	0.278(3)	0.287(4)
	T1-DCI	0.281(1)	0.492(13)	0.503(14)	0.485(12)	0.281(1)	0.393(9)	0.409(10)	0.446(11)	0.281(1)	0.362(7)	0.354(6)	0.364(8)	0.347(5)	0.289(4)
	T1-PCK	0.220(4)	0.224(8)	0.230(9)	0.232(11)	0.222(6)	0.198(1)	0.204(2)	0.210(3)	0.222(5)	0.265(13)	0.237(12)	0.266(14)	0.230(10)	0.223(7)
	T1-SEP	0.336(11)	0.296(4)	0.294(3)	0.284(1)	0.336(12)	0.307(5)	0.311(6)	0.287(2)	0.335(9)	0.369(13)	0.336(10)	0.322(8)	0.311(7)	0.370(14)
	T2-CK	0.358(12)	0.345(4)	0.351(9)	0.343(3)	0.358(11)	0.327(1)	0.339(2)	0.348(5)	0.357(10)	0.394(14)	0.349(6)	0.374(13)	0.350(7)	0.351(8)
	T2-DCI	0.275(1)	0.417(12)	0.426(13)	0.431(14)	0.276(2)	0.350(9)	0.365(11)	0.352(10)	0.277(3)	0.293(4)	0.303(5)	0.312(7)	0.331(8)	0.304(6)
	T2-PCK	0.081(1)	0.114(10)	0.116(11)	0.108(8)	0.082(3)	0.126(13)	0.131(14)	0.121(12)	0.082(2)	0.107(7)	0.089(5)	0.113(9)	0.092(6)	0.085(4)
	T2-SEP	0.076(1)	0.137(14)	0.136(13)	0.130(12)	0.079(3)	0.111(10)	0.105(7)	0.108(8)	0.078(2)	0.116(11)	0.102(6)	0.110(9)	0.090(5)	0.084(4)
	U41Q1	0.315(13)	0.279(6)	0.278(5)	0.273(4)	0.311(11)	0.272(3)	0.270(2)	0.264(1)	0.314(12)	0.290(7)	0.308(10)	0.302(8)	0.305(9)	0.325(14)
	U41Q2	0.253(13)	0.207(3)	0.205(2)	0.199(1)	0.249(12)	0.216(7)	0.215(5)	0.226(10)	0.254(14)	0.220(8)	0.222(9)	0.210(4)	0.215(6)	0.244(11)
	U42Q1	0.174(14)	0.134(4)	0.128(1)	0.135(6)	0.169(9)	0.134(5)	0.129(3)	0.129(2)	0.170(10)	0.168(8)	0.170(11)	0.167(7)	0.173(13)	0.172(12)
	U42Q2	0.081(3)	0.111(13)	0.106(10)	0.102(6)	0.079(2)	0.106(11)	0.106(12)	0.104(9)	0.079(1)	0.102(7)	0.092(5)	0.103(8)	0.089(4)	0.115(14)
shot=0, cot=1	ASAP	0.165(2)	0.240(13)	0.237(12)	0.240(14)	0.173(4)	0.216(10)	0.220(11)	0.197(7)	0.170(3)	0.176(5)	0.210(9)	0.195(6)	0.201(8)	0.158(1)
	Semeval	0.669(7)	0.690(11)	0.700(13)	0.726(14)	0.677(10)	0.622(3)	0.619(2)	0.630(4)	0.674(9)	0.606(1)	0.659(6)	0.649(5)	0.696(12)	0.671(8)
	T1-CK	0.363(11)	0.305(1)	0.314(2)	0.314(3)	0.368(13)	0.324(4)	0.340(5)	0.346(7)	0.367(12)	0.354(8)	0.355(9)	0.373(14)	0.358(10)	0.345(6)
	T1-DCI	0.268(1)	0.388(12)	0.407(14)	0.399(13)	0.268(3)	0.333(10)	0.329(9)	0.338(11)	0.269(4)	0.282(6)	0.283(7)	0.298(8)	0.279(5)	0.268(2)
	T1-PCK	0.192(3)	0.238(12)	0.243(13)	0.252(14)	0.195(5)	0.180(1)	0.189(2)	0.209(10)	0.193(4)	0.206(9)	0.204(8)	0.216(11)	0.198(6)	0.204(7)
	T1-SEP	0.348(7)	0.367(14)	0.364(12)	0.362(11)	0.345(5)	0.351(9)	0.353(10)	0.349(8)	0.347(6)	0.343(4)	0.333(2)	0.340(3)	0.331(1)	0.364(13)
	T2-CK	0.364(4)	0.377(9)	0.378(11)	0.378(10)	0.364(3)	0.389(12)	0.391(13)	0.376(8)	0.364(5)	0.396(14)	0.348(1)	0.374(7)	0.349(2)	0.368(6)
	T2-DCI	0.215(3)	0.305(14)	0.297(13)	0.295(12)	0.213(1)	0.251(8)	0.254(9)	0.284(11)	0.214(2)	0.256(10)	0.215(4)	0.229(7)	0.218(5)	0.226(6)
	T2-PCK	0.092(1)	0.149(10)	0.146(9)	0.150(11)	0.092(2)	0.180(14)	0.177(13)	0.152(12)	0.092(3)	0.116(8)	0.102(5)	0.106(7)	0.103(6)	0.094(4)
	T2-SEP	0.084(1)	0.157(14)	0.156(13)	0.156(12)	0.086(2)	0.132(8)	0.141(10)	0.135(9)	0.086(3)	0.155(11)	0.118(6)	0.112(5)	0.090(4)	0.125(7)
	U41Q1	0.312(10)	0.290(5)	0.293(6)	0.289(4)	0.312(9)	0.281(3)	0.277(1)	0.277(2)	0.313(11)	0.301(8)	0.315(12)	0.316(13)	0.318(14)	0.300(7)
	U41Q2	0.246(13)	0.223(9)	0.220(7)	0.194(2)	0.241(11)	0.207(4)	0.197(3)	0.190(1)	0.242(12)	0.232(10)	0.211(6)	0.209(5)	0.221(8)	0.252(14)
	U42Q1	0.172(13)	0.154(7)	0.151(6)	0.150(5)	0.163(10)	0.139(2)	0.135(1)	0.141(3)	0.167(11)	0.170(12)	0.155(8)	0.160(9)	0.145(4)	0.182(14)
	U42Q2	0.068(4)	0.088(9)	0.090(11)	0.091(13)	0.053(1)	0.078(7)	0.089(10)	0.091(12)	0.056(2)	0.087(8)	0.076(6)	0.073(5)	0.068(3)	0.094(14)
shot=2, cot=1	ASAP	0.145(3)	0.237(13)	0.240(14)	0.228(12)	0.138(2)	0.216(10)	0.225(11)	0.210(9)	0.138(1)	0.155(4)	0.173(8)	0.156(5)	0.163(7)	0.159(6)
	Semeval	0.670(5)	0.683(7)	0.690(10)	0.720(13)	0.688(9)	0.662(3)	0.669(4)	0.659(2)	0.678(6)	0.706(12)	0.686(8)	0.705(11)	0.726(14)	0.657(1)
	T1-CK	0.084(1)	0.097(7)	0.101(9)	0.095(5)	0.095(6)	0.101(10)	0.110(11)	0.098(8)	0.090(4)	0.117(13)	0.117(14)	0.110(12)	0.085(2)	0.086(3)
	T1-DCI	0.270(4)	0.342(12)	0.330(10)	0.316(9)	0.271(6)	0.377(14)	0.358(13)	0.335(11)	0.270(5)	0.302(8)	0.250(1)	0.259(2)	0.263(3)	0.282(7)
	T1-PCK	0.124(3)	0.200(13)	0.203(14)	0.198(12)	0.117(1)	0.174(11)	0.173(10)	0.163(9)	0.122(2)	0.151(6)	0.156(7)	0.135(4)	0.145(5)	0.159(8)
	T1-SEP	0.311(2)	0.328(6)	0.327(5)	0.316(4)	0.312(3)	0.344(8)	0.335(7)	0.351(9)	0.309(1)	0.372(11)	0.409(14)	0.383(12)	0.389(13)	0.371(10)
	T2-CK	0.195(4)	0.191(1)	0.191(2)	0.206(6)	0.197(5)	0.246(12)	0.251(14)	0.249(13)	0.194(3)	0.232(10)	0.241(11)	0.228(9)	0.213(8)	0.209(7)
	T2-DCI	0.183(2)	0.273(11)	0.271(10)	0.284(12)	0.188(4)	0.290(14)	0.288(13)	0.256(9)	0.185(3)	0.243(8)	0.224(6)	0.234(7)	0.214(5)	0.178(1)
	T2-PCK	0.103(3)	0.197(14)	0.187(13)	0.162(12)	0.107(6)	0.137(8)	0.143(10)	0.139(9)	0.105(5)	0.152(11)	0.101(1)	0.105(4)	0.102(2)	0.116(7)
	T2-SEP	0.163(1)	0.202(11)	0.205(12)	0.206(13)	0.176(7)	0.177(8)	0.181(9)	0.185(10)	0.174(6)	0.216(14)	0.164(2)	0.172(5)	0.171(3)	0.172(4)
	U41Q1	0.138(7)	0.135(3)	0.136(6)	0.139(8)	0.134(2)	0.144(12)	0.142(10)	0.135(4)	0.133(1)	0.142(11)	0.145(13)	0.147(14)	0.142(9)	0.135(5)
	U41Q2	0.175(13)	0.168(6)	0.175(12)	0.179(14)	0.174(10)	0.168(7)	0.166(5)	0.151(3)	0.175(11)	0.173(9)	0.140(1)	0.162(4)	0.148(2)	0.171(8)
	U42Q1	0.109(11)	0.103(5)	0.101(4)	0.104(6)	0.109(10)	0.110(12)	0.108(8)	0.111(13)	0.108(7)	0.116(14)	0.088(1)	0.090(2)	0.092(3)	0.108(9)
	U42Q2	0.064(3)	0.086(14)	0.084(13)	0.081(12)	0.061(2)	0.078(9)	0.080(10)	0.072(5)	0.060(1)	0.075(7)	0.073(6)	0.069(4)	0.080(11)	0.075(8)

Table 8: Metric value and ranking results of each uncertainty quantification methods with Ministral-3-14B over 14 questions.

