Is This Collection Worth My LLM's Time? Automatically Measuring Information Potential in Text Corpora

Anonymous ACL submission

Abstract

As large language models (LLMs) converge in capabilities, progress increasingly depends on integrating novel, high-value information sources. Yet deciding which text collections justify the costs of digitization and integration remains difficult. We introduce an automated pipeline that estimates a collection's information potential without retraining. The method generates multiple-choice questions (MCQs) from texts and compares model performance with and without access to the source; the gap serves as a proxy for novel knowledge. We validate this approach on five datasets—EPFL PhD manuscripts, Venetian historical records, two Wikipedia subsets, and a synthetic baseline. Results show that the pipeline reliably distinguishes high-value from redundant corpora, offering a practical tool for prioritizing data acquisition and integration.

1 Introduction

Large language models (LLMs) show convergent capabilities as they scale, largely independent of architecture (Huh et al., 2024). Combined with scaling laws (Kaplan et al., 2020) and the shift toward data-centric AI (Zha et al., 2025), this suggests that future progress depends less on model design and more on integrating high-quality, novel information sources. Yet acquiring such data is costly: digitization, preprocessing, and fine-tuning require substantial resources.

This raises a central question: how can we assess whether a text collection meaningfully expands an LLM's knowledge before investing in its integration? Existing strategies typically require expensive retraining or fine-tuning, while retrieval-augmented generation (RAG) (Lewis et al., 2020) mitigates some costs but still requires careful knowledge curation (Kandpal et al., 2023). Practical methods for estimating the value of new corpora remain limited.

We propose an automated pipeline that estimates a collection's information potential without training. The method generates MCQs from text chunks, applies a two-stage quality filter—(i) context–answer alignment using ROUGE-L/Jaccard and (ii) distractor plausibility via embedding cosine similarity—and then compares LLM performance with and without access to the source. The performance gap serves as a proxy for novel information.

To validate this approach, we test three distinct datasets: (1) EPFL PhD manuscripts, containing specialized academic knowledge, (2) Wikipedia articles, representing widely available content, and (3) a synthetic baseline of model-generated text. Importantly, the pipeline is dataset-agnostic and can be applied to any domain, making it a practical tool for prioritizing corpora for fine-tuning or RAG systems.

2 Related Work

Information Gain. The challenge of efficiently selecting new information sources for LLMs can be viewed through the lens of optimal experiment design, a field pioneered by Fedorov et al. (1972). This framework emphasizes maximizing information gain while being strategic about resource allocation – a particularly relevant consideration given the costs associated with integrating new data into LLM systems. Information gain itself has been conceptualized across various fields: in information theory, it relates to reductions in algorithmic information content (Cover et al., 1989); in machine learning, it quantifies a feature's contribution to model performance (Odhiambo Omuya et al., 2021); and in cognitive science, it represents uncertainty reduction in our experience of the world (Damiano et al., 2021). While these theoretical frameworks provide valuable insights, they have not been previously applied to the specific challenge of evaluating the potential value

of text collections for enhancing LLM knowledge. Our work bridges this gap by proposing a practical, MCQ-based approach that quantifies information gain by measuring an LLM's ability to answer questions about a text collection with and without access to the source material.

Knowledge Detection in LLMs. Prior research has developed several methods to analyze how LLMs process and retain textual information. Work on memorization (Hartmann et al., 2023; Shi et al., 2024) and data contamination (Yax et al., 2024; Golchin and Surdeanu, 2024) focuses on identifying verbatim recall of training data, while hallucination detection (Farquhar et al., 2024) aims to identify when models generate false information. Research on novelty detection has primarily focused on linguistic and semantic novelty (McCoy et al., 2023; Lin et al., 2024), with less attention paid to factual novelty. While these approaches provide valuable insights into model behavior, they are retrospective – analyzing what models have already learned or memorized. In contrast, our work takes a prospective approach, developing metrics to evaluate the potential value of new information sources before investing in their integration into LLM systems.

3 Methods

Multiple Choice Questions (MCQs) are a well-established tool for knowledge assessment, supported by research in cognitive science and educational psychology (Thomas M. Haladyna and Rodriguez, 2002). Their four-option format, consisting of one correct answer and three distractors, offers an optimal balance between assessment reliability and cognitive load (Vyas and Supe, 2007). MCQs are particularly valuable for automated evaluation as they provide objective correctness measures while efficiently testing understanding across diverse topics (Oc and Hassen, 2024). For LLM evaluation, they offer the added advantage of constraining the output space to a finite set of options, eliminating ambiguity in grading and enabling precise comparisons across conditions.

3.1 MCQ Generation

The input text is first divided into manageable chunks of 2000 words to ensure consistent context length across questions. These chunks serve as the basis for LLM-generated MCQs, where each question is crafted to test understanding of specific information within the chunk. The exact prompts used for generation can be found in table 1 of the Supplementary Methods.