AUROC

Control	Question	CE	Embed_Eigen	Embed_GE	Embed_NAD	FSD	JS_Eigen	JS_GE	JS_NAD	MAR	NLI_DSE	NLI_Eigen	NLI_GE	NLI_NAD	Numset
shot=0, cot=0	ASAP	0.757(3)	0.524(14)	0.534(13)	0.567(10)	0.770(2)	0.559(11)	0.556(12)	0.629(8)	0.782(1)	0.584(9)	0.673(7)	0.686(6)	0.711(4)	0.692(5)
	Semeval	0.571(3)	0.462(9)	0.464(8)	0.521(5)	0.609(1)	0.429(12)	0.422(13)	0.480(7)	0.603(2)	0.412(14)	0.460(10)	0.451(11)	0.484(6)	0.542(4)
	T1-CK	0.661(4)	0.491(14)	0.517(12)	0.552(11)	0.652(6)	0.661(3)	0.676(2)	0.732(1)	0.651(7)	0.501(13)	0.624(8)	0.607(10)	0.613(9)	0.657(5)
	T1-DCI	0.607(6)	0.562(9)	0.573(8)	0.558(11)	0.623(3)	0.611(4)	0.607(5)	0.640(1)	0.626(2)	0.424(14)	0.515(12)	0.507(13)	0.559(10)	0.574(7)
	T1-PCK	0.805(3)	0.645(13)	0.668(11)	0.683(10)	0.818(1)	0.736(7)	0.758(4)	0.750(5)	0.813(2)	0.503(14)	0.662(12)	0.696(9)	0.709(8)	0.741(6)
	T1-SEP	0.729(1)	0.564(10)	0.572(9)	0.614(8)	0.708(3)	0.629(7)	0.639(6)	0.684(4)	0.719(2)	0.476(11)	0.419(14)	0.470(13)	0.476(12)	0.663(5)
	T2-CK	0.768(2)	0.657(10)	0.666(9)	0.706(6)	0.763(4)	0.694(8)	0.697(7)	0.826(1)	0.764(3)	0.477(14)	0.577(12)	0.555(13)	0.620(11)	0.738(5)
	T2-DCI	0.840(3)	0.674(13)	0.693(10)	0.680(12)	0.854(1)	0.686(11)	0.694(9)	0.756(6)	0.854(2)	0.517(14)	0.746(7)	0.761(5)	0.806(4)	0.721(8)
	T2-PCK	0.763(1)	0.651(9)	0.671(8)	0.633(12)	0.755(3)	0.649(10)	0.648(11)	0.626(13)	0.759(2)	0.515(14)	0.701(5)	0.700(6)	0.752(4)	0.673(7)
	T2-SEP	0.727(6)	0.582(10)	0.580(11)	0.618(9)	0.758(4)	0.803(3)	0.809(2)	0.841(1)	0.747(5)	0.508(14)	0.515(13)	0.638(8)	0.658(7)	0.565(12)
	U41Q1	0.419(12)	0.612(5)	0.597(6)	0.716(4)	0.274(14)	0.743(2)	0.758(1)	0.732(3)	0.313(13)	0.504(9)	0.440(11)	0.541(7)	0.446(10)	0.508(8)
	U41Q2	0.698(4)	0.650(8)	0.654(7)	0.658(6)	0.581(13)	0.701(2)	0.689(5)	0.716(1)	0.630(9)	0.558(14)	0.590(12)	0.626(10)	0.618(11)	0.699(3)
	U42Q1	0.545(8)	0.624(6)	0.643(4)	0.634(5)	0.538(9)	0.672(3)	0.680(2)	0.685(1)	0.546(7)	0.482(13)	0.412(14)	0.497(11)	0.484(12)	0.529(10)
	U42Q2	0.548(7)	0.514(9)	0.529(8)	0.549(6)	0.598(3)	0.580(5)	0.602(2)	0.612(1)	0.594(4)	0.490(11)	0.432(12)	0.408(14)	0.414(13)	0.497(10)
shot=0, cot=1	ASAP	0.751(2)	0.557(12)	0.563(11)	0.615(5)	0.732(3)	0.592(7)	0.577(9)	0.607(6)	0.754(1)	0.527(14)	0.567(10)	0.554(13)	0.587(8)	0.681(4)
	Semeval	0.628(4)	0.562(7)	0.575(5)	0.631(3)	0.639(1)	0.499(10)	0.484(11)	0.530(8)	0.635(2)	0.500(9)	0.454(14)	0.464(13)	0.466(12)	0.571(6)
	T1-CK	0.569(7)	0.500(12)	0.491(14)	0.539(11)	0.550(9)	0.595(3)	0.594(4)	0.669(1)	0.559(8)	0.498(13)	0.574(6)	0.579(5)	0.647(2)	0.544(10)
	T1-DCI	0.513(13)	0.582(6)	0.588(5)	0.625(4)	0.525(10)	0.651(3)	0.663(2)	0.669(1)	0.519(12)	0.526(9)	0.535(7)	0.528(8)	0.523(11)	0.509(14)
	T1-PCK	0.802(3)	0.506(12)	0.506(12)	0.467(14)	0.805(1)	0.649(9)	0.647(10)	0.710(6)	0.805(2)	0.510(11)	0.698(7)	0.675(8)	0.731(5)	0.755(4)
	T1-SEP	0.748(1)	0.607(10)	0.611(9)	0.621(8)	0.709(3)	0.655(6)	0.644(7)	0.677(5)	0.729(2)	0.448(14)	0.452(12)	0.450(13)	0.496(11)	0.698(4)
	T2-CK	0.741(2)	0.606(11)	0.623(9)	0.669(6)	0.738(4)	0.619(10)	0.633(8)	0.768(1)	0.740(3)	0.524(14)	0.577(13)	0.590(12)	0.637(7)	0.672(5)
	T2-DCI	0.844(3)	0.660(13)	0.686(12)	0.693(11)	0.846(2)	0.751(9)	0.766(8)	0.789(7)	0.852(1)	0.520(14)	0.818(6)	0.820(5)	0.836(4)	0.740(10)
	T2-PCK	0.769(3)	0.557(11)	0.579(9)	0.523(13)	0.800(1)	0.552(12)	0.571(10)	0.588(8)	0.792(2)	0.511(14)	0.696(5)	0.676(6)	0.715(4)	0.652(7)
	T2-SEP	0.719(6)	0.463(14)	0.480(13)	0.573(10)	0.733(5)	0.771(3)	0.791(2)	0.819(1)	0.734(4)	0.497(12)	0.525(11)	0.618(9)	0.685(7)	0.635(8)
	U41Q1	0.463(9)	0.618(4)	0.637(2)	0.654(1)	0.328(14)	0.620(3)	0.601(6)	0.605(5)	0.371(13)	0.513(8)	0.380(12)	0.396(11)	0.458(10)	0.577(7)
	U41Q2	0.635(2)	0.545(11)	0.551(9)	0.593(5)	0.574(6)	0.562(8)	0.570(7)	0.644(1)	0.607(4)	0.453(13)	0.433(14)	0.514(12)	0.549(10)	0.630(3)
	U42Q1	0.582(9)	0.549(11)	0.541(12)	0.571(10)	0.605(8)	0.664(3)	0.654(4)	0.666(1)	0.613(7)	0.491(14)	0.639(5)	0.666(2)	0.618(6)	0.527(13)
	U42Q2	0.609(5)	0.580(8)	0.570(9)	0.591(7)	0.605(6)	0.653(2)	0.655(1)	0.644(3)	0.630(4)	0.494(11)	0.451(14)	0.459(13)	0.465(12)	0.552(10)
shot=2, cot=1	ASAP	0.739(3)	0.627(13)	0.635(11)	0.654(10)	0.747(2)	0.659(9)	0.668(7)	0.689(5)	0.764(1)	0.496(14)	0.627(12)	0.714(4)	0.683(6)	0.663(8)
	Semeval	0.560(6)	0.569(5)	0.585(4)	0.628(1)	0.598(2)	0.547(7)	0.543(8)	0.519(9)	0.591(3)	0.476(12)	0.450(14)	0.464(13)	0.479(11)	0.492(10)
	T1-CK	0.691(4)	0.570(10)	0.579(9)	0.638(5)	0.717(2)	0.627(7)	0.628(6)	0.712(3)	0.722(1)	0.431(14)	0.471(13)	0.478(12)	0.543(11)	0.621(8)
	T1-DCI	0.611(7)	0.536(12)	0.535(13)	0.537(11)	0.637(2)	0.612(6)	0.625(4)	0.670(1)	0.625(3)	0.495(14)	0.594(8)	0.617(5)	0.571(10)	0.571(9)
	T1-PCK	0.822(3)	0.495(14)	0.526(12)	0.521(13)	0.857(1)	0.562(9)	0.549(10)	0.589(8)	0.850(2)	0.531(11)	0.615(7)	0.694(6)	0.719(5)	0.770(4)
	T1-SEP	0.676(2)	0.542(10)	0.547(9)	0.563(8)	0.641(5)	0.634(6)	0.618(7)	0.647(4)	0.656(3)	0.495(11)	0.379(14)	0.426(13)	0.427(12)	0.676(1)
	T2-CK	0.798(1)	0.696(8)	0.715(7)	0.795(2)	0.777(4)	0.658(9)	0.647(10)	0.751(5)	0.794(3)	0.455(14)	0.458(13)	0.461(12)	0.499(11)	0.735(6)
	T2-DCI	0.748(3)	0.469(14)	0.473(13)	0.475(12)	0.783(1)	0.572(9)	0.560(10)	0.599(8)	0.780(2)	0.511(11)	0.696(6)	0.714(5)	0.714(4)	0.665(7)
	T2-PCK	0.682(5)	0.526(13)	0.540(12)	0.581(11)	0.706(2)	0.585(10)	0.598(8)	0.652(7)	0.705(3)	0.524(14)	0.654(6)	0.688(4)	0.713(1)	0.595(9)
	T2-SEP	0.709(3)	0.539(12)	0.530(13)	0.543(11)	0.708(5)	0.707(6)	0.709(4)	0.736(1)	0.721(2)	0.523(14)	0.587(10)	0.662(8)	0.656(9)	0.688(7)
	U41Q1	0.678(5)	0.802(3)	0.808(2)	0.810(1)	0.617(10)	0.668(7)	0.671(6)	0.707(4)	0.651(8)	0.458(13)	0.409(14)	0.513(11)	0.467(12)	0.619(9)
	U41Q2	0.642(1)	0.574(9)	0.559(11)	0.552(12)	0.630(4)	0.621(6)	0.612(7)	0.628(5)	0.642(2)	0.509(13)	0.483(14)	0.584(8)	0.564(10)	0.634(3)
	U42Q1	0.623(3)	0.567(10)	0.576(8)	0.572(9)	0.629(2)	0.605(5)	0.590(7)	0.608(4)	0.640(1)	0.509(13)	0.527(11)	0.522(12)	0.508(14)	0.602(6)
	U42Q2	0.533(3)	0.463(11)	0.457(13)	0.439(14)	0.577(1)	0.505(8)	0.531(5)	0.509(7)	0.567(2)	0.530(6)	0.533(4)	0.463(11)	0.481(10)	0.486(9)

C-index

Control	Question	CE	Embed_Eigen	Embed_GE	Embed_NAD	FSD	JS_Eigen	JS_GE	JS_NAD	MAR	NLI_DSE	NLI_Eigen	NLI_GE	NLI_NAD	Numset
shot=0, cot=0	ASAP	0.645(5)	0.539(13)	0.540(11)	0.549(10)	0.657(1)	0.539(12)	0.527(14)	0.594(8)	0.656(2)	0.580(9)	0.622(6)	0.653(4)	0.654(3)	0.615(7)
	Semeval	0.514(3)	0.434(9)	0.437(8)	0.491(5)	0.548(1)	0.410(12)	0.404(13)	0.451(6)	0.542(2)	0.388(14)	0.433(10)	0.418(11)	0.445(7)	0.494(4)
	T1-CK	0.531(5)	0.446(14)	0.451(13)	0.463(12)	0.519(7)	0.581(3)	0.593(2)	0.657(1)	0.520(6)	0.485(11)	0.511(9)	0.512(8)	0.504(10)	0.549(4)
	T1-DCI	0.591(6)	0.548(11)	0.560(8)	0.549(10)	0.608(3)	0.600(4)	0.594(5)	0.626(1)	0.609(2)	0.427(14)	0.507(12)	0.493(13)	0.550(9)	0.563(7)
	T1-PCK	0.791(3)	0.630(13)	0.651(12)	0.662(10)	0.803(1)	0.712(8)	0.725(5)	0.723(6)	0.800(2)	0.506(14)	0.651(11)	0.693(9)	0.714(7)	0.728(4)
	T1-SEP	0.570(4)	0.513(10)	0.517(9)	0.542(5)	0.582(2)	0.532(7)	0.540(6)	0.585(1)	0.580(3)	0.465(11)	0.383(14)	0.421(12)	0.414(13)	0.527(8)
	T2-CK	0.612(6)	0.601(10)	0.608(8)	0.637(4)	0.607(9)	0.664(2)	0.662(3)	0.734(1)	0.608(7)	0.484(14)	0.496(13)	0.497(12)	0.511(11)	0.627(5)
	T2-DCI	0.727(4)	0.616(12)	0.628(11)	0.610(13)	0.736(3)	0.655(10)	0.663(8)	0.702(5)	0.736(2)	0.514(14)	0.695(7)	0.699(6)	0.752(1)	0.656(9)
	T2-PCK	0.722(2)	0.625(9)	0.640(7)	0.603(12)	0.717(4)	0.619(10)	0.619(11)	0.587(13)	0.719(3)	0.517(14)	0.687(6)	0.695(5)	0.737(1)	0.632(8)
	T2-SEP	0.717(6)	0.569(10)	0.568(11)	0.615(9)	0.748(4)	0.783(3)	0.789(2)	0.825(1)	0.737(5)	0.508(14)	0.514(13)	0.625(8)	0.648(7)	0.554(12)
	U41Q1	0.427(11)	0.568(5)	0.563(6)	0.591(3)	0.374(14)	0.580(4)	0.609(2)	0.611(1)	0.379(13)	0.488(7)	0.412(12)	0.470(9)	0.443(10)	0.486(8)
	U41Q2	0.561(9)	0.666(3)	0.677(2)	0.707(1)	0.491(14)	0.636(6)	0.639(5)	0.665(4)	0.509(13)	0.552(10)	0.582(8)	0.537(12)	0.539(11)	0.605(7)
	U42Q1	0.510(7)	0.621(6)	0.639(4)	0.630(5)	0.495(11)	0.679(2)	0.689(1)	0.667(3)	0.503(8)	0.485(12)	0.448(14)	0.499(10)	0.476(13)	0.502(9)
	U42Q2	0.520(7)	0.502(9)	0.517(8)	0.538(6)	0.563(4)	0.578(3)	0.603(2)	0.625(1)	0.553(5)	0.492(10)	0.456(12)	0.443(14)	0.452(13)	0.483(11)
shot=0, cot=1	ASAP	0.611(3)	0.505(14)	0.525(11)	0.554(6)	0.615(2)	0.545(8)	0.519(13)	0.526(10)	0.616(1)	0.549(7)	0.561(5)	0.524(12)	0.536(9)	0.566(4)
	Semeval	0.614(4)	0.565(6)	0.576(5)	0.624(2)	0.624(1)	0.497(10)	0.485(11)	0.527(8)	0.619(3)	0.500(9)	0.457(14)	0.459(13)	0.458(13)	0.560(7)
	T1-CK	0.517(7)	0.469(12)	0.457(14)	0.460(13)	0.495(10)	0.591(2)	0.586(3)	0.650(1)	0.502(9)	0.495(10)	0.509(8)	0.530(6)	0.557(4)	0.533(5)
	T1-DCI	0.513(13)	0.582(6)	0.588(5)	0.625(4)	0.525(10)	0.651(3)	0.663(2)	0.669(1)	0.519(12)	0.526(9)	0.535(7)	0.528(8)	0.523(11)	0.509(14)
	T1-PCK	0.775(3)	0.509(12)	0.510(11)	0.473(14)	0.786(1)	0.643(9)	0.641(10)	0.700(6)	0.784(2)	0.507(13)	0.679(7)	0.674(8)	0.713(5)	0.724(4)
	T1-SEP	0.610(2)	0.554(5)	0.541(8)	0.529(9)	0.572(4)	0.547(6)	0.523(10)	0.545(7)	0.587(3)	0.453(11)	0.376(13)	0.368(14)	0.395(12)	0.612(1)
	T2-CK	0.599(4)	0.581(9)	0.586(5)	0.619(2)	0.582(8)	0.573(10)	0.583(7)	0.656(1)	0.585(6)	0.520(11)	0.498(14)	0.501(13)	0.509(12)	0.600(3)
	T2-DCI	0.771(4)	0.616(13)	0.633(12)	0.638(11)	0.775(2)	0.704(9)	0.717(8)	0.736(7)	0.777(1)	0.516(14)	0.762(5)	0.761(6)	0.773(3)	0.685(10)
	T2-PCK	0.768(3)	0.560(11)	0.582(9)	0.527(13)	0.802(1)	0.550(12)	0.571(10)	0.588(8)	0.793(2)	0.511(14)	0.696(5)	0.669(6)	0.712(4)	0.648(7)
	T2-SEP	0.689(6)	0.482(14)	0.497(13)	0.571(10)	0.696(5)	0.750(3)	0.773(2)	0.803(1)	0.697(4)	0.498(12)	0.508(11)	0.613(9)	0.652(7)	0.618(8)
	U41Q1	0.467(12)	0.549(5)	0.568(3)	0.603(1)	0.405(14)	0.547(6)	0.557(4)	0.582(2)	0.422(13)	0.511(7)	0.491(11)	0.495(10)	0.510(8)	0.507(9)
	U41Q2	0.573(7)	0.583(5)	0.582(6)	0.622(2)	0.528(10)	0.601(4)	0.609(3)	0.640(1)	0.553(9)	0.436(13)	0.417(14)	0.462(12)	0.497(11)	0.567(8)
	U42Q1	0.538(12)	0.558(10)	0.555(11)	0.582(6)	0.564(9)	0.672(1)	0.662(2)	0.641(3)	0.565(8)	0.492(13)	0.619(5)	0.624(4)	0.566(7)	0.489(14)
	U42Q2	0.536(7)	0.579(6)	0.580(5)	0.597(4)	0.500(11)	0.645(2)	0.653(1)	0.639(3)	0.523(9)	0.504(10)	0.463(12)	0.426(14)	0.433(13)	0.525(8)
shot=2, cot=1	ASAP	0.607(5)	0.589(8)	0.586(10)	0.577(11)	0.624(3)	0.575(12)	0.591(6)	0.589(7)	0.622(4)	0.501(14)	0.589(9)	0.663(1)	0.628(2)	0.559(13)
	Semeval	0.533(8)	0.561(5)	0.574(2)	0.616(1)	0.570(3)	0.551(6)	0.550(7)	0.517(9)	0.562(4)	0.477(10)	0.439(14)	0.457(13)	0.464(12)	0.472(11)
	T1-CK	0.663(4)	0.585(10)	0.595(9)	0.639(5)	0.669(3)	0.617(7)	0.616(8)	0.697(1)	0.677(2)	0.450(12)	0.446(14)	0.449(13)	0.501(11)	0.617(6)
	T1-DCI	0.594(7)	0.523(11)	0.520(13)	0.520(12)	0.621(3)	0.617(4)	0.627(2)	0.670(1)	0.608(5)	0.495(14)	0.580(8)	0.602(6)	0.556(9)	0.555(10)
	T1-PCK	0.735(3)	0.431(14)	0.452(13)	0.469(12)	0.766(1)	0.544(8)	0.539(9)	0.560(7)	0.748(2)	0.498(11)	0.522(10)	0.607(6)	0.629(5)	0.715(4)
	T1-SEP	0.553(7)	0.565(4)	0.565(2)	0.561(6)	0.518(10)	0.565(3)	0.548(8)	0.564(5)	0.530(9)	0.495(11)	0.382(14)	0.399(12)	0.395(13)	0.602(1)
	T2-CK	0.725(2)	0.649(8)	0.671(7)	0.750(1)	0.700(5)	0.614(9)	0.607(10)	0.721(3)	0.715(4)	0.451(12)	0.423(14)	0.431(13)	0.456(11)	0.689(6)
	T2-DCI	0.690(3)	0.461(14)	0.479(13)	0.491(12)	0.712(1)	0.524(9)	0.518(10)	0.560(8)	0.711(2)	0.494(11)	0.636(6)	0.646(5)	0.649(4)	0.620(7)
	T2-PCK	0.603(6)	0.525(14)	0.533(12)	0.576(9)	0.625(4)	0.574(10)	0.586(8)	0.628(3)	0.619(5)	0.530(13)	0.590(7)	0.629(2)	0.635(1)	0.562(11)
	T2-SEP	0.605(7)	0.565(10)	0.549(11)	0.537(13)	0.605(6)	0.663(3)	0.669(2)	0.676(1)	0.612(4)	0.494(14)	0.538(12)	0.587(9)	0.605(5)	0.592(8)
	U41Q1	0.491(8)	0.638(3)	0.653(2)	0.666(1)	0.434(12)	0.594(5)	0.595(4)	0.580(6)	0.447(10)	0.477(9)	0.425(13)	0.440(11)	0.391(14)	0.525(7)
	U41Q2	0.513(9)	0.557(4)	0.547(5)	0.531(7)	0.499(11)	0.610(1)	0.590(2)	0.572(3)	0.503(10)	0.492(12)	0.472(14)	0.525(8)	0.477(13)	0.543(6)
	U42Q1	0.542(10)	0.601(4)	0.609(1)	0.595(5)	0.543(9)	0.605(3)	0.589(6)	0.608(2)	0.549(7)	0.516(11)	0.501(12)	0.484(13)	0.459(14)	0.545(8)
	U42Q2	0.495(7)	0.436(12)	0.429(13)	0.426(14)	0.506(5)	0.487(8)	0.510(3)	0.515(2)	0.501(6)	0.521(1)	0.507(4)	0.456(11)	0.467(10)	0.475(9)

AUARC

Control	Question	CE	Embed_Eigen	Embed_GE	Embed_NAD	FSD	JS_Eigen	JS_GE	JS_NAD	MAR	NLI_DSE	NLI_Eigen	NLI_GE	NLI_NAD	Numset
shot=0, cot=0	ASAP	0.666(3)	0.520(14)	0.525(13)	0.529(10)	0.669(2)	0.529(11)	0.525(12)	0.568(8)	0.677(1)	0.539(9)	0.628(6)	0.609(7)	0.652(4)	0.639(5)
	Semeval	0.553(3)	0.476(8)	0.475(9)	0.491(5)	0.561(1)	0.441(13)	0.428(14)	0.468(11)	0.558(2)	0.481(6)	0.479(7)	0.471(10)	0.465(12)	0.551(4)
	T1-CK	0.252(8)	0.193(13)	0.205(12)	0.208(11)	0.256(5)	0.271(3)	0.278(2)	0.334(1)	0.255(6)	0.185(14)	0.255(7)	0.232(10)	0.247(9)	0.259(4)
	T1-DCI	0.737(4)	0.702(11)	0.708(10)	0.697(12)	0.741(2)	0.736(5)	0.731(6)	0.747(1)	0.741(3)	0.681(13)	0.680(14)	0.708(9)	0.728(8)	0.729(7)
	T1-PCK	0.771(3)	0.682(11)	0.689(10)	0.678(13)	0.776(1)	0.738(6)	0.745(4)	0.731(7)	0.773(2)	0.538(14)	0.679(12)	0.704(9)	0.712(8)	0.741(5)
	T1-SEP	0.404(4)	0.345(11)	0.347(10)	0.366(8)	0.398(6)	0.409(3)	0.418(2)	0.427(1)	0.401(5)	0.353(9)	0.290(14)	0.317(13)	0.332(12)	0.378(7)
	T2-CK	0.530(2)	0.462(10)	0.465(9)	0.509(5)	0.529(3)	0.499(7)	0.489(8)	0.551(1)	0.528(4)	0.328(14)	0.424(12)	0.404(13)	0.434(11)	0.509(6)
	T2-DCI	0.779(3)	0.669(12)	0.681(10)	0.661(13)	0.786(1)	0.681(11)	0.687(9)	0.720(8)	0.784(2)	0.593(14)	0.725(7)	0.734(5)	0.760(4)	0.733(6)
	T2-PCK	0.740(6)	0.738(8)	0.750(4)	0.733(9)	0.739(7)	0.711(12)	0.713(11)	0.691(13)	0.740(5)	0.689(14)	0.766(2)	0.753(3)	0.795(1)	0.716(10)
	T2-SEP	0.729(6)	0.689(10)	0.688(11)	0.713(8)	0.747(4)	0.822(3)	0.827(2)	0.840(1)	0.739(5)	0.517(14)	0.633(12)	0.695(9)	0.720(7)	0.629(13)
	U41Q1	0.044(12)	0.095(5)	0.092(6)	0.113(4)	0.028(14)	0.134(3)	0.143(2)	0.143(1)	0.033(13)	0.064(11)	0.064(9)	0.092(7)	0.064(10)	0.082(8)
	U41Q2	0.230(7)	0.248(2)	0.246(3)	0.242(4)	0.187(13)	0.237(6)	0.238(5)	0.266(1)	0.206(10)	0.140(14)	0.210(9)	0.199(12)	0.204(11)	0.223(8)
	U42Q1	0.425(10)	0.476(6)	0.495(4)	0.492(5)	0.419(12)	0.584(3)	0.589(2)	0.590(1)	0.419(13)	0.451(7)	0.361(14)	0.430(9)	0.422(11)	0.432(8)
	U42Q2	0.421(6)	0.403(9)	0.405(8)	0.413(7)	0.464(1)	0.431(5)	0.442(3)	0.438(4)	0.449(2)	0.400(10)	0.330(13)	0.332(12)	0.323(14)	0.390(11)
shot=0, cot=1	ASAP	0.685(3)	0.533(14)	0.543(13)	0.575(12)	0.688(2)	0.599(6)	0.589(9)	0.595(7)	0.690(1)	0.576(11)	0.593(8)	0.584(10)	0.613(5)	0.656(4)
	Semeval	0.547(4)	0.525(8)	0.530(7)	0.561(1)	0.548(3)	0.472(10)	0.465(11)	0.485(9)	0.549(2)	0.534(6)	0.418(14)	0.438(13)	0.444(12)	0.536(5)
	T1-CK	0.184(10)	0.202(8)	0.189(9)	0.207(6)	0.177(13)	0.210(5)	0.226(3)	0.255(1)	0.184(11)	0.172(14)	0.205(7)	0.218(4)	0.226(2)	0.177(12)
	T1-DCI	0.709(13)	0.729(7)	0.731(6)	0.752(4)	0.712(10)	0.786(3)	0.793(1)	0.792(2)	0.714(9)	0.745(5)	0.712(11)	0.717(8)	0.695(14)	0.710(12)
	T1-PCK	0.799(3)	0.583(11)	0.573(12)	0.537(14)	0.801(1)	0.703(10)	0.704(9)	0.717(8)	0.800(2)	0.569(13)	0.741(6)	0.717(7)	0.748(5)	0.774(4)
	T1-SEP	0.448(1)	0.361(8)	0.347(9)	0.339(10)	0.435(4)	0.404(6)	0.398(7)	0.422(5)	0.438(3)	0.301(13)	0.279(14)	0.309(12)	0.320(11)	0.446(2)
	T2-CK	0.458(5)	0.437(9)	0.444(8)	0.466(2)	0.460(4)	0.452(6)	0.451(7)	0.512(1)	0.460(3)	0.321(14)	0.350(13)	0.374(12)	0.392(11)	0.418(10)
	T2-DCI	0.766(2)	0.670(13)	0.684(12)	0.692(11)	0.764(4)	0.746(9)	0.755(8)	0.758(6)	0.768(1)	0.627(14)	0.759(5)	0.756(7)	0.765(3)	0.713(10)
	T2-PCK	0.811(3)	0.663(13)	0.678(11)	0.654(14)	0.822(1)	0.702(9)	0.709(8)	0.684(10)	0.818(2)	0.674(12)	0.774(6)	0.756(7)	0.793(4)	0.778(5)
	T2-SEP	0.750(7)	0.634(13)	0.643(11)	0.672(9)	0.751(6)	0.807(3)	0.818(2)	0.823(1)	0.760(4)	0.482(14)	0.642(12)	0.728(8)	0.759(5)	0.652(10)
	U41Q1	0.040(10)	0.067(3)	0.070(2)	0.074(1)	0.021(14)	0.052(5)	0.050(6)	0.052(4)	0.027(13)	0.050(7)	0.028(12)	0.040(11)	0.044(9)	0.046(8)
	U41Q2	0.231(2)	0.195(10)	0.201(9)	0.211(8)	0.211(7)	0.180(11)	0.178(12)	0.229(3)	0.220(5)	0.103(14)	0.160(13)	0.217(6)	0.225(4)	0.240(1)
	U42Q1	0.508(9)	0.481(12)	0.469(14)	0.477(13)	0.523(7)	0.577(1)	0.577(3)	0.577(2)	0.519(8)	0.499(10)	0.569(4)	0.559(5)	0.532(6)	0.489(11)
	U42Q2	0.385(6)	0.366(7)	0.356(10)	0.358(9)	0.387(5)	0.460(1)	0.452(2)	0.439(3)	0.402(4)	0.292(14)	0.324(11)	0.308(12)	0.308(13)	0.358(8)
shot=2, cot=1	ASAP	0.706(4)	0.666(9)	0.668(8)	0.651(13)	0.715(2)	0.658(12)	0.668(7)	0.682(6)	0.716(1)	0.585(14)	0.659(10)	0.709(3)	0.704(5)	0.659(11)
	Semeval	0.496(8)	0.535(5)	0.549(3)	0.568(2)	0.513(6)	0.490(10)	0.493(9)	0.475(13)	0.505(7)	0.622(1)	0.477(11)	0.462(14)	0.476(12)	0.541(4)
	T1-CK	0.767(6)	0.738(9)	0.736(10)	0.757(7)	0.776(4)	0.777(3)	0.777(2)	0.812(1)	0.774(5)	0.727(11)	0.680(14)	0.682(13)	0.716(12)	0.749(8)
	T1-DCI	0.772(3)	0.693(12)	0.693(13)	0.693(11)	0.781(1)	0.736(8)	0.735(9)	0.749(5)	0.777(2)	0.640(14)	0.727(10)	0.763(4)	0.742(7)	0.745(6)
	T1-PCK	0.858(3)	0.675(13)	0.691(11)	0.690(12)	0.869(1)	0.729(9)	0.713(10)	0.731(8)	0.869(2)	0.605(14)	0.758(7)	0.797(5)	0.792(6)	0.830(4)
	T1-SEP	0.502(2)	0.391(11)	0.400(10)	0.413(9)	0.490(4)	0.452(6)	0.446(7)	0.484(5)	0.495(3)	0.413(8)	0.314(14)	0.326(13)	0.339(12)	0.514(1)
	T2-CK	0.835(2)	0.781(10)	0.790(7)	0.835(1)	0.827(5)	0.785(8)	0.782(9)	0.832(3)	0.831(4)	0.615(14)	0.667(12)	0.638(13)	0.689(11)	0.816(6)
	T2-DCI	0.710(3)	0.545(13)	0.547(12)	0.528(14)	0.725(1)	0.621(8)	0.603(10)	0.610(9)	0.721(2)	0.559(11)	0.669(6)	0.678(4)	0.676(5)	0.665(7)
	T2-PCK	0.783(4)	0.694(14)	0.709(13)	0.732(10)	0.793(2)	0.721(12)	0.721(11)	0.750(7)	0.786(3)	0.737(9)	0.755(6)	0.783(5)	0.799(1)	0.744(8)
	T2-SEP	0.725(6)	0.596(12)	0.591(13)	0.616(11)	0.725(5)	0.743(3)	0.744(2)	0.766(1)	0.731(4)	0.489(14)	0.680(10)	0.724(7)	0.724(8)	0.683(9)
	U41Q1	0.192(7)	0.288(2)	0.283(3)	0.309(1)	0.188(9)	0.223(4)	0.210(5)	0.210(6)	0.191(8)	0.119(13)	0.094(14)	0.164(11)	0.125(12)	0.172(10)
	U41Q2	0.393(2)	0.334(10)	0.332(11)	0.329(12)	0.376(6)	0.388(3)	0.384(4)	0.376(7)	0.381(5)	0.310(14)	0.319(13)	0.358(8)	0.344(9)	0.398(1)
	U42Q1	0.555(7)	0.535(11)	0.549(10)	0.552(8)	0.551(9)	0.580(1)	0.563(5)	0.570(4)	0.560(6)	0.577(2)	0.500(12)	0.491(13)	0.474(14)	0.572(3)
	U42Q2	0.349(3)	0.295(10)	0.293(12)	0.293(13)	0.373(1)	0.312(8)	0.319(7)	0.302(9)	0.366(2)	0.324(6)	0.334(4)	0.290(14)	0.295(11)	0.332(5)