3.2 MCQ Quality Filtering (Two-Stage)

To avoid trivial or ambiguous items, we apply complementary filters that test (a) grounding in the source and (b) distractor plausibility:

(a) Context–Answer Alignment. For each MCQ, we require the correct answer g to be better aligned with its source chunk c than any distractor d_i. We compute surface-level similarity via Jaccard and ROUGE-L (Lin, 2004):

$$\min_i \left[\text{sim}(c, g) - \text{sim}(c, d_i)\right] \quad , i = 1, 2, 3 \qquad (1)$$

where $\text{sim} \in \{\text{Jaccard}, \text{ROUGE-L}\}$. Items failing either metric are discarded, removing questions whose "correct" answer is not actually supported by the text, or whose distractors inadvertently match the context.

(b) Distractor Plausibility. To prevent elimination-by-obviousness while keeping answers unambiguous, we enforce semantic closeness (but not identity) between g and at least one distractor using cosine similarity over sentence embeddings using the state-of-the-art NV-Embed-v2 (de Souza P. Moreira et al., 2024; Lee et al., 2024):

$$\max_i \left[\text{cos-sim}(g, d_i)\right] \quad , i = 1, 2, 3 \qquad (2)$$

which retains items where distractors are "nearby" enough to be plausible but not paraphrases of the correct answer. Thresholds are set by percentiles (e.g., 40–60th) computed on the pool of generated items, trading coverage for quality.

3.3 Evaluation (Position-Debiased, With/Without Context)

Each filtered MCQ is evaluated in two conditions:

1. **Direct (no context):** The model answers $(q, \{A, B, C, D\})$ without access to the chunk.
2. **With context:** The model receives the chunk c plus the same $(q, \{A, B, C, D\})$.

To mitigate position bias, we evaluate each item four times, rotating the correct option through A, B, C, and D while shuffling distractors. A question

is marked correct in a condition only if the model selects the correct option for all four rotations (a conservative criterion that reduces spurious gains).

We define the information potential (IP) of a collection as the normalized improvement in correctness when adding context:

$$IP = \frac{C_{\text{context}} - C_{\text{direct}}}{|\mathcal{Q}| - (I_{\text{context}} + I_{\text{direct}})}, \quad (3)$$

where C_{context} and C_{direct} count questions answered correctly (with the "4x" position control) with and without context, and $I_{\text{context}}, I_{\text{direct}}$ count incorrect ones. The denominator excludes items missed in both settings to focus on questions that are answerable in at least one condition.

3.4 Data

We evaluate five collections with distinct expected novelty profiles:

- **EPFL PhD manuscripts:** 177 theses with specialized, recent research likely underrepresented in pretraining corpora.
- **Venetian historical records:** Public-domain books on Venice's urban history, many newly digitized and rare online.
- **Wikipedia (EPFL-related) and Wikipedia (Venice-related):** Articles fetched via the Wikipedia API using manuscript titles and Venetian themes as seeds; expected to overlap with pretraining data.
- **Synthetic baseline:** Model-generated chunks on the same themes to establish a lower-bound novelty reference.

All corpora are chunked uniformly; MCQs, filtering metadata, and prompts are released with this submission to support reproducibility.

4 Results

Across five corpora and two models, our pipeline (i) increases MCQ quality via complementary similarity filters, (ii) assigns higher information potential (IP) to specialized/rare collections than to widely available ones, and (iii) yields consistent cross-model patterns. Figure 1 illustrates the impact of different thresholding methods, while Figure 2 summarizes correctness with/without context and the resulting IP per dataset/model.

4.1 Effectiveness of Similarity Thresholding

Cosine similarity. Increasing the cutoff reduces performance without context (NC 4x) by about 10% between 0 and the 50th percentile, while performance with context (WC 4x) stays stable—indicating higher difficulty for no-context answers without hurting context-based performance.

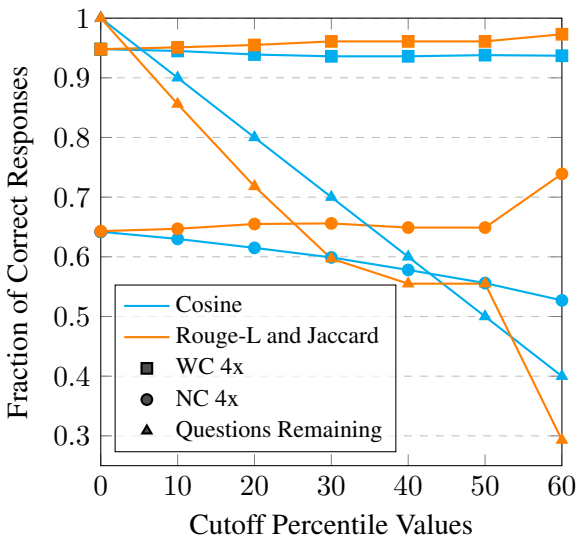


Figure 1: **Llama 3 70B Performance on EPFL Dataset Across Different Cutoff Percentages.** (with Separated Thresholding and Number of MCQs for the Corresponding Percentiles). **NC 4x:** Evaluation with no context correct for all 4 bias mitigation evaluations of each question; **WC 4x:** Evaluation with context correct for all 4 bias mitigation evaluations of each question. **Questions Remaining:** Fraction of the original number of questions still remaining for the given cutoff threshold. **Cosine:** Only cosine thresholding applied. **Rouge-L and Jaccard:** Only Rouge-L and Jaccard thresholding applied.

ROUGE-L and Jaccard. In contrast, these filters slightly improve WC 4x accuracy (+2% up to the 50th percentile) while NC 4x remains stable, suggesting improved question clarity rather than difficulty. However, combined percentile filtering sharply reduces the number of retained questions between the 50th and 60th percentiles.

Comparison. Cosine thresholding mainly increases difficulty, whereas ROUGE-L and Jaccard refine clarity at the cost of coverage.

4.2 Information Potential Across Corpora

Limited-Access Knowledge Collections (IP: 0.211–0.265). EPFL PhD manuscripts and Venetian historical records yield the highest information potential, driven by large gaps between with-context (WC) and no-context (NC)

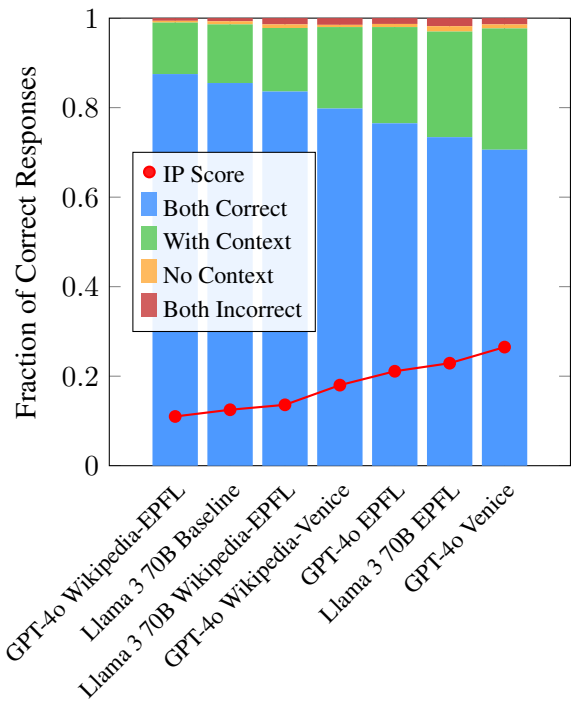


Figure 2: **Information Potential Analysis Across Datasets and Models.** Stacked barplot showing correct response overlap between context-free and context-provided conditions. Overlaying the barplot is a line plot showing the Information Potential (IP) scores. Higher IP scores indicate greater novel information content, with PhD manuscripts (EPFL) showing consistently higher IP (0.211-0.229) compared to Wikipedia (0.110-0.136) and synthetic baseline (0.125). Both open and closed-source models exhibit similar patterns despite architectural differences.

performance. WC accuracy is consistently $> 97\%$ while NC sits around 70–73%, producing IP $\approx$ 0.211–0.265. This suggests these collections contain knowledge that is hard to retrieve from model priors and thus valuable for integration (Fig. 2).

Wikipedia Datasets (IP: 0.110–0.180). Both EPFL-related and Venice-related Wikipedia subsets show moderate IP with smaller WC–NC gaps (WC $\sim$ 97% vs. NC $\sim$ 82–84%). The attenuated gain indicates substantial overlap with pretraining data or readily retrievable priors, aligning with their broad web availability (Fig. 2).

Synthetic Baseline (IP: 0.125). Model-generated text forms a lower bound: WC remains near ceiling ($\sim$ 99%) and NC is high ($\sim$ 85–86%), yielding a small but nonzero IP $\approx$ 0.125. Together with the 4$\times$ rotation consistency checks (Fig. 6), these trends support IP as a stable proxy for identifying corpora with genuinely novel, harder-to-retrieve information.

4.3 Comparison between open and closed source model

Both Llama 3 70B and GPT-4o show similar dataset rankings for information potential: EPFL manuscripts consistently score about twice as high as Wikipedia. GPT-4o yields slightly lower IP than Llama 70B, suggesting stronger baseline knowledge retention, but both models reach near-perfect accuracy ($>$ 97%) with context, confirming robust comprehension when relevant text is provided.

A broader survey of models could clarify how architecture and scale shape IP, but here our focus is on validating it as a general metric across different model families.