AUERC

Control	Question	CE	Embed_Eigen	Embed_GE	Embed_NAD	FSD	JS_Eigen	JS_GE	JS_NAD	MAR	NLI_DSE	NLI_Eigen	NLI_GE	NLI_NAD	Numset
shot=0, cot=0	ASAP	0.125(4)	0.154(10)	0.156(11)	0.158(12)	0.124(3)	0.160(13)	0.163(14)	0.149(8)	0.122(2)	0.152(9)	0.129(6)	0.128(5)	0.121(1)	0.131(7)
	Semeval	0.435(4)	0.495(7)	0.495(8)	0.481(5)	0.428(1)	0.528(13)	0.540(14)	0.508(11)	0.431(2)	0.498(9)	0.489(6)	0.501(10)	0.516(12)	0.434(3)
	T1-CK	0.545(7)	0.597(14)	0.594(13)	0.589(12)	0.543(5)	0.498(3)	0.485(2)	0.437(1)	0.544(6)	0.573(11)	0.551(9)	0.557(10)	0.551(8)	0.531(4)
	T1-DCI	0.253(5)	0.287(11)	0.281(10)	0.291(12)	0.248(2)	0.253(4)	0.258(6)	0.242(1)	0.249(3)	0.307(14)	0.307(13)	0.281(9)	0.260(8)	0.260(7)
	T1-PCK	0.130(3)	0.211(13)	0.204(12)	0.203(11)	0.127(1)	0.169(8)	0.165(5)	0.166(7)	0.128(2)	0.330(14)	0.194(10)	0.175(9)	0.165(6)	0.156(4)
	T1-SEP	0.413(7)	0.431(11)	0.428(9)	0.414(8)	0.408(4)	0.381(3)	0.377(2)	0.373(1)	0.408(5)	0.412(6)	0.454(14)	0.433(12)	0.431(10)	0.444(13)
	T2-CK	0.384(5)	0.406(10)	0.402(9)	0.371(4)	0.385(7)	0.351(2)	0.358(3)	0.321(1)	0.385(6)	0.523(14)	0.468(12)	0.474(13)	0.455(11)	0.389(8)
	T2-DCI	0.133(4)	0.194(12)	0.187(11)	0.197(13)	0.129(1)	0.177(10)	0.174(9)	0.157(7)	0.131(2)	0.232(14)	0.153(5)	0.155(6)	0.133(3)	0.162(8)
	T2-PCK	0.125(8)	0.124(5)	0.119(4)	0.127(9)	0.125(7)	0.138(11)	0.138(10)	0.151(14)	0.125(6)	0.140(12)	0.108(2)	0.112(3)	0.094(1)	0.140(13)
	T2-SEP	0.131(6)	0.154(11)	0.153(10)	0.139(8)	0.122(4)	0.087(3)	0.084(2)	0.077(1)	0.126(5)	0.239(14)	0.180(12)	0.150(9)	0.136(7)	0.184(13)
	U41Q1	0.353(12)	0.304(5)	0.307(6)	0.302(4)	0.362(14)	0.298(3)	0.290(1)	0.293(2)	0.358(13)	0.338(9)	0.342(11)	0.332(7)	0.341(10)	0.332(8)
	U41Q2	0.234(8)	0.203(3)	0.201(2)	0.199(1)	0.250(13)	0.222(6)	0.219(5)	0.214(4)	0.244(11)	0.263(14)	0.230(7)	0.247(12)	0.242(10)	0.236(9)
	U42Q1	0.163(11)	0.135(6)	0.129(4)	0.129(5)	0.165(14)	0.108(2)	0.104(1)	0.112(3)	0.165(13)	0.140(7)	0.165(12)	0.151(8)	0.156(9)	0.162(10)
	U42Q2	0.168(6)	0.173(9)	0.172(8)	0.170(7)	0.156(4)	0.153(3)	0.148(2)	0.148(1)	0.161(5)	0.183(11)	0.190(14)	0.188(12)	0.189(13)	0.182(10)
shot=0, cot=1	ASAP	0.125(3)	0.163(14)	0.159(13)	0.149(11)	0.118(1)	0.133(4)	0.140(8)	0.144(10)	0.119(2)	0.153(12)	0.135(5)	0.143(9)	0.137(7)	0.137(6)
	Semeval	0.443(4)	0.458(8)	0.454(6)	0.426(1)	0.441(3)	0.513(10)	0.519(11)	0.501(9)	0.440(2)	0.455(7)	0.565(14)	0.549(13)	0.544(12)	0.452(5)
	T1-CK	0.606(9)	0.591(7)	0.607(10)	0.612(12)	0.615(13)	0.556(3)	0.548(2)	0.514(1)	0.608(11)	0.617(14)	0.589(6)	0.570(5)	0.560(4)	0.602(8)
	T1-DCI	0.281(13)	0.261(7)	0.259(6)	0.238(4)	0.278(10)	0.204(3)	0.197(1)	0.198(2)	0.276(9)	0.245(5)	0.278(11)	0.273(8)	0.295(14)	0.280(12)
	T1-PCK	0.119(3)	0.264(11)	0.269(12)	0.299(13)	0.116(1)	0.184(10)	0.183(9)	0.175(8)	0.117(2)	0.311(14)	0.163(6)	0.165(7)	0.153(5)	0.143(4)
	T1-SEP	0.346(2)	0.386(8)	0.407(9)	0.413(10)	0.357(4)	0.377(6)	0.382(7)	0.369(5)	0.356(3)	0.447(11)	0.475(14)	0.462(13)	0.453(12)	0.335(1)
	T2-CK	0.462(7)	0.442(6)	0.439(5)	0.420(4)	0.465(9)	0.417(3)	0.417(2)	0.377(1)	0.463(8)	0.534(14)	0.532(13)	0.521(12)	0.508(11)	0.478(10)
	T2-DCI	0.144(8)	0.189(13)	0.182(12)	0.180(11)	0.144(7)	0.143(6)	0.137(1)	0.138(2)	0.142(3)	0.208(14)	0.143(4)	0.147(9)	0.143(5)	0.177(10)
	T2-PCK	0.090(3)	0.164(13)	0.157(11)	0.169(14)	0.084(1)	0.147(9)	0.142(8)	0.155(10)	0.086(2)	0.160(12)	0.109(6)	0.124(7)	0.100(4)	0.107(5)
	T2-SEP	0.120(6)	0.168(12)	0.166(10)	0.151(9)	0.120(7)	0.091(3)	0.085(2)	0.082(1)	0.116(4)	0.257(14)	0.175(13)	0.126(8)	0.117(5)	0.168(11)
	U41Q1	0.351(11)	0.312(3)	0.310(2)	0.309(1)	0.358(14)	0.323(5)	0.323(6)	0.318(4)	0.356(13)	0.340(9)	0.341(10)	0.336(8)	0.326(7)	0.352(12)
	U41Q2	0.226(6)	0.224(3)	0.224(4)	0.219(2)	0.234(10)	0.228(8)	0.227(7)	0.216(1)	0.229(9)	0.287(14)	0.258(13)	0.241(12)	0.235(11)	0.224(5)
	U42Q1	0.134(13)	0.126(8)	0.128(10)	0.125(7)	0.129(11)	0.102(1)	0.103(2)	0.110(4)	0.130(12)	0.122(6)	0.110(3)	0.119(5)	0.128(9)	0.139(14)
	U42Q2	0.166(8)	0.160(6)	0.160(4)	0.160(5)	0.171(10)	0.130(1)	0.131(2)	0.137(3)	0.165(7)	0.184(12)	0.172(11)	0.185(14)	0.184(13)	0.171(9)
shot=2, cot=1	ASAP	0.125(9)	0.118(5)	0.120(7)	0.131(11)	0.116(3)	0.131(12)	0.124(8)	0.126(10)	0.119(6)	0.164(14)	0.117(4)	0.097(1)	0.103(2)	0.147(13)
	Semeval	0.493(10)	0.447(5)	0.434(3)	0.415(2)	0.476(6)	0.485(9)	0.481(7)	0.503(11)	0.484(8)	0.362(1)	0.507(12)	0.520(14)	0.509(13)	0.445(4)
	T1-CK	0.116(6)	0.126(9)	0.125(8)	0.117(7)	0.114(4)	0.111(2)	0.112(3)	0.090(1)	0.115(5)	0.150(11)	0.175(14)	0.174(13)	0.154(12)	0.128(10)
	T1-DCI	0.218(3)	0.295(11)	0.296(12)	0.296(13)	0.208(1)	0.246(7)	0.249(9)	0.236(5)	0.213(2)	0.332(14)	0.262(10)	0.226(4)	0.247(8)	0.245(6)
	T1-PCK	0.066(3)	0.153(13)	0.147(12)	0.144(11)	0.061(1)	0.118(8)	0.121(10)	0.121(9)	0.062(2)	0.167(14)	0.117(7)	0.096(5)	0.099(6)	0.072(4)
	T1-SEP	0.339(6)	0.331(5)	0.327(2)	0.329(4)	0.348(8)	0.351(9)	0.360(10)	0.329(3)	0.345(7)	0.389(11)	0.432(12)	0.447(14)	0.441(13)	0.313(1)
	T2-CK	0.098(3)	0.129(10)	0.121(7)	0.090(1)	0.103(5)	0.126(8)	0.127(9)	0.091(2)	0.100(4)	0.230(14)	0.203(12)	0.211(13)	0.185(11)	0.105(6)
	T2-DCI	0.180(3)	0.265(13)	0.256(12)	0.268(14)	0.173(1)	0.240(9)	0.248(10)	0.235(8)	0.176(2)	0.252(11)	0.216(7)	0.215(6)	0.209(5)	0.200(4)
	T2-PCK	0.103(5)	0.131(14)	0.125(10)	0.113(6)	0.097(1)	0.128(12)	0.130(13)	0.117(7)	0.102(4)	0.128(11)	0.119(9)	0.101(3)	0.097(2)	0.118(8)
	T2-SEP	0.130(8)	0.166(12)	0.170(13)	0.164(11)	0.129(7)	0.111(3)	0.109(2)	0.103(1)	0.127(6)	0.219(14)	0.141(9)	0.126(5)	0.122(4)	0.148(10)
	U41Q1	0.268(9)	0.211(3)	0.209(2)	0.194(1)	0.273(12)	0.223(4)	0.229(5)	0.241(6)	0.272(11)	0.266(8)	0.286(13)	0.269(10)	0.294(14)	0.265(7)
	U41Q2	0.178(8)	0.172(4)	0.173(5)	0.176(7)	0.184(11)	0.154(1)	0.157(2)	0.169(3)	0.182(10)	0.192(13)	0.190(12)	0.181(9)	0.192(14)	0.174(6)
	U42Q1	0.121(10)	0.103(6)	0.100(2)	0.101(4)	0.122(11)	0.100(1)	0.105(7)	0.103(5)	0.120(9)	0.100(3)	0.126(12)	0.134(13)	0.139(14)	0.113(8)
	U42Q2	0.175(6)	0.193(14)	0.192(13)	0.189(12)	0.171(1)	0.180(7)	0.175(5)	0.174(4)	0.173(2)	0.180(9)	0.173(3)	0.186(11)	0.185(10)	0.180(8)