5 Conclusion

This work introduced a lightweight pipeline to estimate the information potential (IP) of text collections for LLMs without retraining or fine-tuning. Our contributions are:

1. **Evaluation pipeline:** Automated MCQ generation, filtering, and position-debiased testing provide a scalable measure of collection value.
2. **Empirical validation:** Results align with expectations—IP rises from synthetic baseline (0.125) to Wikipedia (0.136–0.180) to PhD manuscripts (0.229) and Venetian records (0.265). Larger models (GPT-4o) show lower IP than smaller ones (Llama 3 70B), consistent with stronger prior knowledge.

By highlighting where LLMs benefit most from external context, our method offers a practical screening tool for prioritizing digitization and integration of specialized corpora, from recent academic research to rare historical sources.

6 Limitations

Practical Considerations. Our pipeline generates MCQs before filtering; integrating quality checks directly into generation could improve efficiency. A further challenge is verifying whether corpora truly lie outside LLM training data, given opaque and ever-growing pretraining sets. Finally, our method identifies promising collections but does

not address how to integrate them into models without disrupting existing knowledge (Moiseev et al., 2022; Liu et al., 2024).

Evaluation Scope. This study omits human annotation and pretraining experiments to keep the method lightweight. Instead, we rely on safeguards such as similarity filtering, position-debiased testing, and cross-model evaluation. While fine-tuning or human validation could strengthen findings, our results already align with intuitive knowledge gaps (e.g., low IP for synthetic, high IP for specialized corpora).

Broader Perspective. The method requires digitized text but supports sampling-based evaluation—e.g., testing a few chapters before digitizing entire collections—making it useful for resource-constrained projects. More generally, IP highlights areas where models lack or fail to retrieve knowledge, raising broader questions about the line between knowledge possession and competence, and offering a practical complement to theoretical analyses of internal representations.

References

Thomas M. Cover, Peter Gács, and Robert M. Gray. 1989. Kolmogorov's contributions to information theory and algorithmic complexity. *Annals of Probability*, 17(3):840–865.

Claudia Damiano, Sander Van de Cruys, Yannick Boddez, Magdalena Król, Lore Goetschalckx, and Johan Wagemans. 2021. Visual affects: Linking curiosity, aha-erlebnis, and memory through information gain. In *Vision Sciences Society Annual Meeting Abstracts*.

Gabriel de Souza P. Moreira, Radek Osmulski, Mengyao Xu, Ronay Ak, Benedikt Schifferer, and Even Oldridge. 2024. Nv-retriever: Improving text embedding models with effective hard-negative mining. *Preprint*, arXiv:2407.15831.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630.

V. V. Fedorov, W. J. Studden, and E. M. Klimko. 1972. *Theory of Optimal Experiments*. Academic Press, New York, NY.

Shahriar Golchin and Mihai Surdeanu. 2024. Data contamination quiz: A tool to detect and estimate contamination in large language models. *Preprint*, arXiv:2311.06233.

Valentin Hartmann, Anshuman Suri, Vincent Bindschaedler, David Evans, Shruti Tople, and Robert West. 2023. Sok: Memorization in general-purpose large language models. *Preprint*, arXiv:2310.18362.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. 2024. The platonic representation hypothesis. *Preprint*, arXiv:2405.07987.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. Large language models struggle to learn long-tail knowledge. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 15696–15707. PMLR.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *Preprint*, arXiv:2001.08361.

Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2024. Nv-embed: Improved techniques for training llms as generalist embedding models. *Preprint*, arXiv:2405.17428.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Ethan Lin, Zhiyuan Peng, and Yi Fang. 2024. Evaluating and enhancing large language models for novelty assessment in scholarly publications. *Preprint*, arXiv:2409.16605.

Yantao Liu, Zijun Yao, Xin Lv, Yuchen Fan, Shulin Cao, Jifan Yu, Lei Hou, and Juanzi Li. 2024. Untangle the KNOT: Interweaving conflicting knowledge and reasoning skills in large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17186–17204, Torino, Italia. ELRA and ICCL.

R. McCoy, Paul Smolensky, Tal Linzen, Jianfeng Gao, and Celikyilmaz Asli. 2023. How much do language models copy from their training data? evaluating linguistic novelty in text generation using raven. *Transactions of the Association for Computational Linguistics*, 11:652–670.

Fedor Moiseev, Zhe Dong, Enrique Alfonseca, and Martin Jaggi. 2022. SKILL: Structured knowledge infusion for large language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1581–1588, Seattle, United States. Association for Computational Linguistics.

Yusuf Oc and Hela Hassen. 2024. Comparing the effectiveness of multiple-answer and single-answer multiple-choice questions in assessing student learning. *Marketing Education Review*, 0(0):1–14.

Erick Odhiambo Omuya, George Onyango Okeyo, and Michael Waema Kimwele. 2021. Feature selection for classification using principal component analysis and information gain. *Expert Systems with Applications*, 174:114765.

Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. 2024. Detecting pretraining data from large language models. *Preprint*, arXiv:2310.16789.

Steven M. Downing Thomas M. Haladyna and Michael C. Rodriguez. 2002. A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education*, 15(3):309–333.

Rashmi Vyas and Avinash Supe. 2007. Multiple choice questions: A literature review on the optimal number of options. *The National medical journal of India*, 21:130–3.

Nicolas Yax, Pierre-Yves Oudeyer, and Stefano Palminteri. 2024. Assessing contamination in large language models: Introducing the logprober method. *Preprint*, arXiv:2408.14352.

Daochen Zha, Zaid Pervaiz Bhat, Kwei-Herng Lai, Fan Yang, Zhimeng Jiang, Shaochen Zhong, and Xia Hu. 2025. Data-centric artificial intelligence: A survey. *ACM Comput. Surv.*, 57(5).

A Appendix

A.1 Supplementary Methods

A.1.1 LLM Prompts

Task	LLM Prompt
Multiple-Choice Question Generation	From the following piece of a scientific PhD manuscript: 'TEXT_HERE' Design a multiple-choice question with four answers: 'A', 'B', 'C', 'D'. Please provide the correct answer. The question needs to be difficult, but answers should not be ambiguous. Start the question with ['QUESTION'] and the answers with 'A', 'B', 'C', 'D'. Be concise! Please generate a total of 10 MCQs. Avoid references to the manuscript itself (e.g., do not use phrases like 'according to the text,' 'as stated in the manuscript,' or 'based on the passage' etc.). Use the following format: '[QUESTION] <question> A) <option A> B) <option B> C) <option C> D) <option D> Correct answer: <correct answer letter>) <correct answer>'
Multiple Choice Question Answer Generation	For the following multiple choice question: 'QUESTION_TEXT_HERE' Please write which answer option (A, B, C, or D) is the correct one. Answer in the following format: 'Correct answer: <answer letter>.'
Context-Based Multiple Choice Question Answer Generation	Using the information of the following passage: 'PASSAGE_TEXT_HERE' Answer the following multiple-choice question: 'QUESTION_TEXT_HERE' Please write which answer option (A, B, C, or D) is the correct one. Answer in the following format: 'Correct answer: <answer letter>.'
Baseline Subtopic List Generation	For the following topic: 'TOPIC_HERE' Please generate a list of 5 subtopics that could be used to create a comprehensive PhD manuscript about this topic. List them in order and number them in the following format: '1) <write subtopic 1 here> 2) <write subtopic 2 here> 3) <write subtopic 3 here> 4) <write subtopic 4 here> 5) <write subtopic 5 here> <end>'
Baseline Chapter Generation	For a scientific manuscript with the following title: 'MANUSCRIPT_TITLE_HERE' Please generate a comprehensive chapter that covers the following subtopic: 'SUBTOPIC_HERE'. Aim for around 600 words, include facts and numbers, and focus solely on substantial content. Omit any introductory or closing remarks and just output the content that this chapter would have.

Table 1: LLM Task Prompt Templates

A.2 Supplementary Results

In the following we present additional results for the baseline, Wikipedia-EPFL, and EPFL datasets.

A.2.1 Positional Bias

For all three generated MCQ datasets, GPT-4o shows a strong tendency to place the correct answer in the answer options B and C over A and D in around 80% of the time. This tendency may come from training biases where datasets exhibited a similar distribution in MCQ formats.

When ignored, this positional bias may skew a model's performance during evaluation. To counteract this effect, this project employed the rotation of the correct answer position and evaluated each question four times independently. This ensures a balanced distribution of the correct answer among the four positions and reduces the risk of skewing the evaluation statistics with the positional MCQ generation bias. After the positional bias mitagation strategy is applied, the correct answer is distrubuted evenly, appearing in each option 25% of the time. During evaluation, the models also show some levels of positional bias, however on a lower scale than during the MCQ generation.

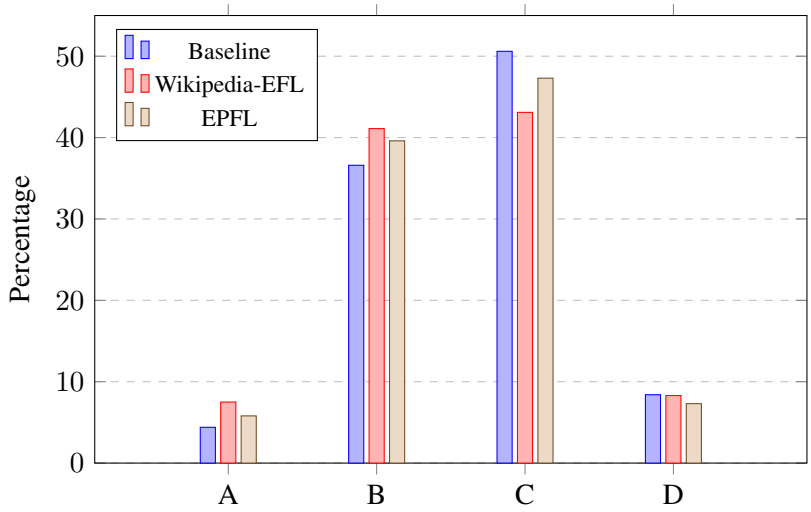


Figure 3: Distribution of the Correct Answer Among the Answer Options for the MCQ Dataset Generated with GPT-4o Before Positional Bias Mitigation.