Table 9: Metric value and ranking results of each uncertainty quantification methods with Deepseek-R1-Distill-Qwen-32B over 14 questions.

AUROC

Control	Question	CE	Embed_Eigen	Embed_GE	Embed_NAD	FSD	JS_Eigen	JS_GE	JS_NAD	MAR	NLI_DSE	NLI_Eigen	NLI_GE	NLI_NAD	Numset
shot=0, cot=0	ASAP	0.718(2)	0.613(10)	0.628(8)	0.678(5)	0.691(4)	0.533(12)	0.527(13)	0.546(11)	0.702(3)	0.500(14)	0.626(9)	0.646(7)	0.664(6)	0.720(1)
	Semeval	0.708(2)	0.653(7)	0.671(5)	0.711(1)	0.697(4)	0.582(12)	0.586(11)	0.589(10)	0.702(3)	0.500(14)	0.603(9)	0.568(13)	0.637(8)	0.664(6)
	T1-CK	0.689(3)	0.609(10)	0.614(9)	0.623(8)	0.682(5)	0.543(11)	0.528(12)	0.518(13)	0.686(4)	0.500(14)	0.672(6)	0.707(2)	0.720(1)	0.634(7)
	T1-DCI	0.614(10)	0.641(5)	0.650(4)	0.636(6)	0.614(10)	0.653(3)	0.661(2)	0.668(1)	0.614(10)	0.500(14)	0.623(8)	0.590(13)	0.629(7)	0.617(9)
	T1-PCK	0.509(11)	0.522(9)	0.529(8)	0.546(7)	0.509(11)	0.606(3)	0.599(4)	0.569(5)	0.509(11)	0.500(14)	0.563(6)	0.666(1)	0.661(2)	0.519(10)
	T1-SEP	0.710(3)	0.679(8)	0.680(7)	0.744(1)	0.707(4)	0.548(13)	0.550(12)	0.602(11)	0.712(2)	0.500(14)	0.655(10)	0.690(6)	0.694(5)	0.669(9)
	T2-CK	0.711(1)	0.687(5)	0.685(6)	0.646(8)	0.694(3)	0.569(12)	0.556(13)	0.579(11)	0.704(2)	0.500(14)	0.621(10)	0.676(7)	0.691(4)	0.643(9)
	T2-DCI	0.677(7)	0.592(13)	0.602(12)	0.637(8)	0.694(4)	0.690(5)	0.708(1)	0.701(2)	0.690(6)	0.500(14)	0.695(3)	0.624(11)	0.636(9)	0.635(10)
	T2-PCK	0.566(6)	0.548(10)	0.556(9)	0.566(7)	0.569(4)	0.483(12)	0.476(13)	0.436(14)	0.568(5)	0.500(11)	0.641(3)	0.654(2)	0.705(1)	0.563(8)
	T2-SEP	0.564(6)	0.547(7)	0.544(8)	0.499(14)	0.582(3)	0.522(11)	0.528(10)	0.537(9)	0.580(4)	0.500(13)	0.567(5)	0.644(2)	0.677(1)	0.506(12)
	U41Q1	0.618(4)	0.526(11)	0.528(10)	0.553(7)	0.629(2)	0.533(9)	0.567(5)	0.630(1)	0.625(3)	0.500(14)	0.534(8)	0.505(13)	0.517(12)	0.557(6)
	U41Q2	0.756(1)	0.549(9)	0.583(6)	0.643(5)	0.741(3)	0.462(14)	0.478(13)	0.519(11)	0.751(2)	0.500(12)	0.555(8)	0.546(10)	0.577(7)	0.710(4)
	U42Q1	0.756(2)	0.728(5)	0.752(3)	0.786(1)	0.727(6)	0.488(14)	0.491(13)	0.505(11)	0.741(4)	0.500(12)	0.661(8)	0.660(9)	0.704(7)	0.643(10)
	U42Q2	0.668(3)	0.541(13)	0.569(12)	0.651(4)	0.671(2)	0.601(9)	0.605(8)	0.608(7)	0.674(1)	0.500(14)	0.637(5)	0.569(11)	0.631(6)	0.586(10)
shot=0, cot=1	ASAP	0.686(5)	0.616(10)	0.633(9)	0.697(1)	0.692(2)	0.518(13)	0.528(12)	0.585(11)	0.691(4)	0.500(14)	0.643(8)	0.658(6)	0.692(3)	0.651(7)
	Semeval	0.656(5)	0.649(7)	0.685(2)	0.719(1)	0.639(9)	0.536(13)	0.542(12)	0.596(11)	0.643(8)	0.500(14)	0.668(4)	0.673(3)	0.650(6)	0.631(10)
	T1-CK	0.736(1)	0.675(9)	0.693(4)	0.683(8)	0.691(5)	0.607(13)	0.615(12)	0.621(11)	0.713(2)	0.500(14)	0.690(6)	0.641(10)	0.687(7)	0.702(3)
	T1-DCI	0.622(8)	0.665(5)	0.667(4)	0.653(6)	0.622(8)	0.693(2)	0.702(1)	0.690(3)	0.622(8)	0.500(14)	0.605(11)	0.573(13)	0.588(12)	0.629(7)
	T1-PCK	0.510(9)	0.504(10)	0.517(5)	0.518(4)	0.512(6)	0.512(7)	0.499(13)	0.489(14)	0.511(8)	0.500(12)	0.581(3)	0.647(1)	0.647(2)	0.502(11)
	T1-SEP	0.747(3)	0.629(10)	0.658(7)	0.731(4)	0.749(2)	0.562(11)	0.539(13)	0.555(12)	0.750(1)	0.500(14)	0.642(9)	0.667(6)	0.670(5)	0.643(8)
	T2-CK	0.732(3)	0.657(9)	0.681(7)	0.683(6)	0.707(5)	0.583(12)	0.568(13)	0.602(11)	0.717(4)	0.500(14)	0.643(10)	0.737(2)	0.748(1)	0.670(8)
	T2-DCI	0.721(2)	0.633(9)	0.648(8)	0.714(4)	0.719(3)	0.621(12)	0.661(7)	0.709(5)	0.724(1)	0.500(14)	0.622(11)	0.573(13)	0.627(10)	0.683(6)
	T2-PCK	0.602(8)	0.614(4)	0.612(5)	0.594(10)	0.605(6)	0.592(11)	0.572(12)	0.475(14)	0.603(7)	0.500(13)	0.666(2)	0.646(3)	0.733(1)	0.595(9)
	T2-SEP	0.558(5)	0.516(12)	0.516(13)	0.536(9)	0.574(3)	0.529(10)	0.548(7)	0.552(6)	0.574(3)	0.500(14)	0.521(11)	0.586(2)	0.636(1)	0.541(8)
	U41Q1	0.560(10)	0.587(7)	0.578(8)	0.605(4)	0.599(5)	0.711(2)	0.715(1)	0.704(3)	0.595(6)	0.500(14)	0.528(12)	0.549(11)	0.569(9)	0.503(13)
	U41Q2	0.762(1)	0.629(9)	0.632(8)	0.665(6)	0.759(3)	0.537(12)	0.525(13)	0.560(11)	0.761(2)	0.500(14)	0.675(5)	0.623(10)	0.660(7)	0.691(4)
	U42Q1	0.683(3)	0.620(8)	0.643(7)	0.662(4)	0.685(2)	0.538(12)	0.541(11)	0.500(14)	0.686(1)	0.500(13)	0.614(9)	0.643(6)	0.606(10)	0.644(5)
	U42Q2	0.715(6)	0.629(12)	0.653(11)	0.727(5)	0.736(4)	0.659(10)	0.675(9)	0.681(8)	0.740(3)	0.500(14)	0.750(2)	0.714(7)	0.766(1)	0.562(13)
shot=2, cot=1	ASAP	0.671(5)	0.577(10)	0.590(9)	0.615(8)	0.686(2)	0.461(14)	0.475(13)	0.550(11)	0.680(4)	0.500(12)	0.646(6)	0.681(3)	0.695(1)	0.626(7)
	Semeval	0.626(4)	0.589(9)	0.583(10)	0.605(6)	0.599(8)	0.508(11)	0.487(13)	0.464(14)	0.609(5)	0.500(12)	0.729(1)	0.626(3)	0.647(2)	0.601(7)
	T1-CK	0.725(7)	0.810(3)	0.823(2)	0.826(1)	0.626(13)	0.695(10)	0.688(11)	0.683(12)	0.701(9)	0.500(14)	0.753(5)	0.724(8)	0.753(6)	0.768(4)
	T1-DCI	0.591(8)	0.603(7)	0.607(6)	0.588(12)	0.591(8)	0.636(2)	0.625(4)	0.627(3)	0.591(8)	0.500(14)	0.623(5)	0.581(13)	0.653(1)	0.589(11)
	T1-PCK	0.587(6)	0.505(12)	0.524(9)	0.531(8)	0.593(5)	0.522(10)	0.519(11)	0.497(14)	0.596(4)	0.500(13)	0.632(3)	0.676(2)	0.695(1)	0.558(7)
	T1-SEP	0.804(3)	0.661(10)	0.666(9)	0.728(5)	0.815(1)	0.522(11)	0.507(12)	0.474(14)	0.813(2)	0.500(13)	0.684(8)	0.720(6)	0.686(7)	0.747(4)
	T2-CK	0.904(1)	0.799(10)	0.816(9)	0.869(4)	0.837(8)	0.661(13)	0.671(12)	0.714(11)	0.891(2)	0.500(14)	0.848(5)	0.840(7)	0.884(3)	0.846(6)
	T2-DCI	0.723(5)	0.606(13)	0.640(12)	0.721(6)	0.737(2)	0.678(10)	0.688(8)	0.727(4)	0.733(3)	0.500(14)	0.687(9)	0.694(7)	0.751(1)	0.643(11)
	T2-PCK	0.627(7)	0.594(9)	0.623(8)	0.663(3)	0.641(5)	0.541(12)	0.541(11)	0.506(13)	0.639(6)	0.500(14)	0.653(4)	0.709(1)	0.704(2)	0.568(10)
	T2-SEP	0.694(4)	0.583(12)	0.605(10)	0.677(5)	0.733(1)	0.698(3)	0.670(6)	0.669(7)	0.722(2)	0.488(14)	0.528(13)	0.599(11)	0.624(9)	0.641(8)
	U41Q1	0.673(4)	0.678(3)	0.669(6)	0.671(5)	0.692(1)	0.634(7)	0.611(8)	0.501(13)	0.687(2)	0.500(14)	0.527(12)	0.534(11)	0.560(10)	0.574(9)
	U41Q2	0.757(1)	0.698(5)	0.701(4)	0.697(6)	0.730(3)	0.538(13)	0.559(12)	0.612(10)	0.741(2)	0.500(14)	0.633(9)	0.601(11)	0.644(8)	0.696(7)
	U42Q1	0.659(4)	0.647(6)	0.650(5)	0.673(3)	0.676(1)	0.549(13)	0.558(11)	0.583(10)	0.674(2)	0.500(14)	0.552(12)	0.592(9)	0.621(7)	0.593(8)
	U42Q2	0.696(2)	0.564(13)	0.585(11)	0.625(9)	0.684(4)	0.653(7)	0.654(6)	0.573(12)	0.695(3)	0.500(14)	0.703(1)	0.649(8)	0.680(5)	0.619(10)

C-index

Control	Question	CE	Embed_Eigen	Embed_GE	Embed_NAD	FSD	JS_Eigen	JS_GE	JS_NAD	MAR	NLI_DSE	NLI_Eigen	NLI_GE	NLI_NAD	Numset
shot=0, cot=0	ASAP	0.633(2)	0.599(9)	0.606(6)	0.626(3)	0.613(5)	0.550(12)	0.543(13)	0.570(11)	0.621(4)	0.500(14)	0.591(10)	0.603(8)	0.603(7)	0.642(1)
	Semeval	0.708(2)	0.653(7)	0.671(5)	0.711(1)	0.697(4)	0.582(12)	0.586(11)	0.589(10)	0.702(3)	0.500(14)	0.603(9)	0.568(13)	0.637(8)	0.664(6)
	T1-CK	0.660(4)	0.585(10)	0.588(9)	0.593(8)	0.658(5)	0.524(11)	0.514(12)	0.504(13)	0.660(3)	0.500(14)	0.642(6)	0.673(2)	0.685(1)	0.610(7)
	T1-DCI	0.614(10)	0.641(5)	0.650(4)	0.636(6)	0.614(10)	0.653(3)	0.661(2)	0.668(1)	0.614(10)	0.500(14)	0.623(8)	0.590(13)	0.629(7)	0.617(9)
	T1-PCK	0.509(11)	0.522(9)	0.529(8)	0.546(7)	0.509(11)	0.606(3)	0.599(4)	0.569(5)	0.509(11)	0.500(14)	0.563(6)	0.666(1)	0.661(2)	0.519(10)
	T1-SEP	0.693(3)	0.663(7)	0.664(5)	0.725(1)	0.687(4)	0.535(13)	0.540(12)	0.594(11)	0.694(2)	0.500(14)	0.634(10)	0.657(8)	0.664(6)	0.652(9)
	T2-CK	0.705(1)	0.680(5)	0.678(6)	0.640(9)	0.687(3)	0.567(12)	0.556(13)	0.579(11)	0.697(2)	0.500(14)	0.616(10)	0.670(7)	0.683(4)	0.641(8)
	T2-DCI	0.573(7)	0.524(13)	0.530(12)	0.546(11)	0.593(5)	0.625(4)	0.640(1)	0.629(3)	0.584(6)	0.500(14)	0.630(2)	0.547(10)	0.549(9)	0.550(8)
	T2-PCK	0.566(6)	0.548(10)	0.556(9)	0.566(7)	0.569(4)	0.483(12)	0.476(13)	0.436(14)	0.568(5)	0.500(11)	0.641(3)	0.654(2)	0.705(1)	0.563(8)
	T2-SEP	0.564(6)	0.547(7)	0.544(8)	0.499(14)	0.582(3)	0.522(11)	0.528(10)	0.537(9)	0.580(4)	0.500(13)	0.567(5)	0.644(2)	0.677(1)	0.506(12)
	U41Q1	0.572(5)	0.536(9)	0.538(8)	0.560(7)	0.587(3)	0.565(6)	0.594(2)	0.656(1)	0.581(4)	0.500(13)	0.523(10)	0.488(14)	0.502(12)	0.519(11)
	U41Q2	0.715(1)	0.503(10)	0.535(8)	0.594(5)	0.707(3)	0.460(14)	0.475(13)	0.495(12)	0.713(2)	0.500(11)	0.550(7)	0.530(9)	0.566(6)	0.669(4)
	U42Q1	0.689(4)	0.715(3)	0.737(2)	0.762(1)	0.662(6)	0.490(14)	0.491(13)	0.499(12)	0.670(5)	0.500(11)	0.600(8)	0.598(9)	0.640(7)	0.596(10)
	U42Q2	0.634(3)	0.511(13)	0.536(11)	0.618(4)	0.647(1)	0.587(8)	0.590(7)	0.584(9)	0.644(2)	0.500(14)	0.601(5)	0.535(12)	0.598(6)	0.550(10)
shot=0, cot=1	ASAP	0.628(7)	0.581(11)	0.594(10)	0.657(1)	0.634(3)	0.528(13)	0.548(12)	0.597(9)	0.633(4)	0.500(14)	0.632(5)	0.631(6)	0.647(2)	0.604(8)
	Semeval	0.656(5)	0.649(7)	0.685(2)	0.719(1)	0.639(9)	0.536(13)	0.542(12)	0.596(11)	0.643(8)	0.500(14)	0.668(4)	0.673(3)	0.650(6)	0.631(10)
	T1-CK	0.694(1)	0.638(9)	0.657(6)	0.644(8)	0.658(5)	0.579(13)	0.582(11)	0.580(12)	0.674(2)	0.500(14)	0.672(3)	0.613(10)	0.649(7)	0.668(4)
	T1-DCI	0.622(8)	0.665(5)	0.667(4)	0.653(6)	0.622(8)	0.693(2)	0.702(1)	0.690(3)	0.622(8)	0.500(14)	0.605(11)	0.573(13)	0.588(12)	0.629(7)
	T1-PCK	0.510(9)	0.504(10)	0.517(5)	0.518(4)	0.512(6)	0.512(7)	0.499(13)	0.489(14)	0.511(8)	0.500(12)	0.581(3)	0.647(1)	0.647(2)	0.502(11)
	T1-SEP	0.709(1)	0.615(10)	0.640(5)	0.702(3)	0.701(4)	0.549(11)	0.530(13)	0.538(12)	0.704(2)	0.500(14)	0.617(9)	0.630(7)	0.632(6)	0.621(8)
	T2-CK	0.706(3)	0.639(9)	0.661(7)	0.663(6)	0.681(5)	0.580(12)	0.566(13)	0.594(11)	0.691(4)	0.500(14)	0.623(10)	0.718(2)	0.727(1)	0.647(8)
	T2-DCI	0.568(7)	0.531(11)	0.534(10)	0.579(5)	0.583(4)	0.593(3)	0.622(2)	0.646(1)	0.577(6)	0.500(13)	0.546(9)	0.499(14)	0.528(12)	0.554(8)
	T2-PCK	0.602(8)	0.614(4)	0.612(5)	0.594(10)	0.605(6)	0.592(11)	0.572(12)	0.475(14)	0.603(7)	0.500(13)	0.666(2)	0.646(3)	0.733(1)	0.595(9)
	T2-SEP	0.541(5)	0.500(12)	0.499(14)	0.519(9)	0.558(3)	0.512(10)	0.531(7)	0.534(6)	0.556(4)	0.500(12)	0.510(11)	0.570(2)	0.620(1)	0.526(8)
	U41Q1	0.481(12)	0.573(5)	0.568(6)	0.594(4)	0.522(7)	0.694(2)	0.696(1)	0.677(3)	0.517(9)	0.500(10)	0.476(13)	0.497(11)	0.518(8)	0.442(14)
	U41Q2	0.754(1)	0.623(9)	0.626(8)	0.664(5)	0.749(3)	0.532(12)	0.521(13)	0.576(11)	0.752(2)	0.500(14)	0.659(6)	0.614(10)	0.652(7)	0.685(4)
	U42Q1	0.681(2)	0.619(8)	0.643(6)	0.662(4)	0.677(3)	0.537(11)	0.537(12)	0.498(14)	0.681(1)	0.500(13)	0.604(9)	0.632(7)	0.593(10)	0.646(5)
	U42Q2	0.670(7)	0.606(12)	0.630(11)	0.695(5)	0.706(3)	0.647(10)	0.658(8)	0.649(9)	0.696(4)	0.500(14)	0.726(2)	0.674(6)	0.731(1)	0.516(13)
shot=2, cot=1	ASAP	0.621(5)	0.550(11)	0.557(9)	0.579(8)	0.632(3)	0.491(14)	0.503(12)	0.552(10)	0.629(4)	0.500(13)	0.615(6)	0.639(2)	0.644(1)	0.582(7)
	Semeval	0.626(4)	0.589(9)	0.583(10)	0.605(6)	0.599(8)	0.508(11)	0.487(13)	0.464(14)	0.609(5)	0.500(12)	0.729(1)	0.626(3)	0.647(2)	0.601(7)
	T1-CK	0.682(10)	0.778(3)	0.790(2)	0.794(1)	0.590(13)	0.711(7)	0.693(8)	0.677(11)	0.655(12)	0.500(14)	0.731(5)	0.686(9)	0.716(6)	0.734(4)
	T1-DCI	0.591(8)	0.603(7)	0.607(6)	0.588(12)	0.591(8)	0.636(2)	0.625(4)	0.627(3)	0.591(8)	0.500(14)	0.623(5)	0.581(13)	0.653(1)	0.589(11)
	T1-PCK	0.596(5)	0.477(14)	0.501(11)	0.524(8)	0.589(6)	0.516(9)	0.515(10)	0.487(13)	0.598(4)	0.500(12)	0.648(3)	0.676(2)	0.701(1)	0.567(7)
	T1-SEP	0.739(3)	0.643(8)	0.644(7)	0.689(5)	0.741(1)	0.542(11)	0.524(12)	0.492(14)	0.741(2)	0.500(13)	0.636(9)	0.651(6)	0.620(10)	0.693(4)
	T2-CK	0.885(1)	0.809(8)	0.822(7)	0.870(2)	0.808(9)	0.673(13)	0.690(12)	0.723(11)	0.870(2)	0.500(14)	0.832(5)	0.794(10)	0.851(4)	0.830(6)
	T2-DCI	0.653(9)	0.590(12)	0.612(11)	0.669(7)	0.681(5)	0.682(4)	0.697(3)	0.740(1)	0.673(6)	0.500(14)	0.653(8)	0.639(10)	0.698(2)	0.588(13)
	T2-PCK	0.614(8)	0.590(9)	0.616(7)	0.656(3)	0.629(5)	0.533(11)	0.532(12)	0.499(14)	0.626(6)	0.500(13)	0.640(4)	0.694(1)	0.689(2)	0.554(10)
	T2-SEP	0.646(7)	0.568(12)	0.589(10)	0.649(6)	0.688(4)	0.708(1)	0.693(3)	0.700(2)	0.674(5)	0.489(14)	0.506(13)	0.578(11)	0.600(8)	0.599(9)
	U41Q1	0.657(4)	0.664(3)	0.656(5)	0.656(5)	0.681(1)	0.626(7)	0.604(8)	0.496(14)	0.672(2)	0.500(13)	0.522(12)	0.524(11)	0.557(10)	0.560(9)
	U41Q2	0.706(1)	0.661(4)	0.661(4)	0.651(7)	0.676(3)	0.496(14)	0.526(12)	0.571(11)	0.685(2)	0.500(13)	0.623(8)	0.595(10)	0.621(9)	0.652(6)
	U42Q1	0.646(4)	0.640(6)	0.643(5)	0.665(1)	0.654(2)	0.542(12)	0.546(11)	0.564(10)	0.654(3)	0.500(14)	0.540(13)	0.587(9)	0.607(7)	0.594(8)
	U42Q2	0.640(2)	0.503(13)	0.523(12)	0.572(10)	0.636(5)	0.612(7)	0.616(6)	0.545(11)	0.637(4)	0.500(14)	0.674(1)	0.603(8)	0.637(3)	0.574(9)