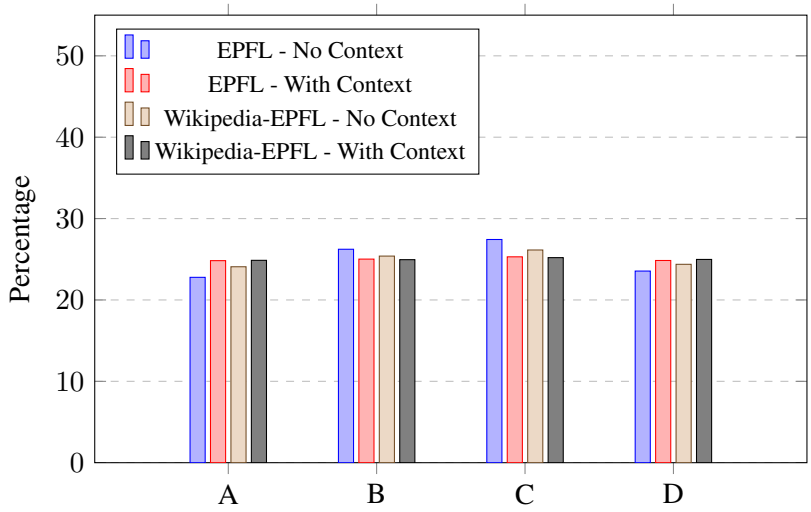


Figure 4: Distribution of Correct Answer Letter Prediction for EPFL and Wikipedia MCQ Datasets Evaluated on GPT-4o After Positional Bias Mitigation.

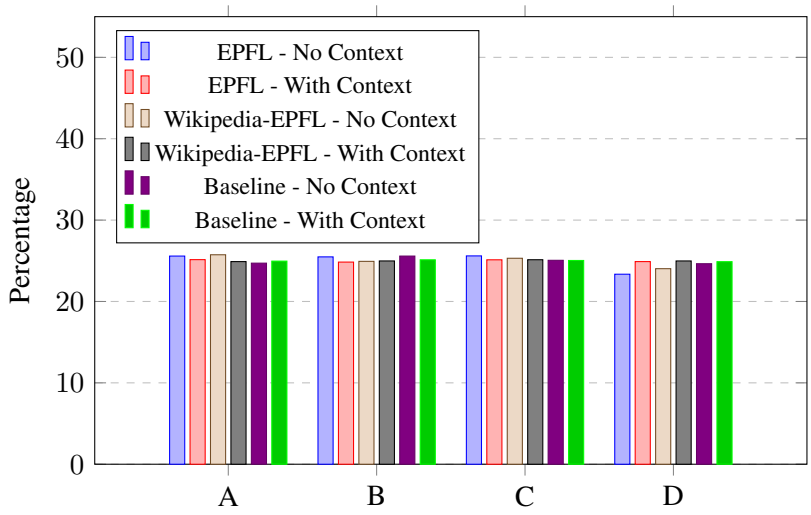
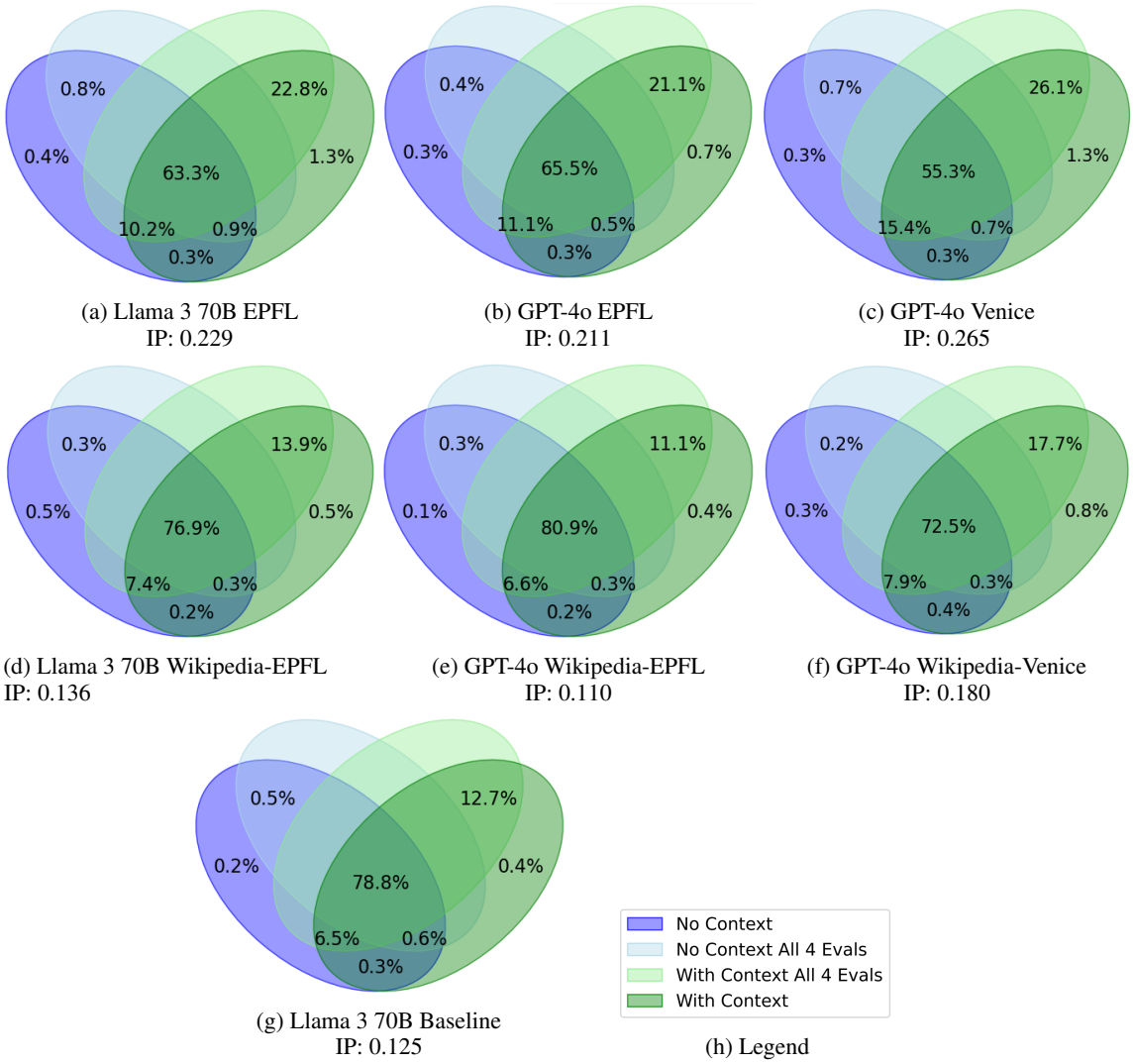


Figure 5: Distribution of Correct Answer Letter Prediction for EPFL, Wikipedia, and Baseline MCQ Datasets Evaluated on Llama 3 70B After Positional Bias Mitigation.

A.2.2 Information Potential

Figure 6 shows the models' performance with an emphasis on the positional bias mitigation strategy. As every MCQ is evaluated four times, this allows the analysis of a model's consistency. It is clearly visible that the models become less consistent without context in datasets with higher information potential while this trend is less pronounced with context. Figure 7 shows the performance of Llama 3 70B along with the information potential across the datasets and cutoff percentiles.



(a) Llama 3 70B EPFL
IP: 0.229

(b) GPT-4o EPFL
IP: 0.211

(c) GPT-4o Venice
IP: 0.265

(d) Llama 3 70B Wikipedia-EPFL
IP: 0.136

(e) GPT-4o Wikipedia-EPFL
IP: 0.110

(f) GPT-4o Wikipedia-Venice
IP: 0.180

(g) Llama 3 70B Baseline
IP: 0.125

(h) Legend

Figure 6: Statistics of Model Performances Including whether Model was Correct on all Four Evaluations of Each MCQ
Venn diagrams of the number of correctly answered questions when answered with or without context.
IP: Information Potential computed with Equation 3

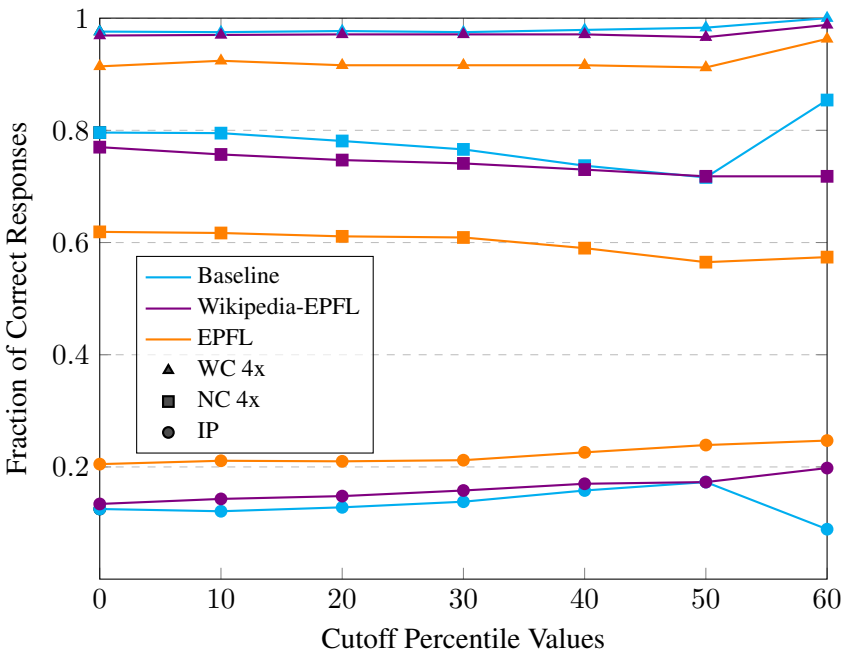


Figure 7: Llama 3 70B Performances Across Different Cutoff Percentiles with Joint Cosine and Jaccard-ROUGE-l Metric Thresholding and Number of MCQs for the Corresponding Percentiles.
NC 4x: Evaluation with no context correct for all 4 bias mitigation evaluations of each question.
WC 4x: Evaluation with context correct for all 4 bias mitigation evaluations of each question.
IP Score: The Information Potential computed with Equation 3

A.2.3 Qualitative Analysis of High-Value Information

The MCQs presented in Table 2 exemplify three key patterns in identifying valuable information:

1. **Technical Terminology:** The question about articulatory features demonstrates the significant performance gap when dealing with specialized terminology. This question, drawn from speech recognition research, requires specific context to understand how articulatory features differ from phone posterior features in their prediction approach. The model's inability to answer correctly without context highlights how technical domains in PhD manuscripts contain specialized knowledge not captured in pre-training. Such terminology-heavy questions serve as reliable indicators of domain-specific knowledge.
2. **Unique Mentions:** The question regarding "interference errors" exemplifies how PhD manuscripts capture recent research outcomes. The detailed distinction between interference errors and reduction of intentionality errors represents novel academic insights that weren't available during model pre-training. This type of question effectively identifies valuable new knowledge contributions from academic manuscripts. The consistency with which models fail these questions without context, despite their strong general reasoning capabilities, suggests genuine knowledge gaps rather than reasoning limitations.