AUARC

Control	Question	CE	Embed_Eigen	Embed_GE	Embed_NAD	FSD	JS_Eigen	JS_GE	JS_NAD	MAR	NLI_DSE	NLI_Eigen	NLI_GE	NLI_NAD	Numset
shot=0, cot=0	ASAP	0.595(2)	0.518(11)	0.532(9)	0.564(8)	0.585(4)	0.448(14)	0.454(13)	0.489(12)	0.591(3)	0.522(10)	0.566(7)	0.576(6)	0.584(5)	0.605(1)
	Semeval	0.752(2)	0.729(7)	0.740(5)	0.760(1)	0.750(4)	0.641(13)	0.657(9)	0.642(12)	0.751(3)	0.643(11)	0.652(10)	0.632(14)	0.676(8)	0.736(6)
	T1-CK	0.780(4)	0.731(10)	0.733(9)	0.741(8)	0.778(6)	0.708(11)	0.699(12)	0.687(13)	0.779(5)	0.672(14)	0.783(3)	0.795(2)	0.803(1)	0.770(7)
	T1-DCI	0.678(6)	0.673(10)	0.676(9)	0.691(4)	0.678(6)	0.712(3)	0.718(2)	0.737(1)	0.678(6)	0.635(14)	0.670(12)	0.647(13)	0.670(11)	0.680(5)
	T1-PCK	0.529(10)	0.524(13)	0.520(14)	0.553(7)	0.531(9)	0.572(4)	0.557(6)	0.570(5)	0.528(11)	0.532(8)	0.643(3)	0.690(2)	0.695(1)	0.528(12)
	T1-SEP	0.769(4)	0.753(7)	0.753(8)	0.776(1)	0.768(5)	0.656(14)	0.660(12)	0.693(11)	0.770(3)	0.658(13)	0.731(10)	0.762(6)	0.770(2)	0.744(9)
	T2-CK	0.756(3)	0.728(9)	0.736(7)	0.730(8)	0.750(5)	0.657(12)	0.650(13)	0.676(11)	0.753(4)	0.593(14)	0.742(6)	0.763(2)	0.777(1)	0.727(10)
	T2-DCI	0.588(6)	0.560(10)	0.570(9)	0.578(7)	0.594(4)	0.610(3)	0.628(1)	0.617(2)	0.592(5)	0.408(14)	0.558(11)	0.538(12)	0.535(13)	0.574(8)
	T2-PCK	0.721(7)	0.684(10)	0.686(9)	0.695(8)	0.722(6)	0.626(12)	0.620(13)	0.594(14)	0.722(5)	0.678(11)	0.770(3)	0.774(2)	0.802(1)	0.736(4)
	T2-SEP	0.712(6)	0.700(10)	0.712(7)	0.704(8)	0.715(4)	0.702(9)	0.682(12)	0.676(13)	0.714(5)	0.648(14)	0.748(3)	0.775(2)	0.799(1)	0.699(11)
	U41Q1	0.612(3)	0.527(14)	0.535(12)	0.567(7)	0.627(2)	0.532(13)	0.538(10)	0.607(4)	0.629(1)	0.537(11)	0.567(6)	0.552(9)	0.556(8)	0.575(5)
	U41Q2	0.844(2)	0.739(8)	0.757(6)	0.786(5)	0.842(3)	0.670(14)	0.678(13)	0.704(11)	0.845(1)	0.705(10)	0.735(9)	0.697(12)	0.745(7)	0.836(4)
	U42Q1	0.807(4)	0.816(3)	0.823(2)	0.831(1)	0.801(6)	0.642(13)	0.634(14)	0.643(12)	0.806(5)	0.697(11)	0.753(9)	0.735(10)	0.764(8)	0.769(7)
	U42Q2	0.786(1)	0.718(12)	0.731(11)	0.774(4)	0.784(2)	0.762(9)	0.763(7)	0.762(10)	0.781(3)	0.624(14)	0.773(5)	0.706(13)	0.762(8)	0.763(6)
shot=0, cot=1	ASAP	0.587(3)	0.511(11)	0.523(9)	0.574(5)	0.589(2)	0.457(14)	0.466(13)	0.511(10)	0.586(4)	0.491(12)	0.561(8)	0.573(6)	0.592(1)	0.572(7)
	Semeval	0.717(2)	0.666(10)	0.684(7)	0.702(6)	0.714(4)	0.598(13)	0.596(14)	0.608(12)	0.713(5)	0.649(11)	0.725(1)	0.680(8)	0.672(9)	0.715(3)
	T1-CK	0.792(1)	0.740(9)	0.744(8)	0.739(10)	0.777(5)	0.724(11)	0.718(12)	0.693(13)	0.786(2)	0.667(14)	0.783(4)	0.751(7)	0.769(6)	0.785(3)
	T1-DCI	0.765(2)	0.703(8)	0.693(10)	0.681(11)	0.765(2)	0.746(6)	0.751(5)	0.741(7)	0.765(2)	0.699(9)	0.626(13)	0.635(12)	0.610(14)	0.768(1)
	T1-PCK	0.492(14)	0.531(7)	0.541(4)	0.540(5)	0.492(12)	0.514(8)	0.506(10)	0.513(9)	0.492(13)	0.538(6)	0.616(3)	0.655(2)	0.661(1)	0.506(11)
	T1-SEP	0.768(3)	0.704(10)	0.721(7)	0.751(4)	0.770(1)	0.627(11)	0.610(13)	0.606(14)	0.769(2)	0.624(12)	0.720(8)	0.726(6)	0.731(5)	0.709(9)
	T2-CK	0.740(5)	0.731(8)	0.742(4)	0.745(3)	0.732(7)	0.681(12)	0.667(13)	0.681(11)	0.734(6)	0.571(14)	0.716(9)	0.758(2)	0.777(1)	0.715(10)
	T2-DCI	0.618(3)	0.558(10)	0.566(8)	0.603(5)	0.618(4)	0.564(9)	0.588(7)	0.623(1)	0.620(2)	0.446(14)	0.546(11)	0.506(13)	0.527(12)	0.599(6)
	T2-PCK	0.707(8)	0.708(5)	0.704(9)	0.683(11)	0.708(6)	0.692(10)	0.667(12)	0.587(14)	0.708(6)	0.651(13)	0.769(2)	0.759(3)	0.803(1)	0.716(4)
	T2-SEP	0.700(11)	0.719(6)	0.719(7)	0.719(4)	0.705(8)	0.698(12)	0.705(9)	0.721(3)	0.703(10)	0.628(14)	0.719(5)	0.755(2)	0.785(1)	0.693(13)
	U41Q1	0.536(11)	0.571(5)	0.571(6)	0.574(4)	0.546(8)	0.649(3)	0.651(2)	0.651(1)	0.545(9)	0.465(14)	0.543(10)	0.514(13)	0.549(7)	0.529(12)
	U41Q2	0.843(2)	0.775(8)	0.775(9)	0.782(7)	0.843(3)	0.717(13)	0.703(14)	0.721(11)	0.842(4)	0.719(12)	0.790(6)	0.765(10)	0.799(5)	0.846(1)
	U42Q1	0.794(1)	0.774(6)	0.778(5)	0.785(4)	0.793(3)	0.715(13)	0.724(11)	0.695(14)	0.794(2)	0.721(12)	0.738(9)	0.753(8)	0.733(10)	0.771(7)
	U42Q2	0.822(5)	0.775(11)	0.789(8)	0.819(6)	0.826(4)	0.768(12)	0.782(9)	0.781(10)	0.830(2)	0.628(14)	0.828(3)	0.807(7)	0.833(1)	0.731(13)
shot=2, cot=1	ASAP	0.576(5)	0.475(14)	0.484(11)	0.491(10)	0.580(3)	0.482(12)	0.477(13)	0.526(8)	0.580(4)	0.496(9)	0.564(7)	0.592(1)	0.587(2)	0.576(6)
	Semeval	0.739(5)	0.720(9)	0.716(11)	0.719(10)	0.731(8)	0.644(12)	0.624(13)	0.612(14)	0.733(7)	0.736(6)	0.808(1)	0.758(3)	0.760(2)	0.753(4)
	T1-CK	0.918(7)	0.925(5)	0.928(3)	0.929(2)	0.896(13)	0.914(9)	0.912(10)	0.906(12)	0.910(11)	0.834(14)	0.934(1)	0.917(8)	0.924(6)	0.925(4)
	T1-DCI	0.723(3)	0.708(8)	0.712(7)	0.699(10)	0.723(3)	0.706(9)	0.698(11)	0.698(12)	0.723(3)	0.649(14)	0.733(2)	0.694(13)	0.735(1)	0.721(6)
	T1-PCK	0.663(6)	0.593(10)	0.600(8)	0.600(9)	0.667(4)	0.558(14)	0.573(12)	0.566(13)	0.663(5)	0.586(11)	0.721(3)	0.749(2)	0.765(1)	0.651(7)
	T1-SEP	0.807(3)	0.733(7)	0.732(8)	0.755(5)	0.811(1)	0.598(11)	0.580(13)	0.554(14)	0.809(2)	0.595(12)	0.713(10)	0.742(6)	0.721(9)	0.788(4)
	T2-CK	0.962(1)	0.946(9)	0.949(8)	0.957(4)	0.952(5)	0.899(13)	0.905(12)	0.917(11)	0.961(2)	0.835(14)	0.951(6)	0.950(7)	0.959(3)	0.940(10)
	T2-DCI	0.716(4)	0.652(13)	0.673(12)	0.719(3)	0.727(1)	0.684(11)	0.697(7)	0.715(5)	0.720(2)	0.563(14)	0.693(8)	0.687(9)	0.713(6)	0.687(10)
	T2-PCK	0.758(7)	0.747(9)	0.755(8)	0.776(4)	0.759(6)	0.691(11)	0.687(12)	0.663(14)	0.759(5)	0.666(13)	0.790(3)	0.818(1)	0.812(2)	0.735(10)
	T2-SEP	0.782(4)	0.717(12)	0.726(11)	0.758(7)	0.795(1)	0.790(3)	0.777(5)	0.776(6)	0.791(2)	0.587(14)	0.707(13)	0.746(9)	0.757(8)	0.739(10)
	U41Q1	0.752(4)	0.755(3)	0.752(5)	0.751(6)	0.761(1)	0.743(7)	0.734(8)	0.678(11)	0.761(2)	0.690(10)	0.655(14)	0.662(13)	0.677(12)	0.730(9)
	U41Q2	0.812(1)	0.785(6)	0.789(4)	0.779(7)	0.803(3)	0.659(14)	0.673(13)	0.691(12)	0.807(2)	0.702(11)	0.714(10)	0.715(9)	0.731(8)	0.786(5)
	U42Q1	0.693(6)	0.702(3)	0.706(2)	0.715(1)	0.697(5)	0.620(12)	0.618(13)	0.646(8)	0.699(4)	0.597(14)	0.626(11)	0.632(10)	0.662(7)	0.641(9)
	U42Q2	0.764(4)	0.696(13)	0.704(12)	0.723(9)	0.762(5)	0.756(8)	0.760(6)	0.715(11)	0.768(3)	0.571(14)	0.787(1)	0.757(7)	0.772(2)	0.720(10)

AUERC

Control	Question	CE	Embed_Eigen	Embed_GE	Embed_NAD	FSD	JS_Eigen	JS_GE	JS_NAD	MAR	NLI_DSE	NLI_Eigen	NLI_GE	NLI_NAD	Numset
shot=0, cot=0	ASAP	0.179(6)	0.189(10)	0.185(9)	0.177(5)	0.182(8)	0.219(14)	0.215(13)	0.199(11)	0.180(7)	0.209(12)	0.177(4)	0.175(3)	0.175(2)	0.171(1)
	Semeval	0.238(2)	0.261(7)	0.250(5)	0.230(1)	0.240(4)	0.349(13)	0.333(9)	0.348(12)	0.239(3)	0.347(11)	0.338(10)	0.358(14)	0.314(8)	0.254(6)
	T1-CK	0.128(4)	0.159(9)	0.161(10)	0.156(8)	0.129(6)	0.169(11)	0.172(12)	0.187(13)	0.128(5)	0.197(14)	0.128(3)	0.121(2)	0.117(1)	0.135(7)
	T1-DCI	0.312(6)	0.317(10)	0.314(9)	0.299(4)	0.312(6)	0.278(3)	0.272(2)	0.253(1)	0.312(6)	0.355(14)	0.320(12)	0.343(13)	0.320(11)	0.310(5)
	T1-PCK	0.230(10)	0.233(13)	0.235(14)	0.218(7)	0.230(9)	0.209(4)	0.217(6)	0.210(5)	0.231(11)	0.229(8)	0.173(3)	0.150(2)	0.148(1)	0.231(12)
	T1-SEP	0.141(2)	0.157(8)	0.157(7)	0.142(4)	0.143(5)	0.222(14)	0.218(13)	0.192(11)	0.141(1)	0.216(12)	0.165(10)	0.146(6)	0.142(3)	0.162(9)
	T2-CK	0.121(3)	0.136(10)	0.133(7)	0.136(9)	0.124(5)	0.169(12)	0.172(13)	0.159(11)	0.122(4)	0.216(14)	0.128(6)	0.117(2)	0.111(1)	0.136(8)
	T2-DCI	0.355(6)	0.367(9)	0.358(7)	0.359(8)	0.349(4)	0.319(3)	0.306(1)	0.316(2)	0.352(5)	0.473(14)	0.370(11)	0.393(12)	0.403(13)	0.367(10)
	T2-PCK	0.134(7)	0.153(10)	0.152(9)	0.147(8)	0.134(6)	0.182(12)	0.185(13)	0.198(14)	0.134(5)	0.156(11)	0.110(3)	0.108(2)	0.094(1)	0.127(4)
	T2-SEP	0.139(6)	0.145(10)	0.139(7)	0.143(8)	0.137(4)	0.144(9)	0.154(12)	0.157(13)	0.138(5)	0.171(14)	0.121(3)	0.108(2)	0.095(1)	0.145(11)
	U41Q1	0.099(4)	0.113(12)	0.111(8)	0.103(5)	0.094(2)	0.107(7)	0.106(6)	0.089(1)	0.094(3)	0.112(11)	0.111(9)	0.113(13)	0.114(14)	0.111(10)
	U41Q2	0.037(2)	0.068(9)	0.063(7)	0.054(5)	0.037(3)	0.083(14)	0.079(12)	0.073(11)	0.037(1)	0.080(13)	0.063(8)	0.072(10)	0.060(6)	0.040(4)
	U42Q1	0.050(4)	0.043(3)	0.041(2)	0.040(1)	0.052(6)	0.090(12)	0.094(14)	0.093(13)	0.050(5)	0.067(8)	0.072(10)	0.077(11)	0.069(9)	0.058(7)
	U42Q2	0.052(2)	0.070(12)	0.066(11)	0.056(4)	0.051(1)	0.056(5)	0.056(6)	0.057(7)	0.052(3)	0.086(14)	0.060(8)	0.074(13)	0.060(9)	0.060(10)
shot=0, cot=1	ASAP	0.174(7)	0.200(11)	0.193(10)	0.168(2)	0.172(4)	0.212(13)	0.206(12)	0.186(9)	0.173(6)	0.219(14)	0.173(5)	0.171(3)	0.167(1)	0.180(8)
	Semeval	0.273(2)	0.324(10)	0.306(7)	0.288(6)	0.276(4)	0.392(13)	0.394(14)	0.382(12)	0.277(5)	0.341(11)	0.265(1)	0.310(8)	0.318(9)	0.275(3)
	T1-CK	0.118(3)	0.146(9)	0.143(7)	0.148(10)	0.125(5)	0.159(11)	0.165(12)	0.178(13)	0.121(4)	0.198(14)	0.115(1)	0.144(8)	0.136(6)	0.116(2)
	T1-DCI	0.225(2)	0.287(8)	0.297(10)	0.309(11)	0.225(2)	0.244(6)	0.239(5)	0.249(7)	0.225(2)	0.291(9)	0.364(13)	0.355(12)	0.380(14)	0.222(1)
	T1-PCK	0.249(14)	0.230(7)	0.224(4)	0.225(5)	0.249(12)	0.238(8)	0.242(10)	0.239(9)	0.249(13)	0.226(6)	0.187(3)	0.167(2)	0.165(1)	0.242(11)
	T1-SEP	0.140(1)	0.168(8)	0.161(5)	0.144(4)	0.141(3)	0.216(11)	0.227(12)	0.234(14)	0.141(2)	0.228(13)	0.168(7)	0.171(9)	0.163(6)	0.172(10)
	T2-CK	0.144(6)	0.142(5)	0.136(4)	0.134(3)	0.148(9)	0.158(12)	0.164(13)	0.158(11)	0.147(8)	0.225(14)	0.145(7)	0.124(2)	0.114(1)	0.155(10)
	T2-DCI	0.336(7)	0.373(10)	0.369(9)	0.336(6)	0.332(4)	0.330(3)	0.310(2)	0.291(1)	0.333(5)	0.405(12)	0.378(11)	0.413(14)	0.408(13)	0.351(8)
	T2-PCK	0.142(8)	0.141(5)	0.143(9)	0.154(11)	0.141(6)	0.149(10)	0.162(12)	0.201(14)	0.141(6)	0.169(13)	0.111(2)	0.116(3)	0.094(1)	0.137(4)
	T2-SEP	0.145(11)	0.135(6)	0.135(7)	0.135(5)	0.143(8)	0.146(12)	0.143(9)	0.135(4)	0.144(10)	0.180(14)	0.134(3)	0.117(2)	0.102(1)	0.148(13)
	U41Q1	0.125(11)	0.103(6)	0.103(5)	0.102(4)	0.122(7)	0.084(2)	0.084(1)	0.085(3)	0.122(8)	0.134(14)	0.125(10)	0.131(13)	0.123(9)	0.128(12)
	U41Q2	0.037(2)	0.055(8)	0.055(9)	0.052(6)	0.038(3)	0.070(12)	0.074(14)	0.067(11)	0.038(4)	0.071(13)	0.053(7)	0.058(10)	0.049(5)	0.037(1)
	U42Q1	0.052(1)	0.056(7)	0.055(5)	0.053(4)	0.053(3)	0.073(12)	0.073(11)	0.080(14)	0.052(2)	0.070(10)	0.070(9)	0.068(8)	0.074(13)	0.056(6)
	U42Q2	0.042(4)	0.052(10)	0.049(7)	0.042(3)	0.040(1)	0.053(12)	0.050(8)	0.052(9)	0.040(2)	0.085(14)	0.045(6)	0.052(11)	0.044(5)	0.067(13)
shot=2, cot=1	ASAP	0.187(6)	0.223(14)	0.219(13)	0.216(11)	0.186(5)	0.208(9)	0.212(10)	0.192(8)	0.186(4)	0.218(12)	0.180(3)	0.172(1)	0.176(2)	0.188(7)
	Semeval	0.251(5)	0.270(9)	0.274(11)	0.271(10)	0.259(8)	0.346(12)	0.366(13)	0.378(14)	0.257(7)	0.254(6)	0.182(1)	0.232(3)	0.230(2)	0.237(4)
	T1-CK	0.036(9)	0.032(5)	0.031(3)	0.031(2)	0.046(13)	0.033(7)	0.035(8)	0.039(11)	0.040(12)	0.074(14)	0.027(1)	0.037(10)	0.033(6)	0.032(4)
	T1-DCI	0.267(3)	0.282(8)	0.278(7)	0.291(10)	0.267(3)	0.284(9)	0.292(11)	0.292(12)	0.267(3)	0.341(14)	0.257(2)	0.296(13)	0.255(1)	0.269(6)
	T1-PCK	0.162(5)	0.205(11)	0.199(10)	0.196(9)	0.162(4)	0.220(14)	0.210(12)	0.214(13)	0.163(6)	0.194(8)	0.132(3)	0.121(2)	0.112(1)	0.165(7)
	T1-SEP	0.129(3)	0.159(6)	0.160(7)	0.150(5)	0.128(1)	0.228(11)	0.244(12)	0.271(13)	0.129(2)	0.275(14)	0.195(9)	0.184(8)	0.196(10)	0.136(4)
	T2-CK	0.018(2)	0.022(6)	0.021(5)	0.016(1)	0.023(7)	0.044(13)	0.040(12)	0.036(11)	0.018(3)	0.094(14)	0.025(8)	0.028(9)	0.021(4)	0.035(10)
	T2-DCI	0.192(9)	0.216(13)	0.206(10)	0.180(5)	0.177(4)	0.175(3)	0.165(2)	0.150(1)	0.185(6)	0.270(14)	0.191(8)	0.206(11)	0.187(7)	0.207(12)
	T2-PCK	0.116(7)	0.121(9)	0.117(8)	0.106(4)	0.115(5)	0.149(11)	0.151(12)	0.163(13)	0.116(6)	0.174(14)	0.100(3)	0.086(1)	0.089(2)	0.130(10)
	T2-SEP	0.107(6)	0.133(12)	0.129(10)	0.114(7)	0.099(4)	0.091(1)	0.094(3)	0.093(2)	0.102(5)	0.199(14)	0.139(13)	0.120(9)	0.114(8)	0.130(11)
	U41Q1	0.059(5)	0.059(3)	0.059(4)	0.060(6)	0.057(1)	0.061(7)	0.063(8)	0.077(11)	0.057(2)	0.074(10)	0.083(14)	0.082(13)	0.077(12)	0.065(9)
	U41Q2	0.044(1)	0.051(6)	0.050(5)	0.053(7)	0.048(3)	0.084(14)	0.079(13)	0.074(12)	0.046(2)	0.068(11)	0.065(10)	0.065(9)	0.062(8)	0.050(4)
	U42Q1	0.078(5)	0.075(3)	0.074(2)	0.071(1)	0.078(6)	0.099(11)	0.100(12)	0.094(9)	0.077(4)	0.103(14)	0.101(13)	0.096(10)	0.090(8)	0.090(7)
	U42Q2	0.056(5)	0.073(13)	0.072(12)	0.066(9)	0.057(6)	0.057(8)	0.056(4)	0.066(10)	0.056(3)	0.090(14)	0.049(1)	0.057(7)	0.053(2)	0.066(11)

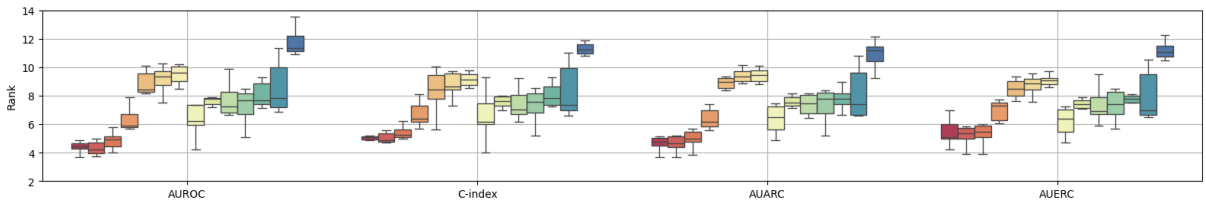
```
As a virtual teaching assistant, your role
involves assisting in the evaluation of
student assignment submissions. Please be
impartial and only use the grading guidelines,
the student`s submission, and the task in your
evaluation.
Analyze the submission and evaluate how well
the student`s submission answers the task.
Grade this submission of a student in 10th,
with a score in {grade_range}
# COT Command:
  In this task, you should use
chain-of-thought to solve the problem
step by step before giving the response.
# Question (the students will work on):
  {question}
# Prompt
  {grading_rubrics}
# Submission (to be evaluated):
  {student_answer}
# Output Format:
  `text (reasons for your decisions)...  ##
Final: final_score`.
```

Figure 9: Prompt template for Zero-shot with COT Setting.

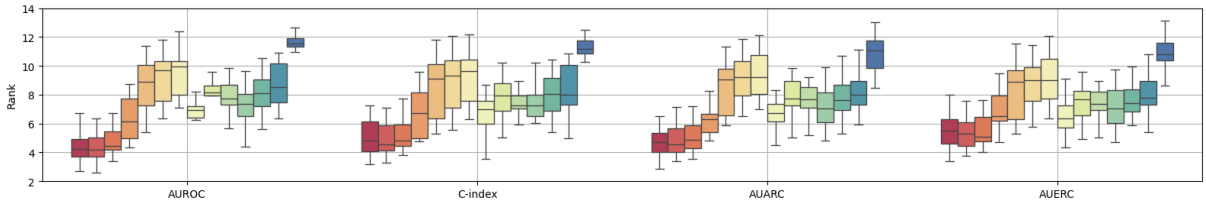
95% stability threshold after approximately 10 repeats.

```
As a virtual teaching assistant, your role
involves assisting in the evaluation of
student assignment submissions. Please be
impartial and only use the grading guidelines,
the student`s submission, and the task in your
evaluation.
Analyze the submission and evaluate how well
the student`s submission answers the task.
Grade this submission of a student in 10th,
with a score in {grade_range}
# COT Command:
  In this task, you should use
chain-of-thought to solve the problem
step by step before giving the response.
# Question (the students will work on):
  {question}
# Prompt
  {grading_rubrics}
# Few-shot Information (examples to follow):
  {example_pairs_of_answer_grading}
# Submission (to be evaluated):
  {student_answer}
# Output Format:
  `text (reasons for your decisions)...  ##
Final: final_score`.
```

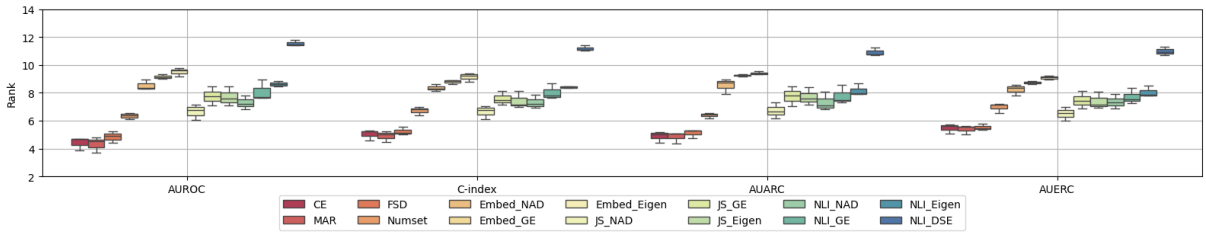
Figure 10: Prompt template for Few-shot with COT Setting.



(a) Models.

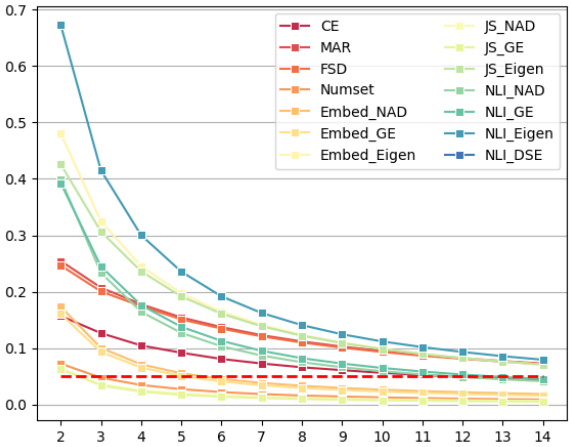


(b) Questions.

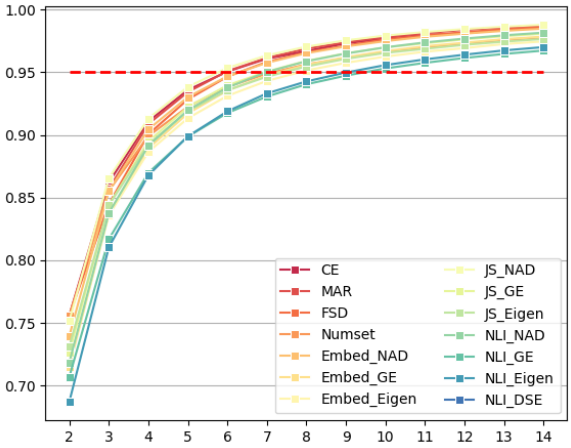


(c) Generation strategies.

Figure 11: Distribution of various in-group average rank for each model, question and generation strategies variants.



(a) Diffratio.



(b) Spearmanr.

Figure 12: Aggregated stability metrics, DiffRatio and Spearmanr, as functions of the number of repeats across different UQ methods, models, questions, and generation strategies.