3. **Complex Relationships:** The question about repetition errors showcases the importance of precise contextual information in understanding intricate conceptual relationships. The distinction between repetition and other error types (omission, timing, sequence) requires careful understanding of how these concepts interrelate. This category highlights a key challenge in assessing knowledge: the line between pure knowledge recall and reasoning ability becomes blurred when concepts are interconnected in complex ways. The model's performance on such questions suggests that even sophisticated reasoning capabilities cannot compensate for missing foundational knowledge.

Additional examples of these patterns can be found in Table 3.

A.2.4 Qualitative Analysis of EPFL MCQ Dataset

MCQ (correct answer in italics)	**Relevant Context Passages**
What is a common feature of 'interference errors' and 'reduction of intentionality errors'? *A) Both involve replacing the correct subject with another similar one.* B) Both typically result from incorrect or incomplete mental models. C) Both involve action reversals. D) Both require complex decision-making at the knowledge-based level.	[...] Interference errors. These errors occur when people multi-task, i.e., when multiple sequences of action are active at the same time, which can result in a person combining them. [...] this error and the ones for reduced intentionality and perceptual confusions can be modeled by replacing the correct subject with another one. [...] Reversals [...] cause a person to undo a previously performed action
Which error describes memorizing an action without proceeding to its next logical step? *A) Repetition.* B) Omission. C) Timing error. D) Sequence error.	[...] Repetitions. These errors cause a person to misjudge the progress of a sequence of actions, making them perform an action already carried on. [...] Omission, when a user skips the current action and execution continues with its successor, e.g., jumping from action i to action i + 1 [...] Timing errors, when users interact with a system at the wrong time, e.g., too early or too late [...] Sequence errors, when users execute an action out of order [...]
How are articulatory features (AF) different from phone posterior features in terms of prediction? A) AFs rely on spectral analysis. B) AFs use a frame-to-phoneme alignment. *C) AFs map phonemes to articulatory features.* D) AFs predict phonemes directly.	[...] There are different ways to represent phonemes as articulatory features, e.g. as binary features (Chomsky and Halle, 1968) or multi-valued features (Ladefoged, 1993). Similar to phone posterior features, they are trained from a frame-to-phoneme alignment. However, instead of predicting phonemes, a mapping from phones to AF is used as targets of the predictor. [...] AFs are modeled by 18 off-the-shelf recurrent neural networks (RNN) based binary classifiers, i.e. $D = 18 \times 2$. The RNNs take as input log energies of 33-dimensional Mel filterbank energies. [...] Similar to phone posterior features, [AFs] are trained from a frame-to-phoneme alignment [...]

Table 2: **Examples of technical terminology, unique mentions, and complex relationships in EPFL PhD manuscripts.** MCQs requiring context for correct model responses

MCQ (correct answer in italics)	**Relevant Context Passages**	**Category**
What is the only possible scheme for Bernstein wave excitation in the TCV tokamak due to its plasma equilibria? *A) O-SX-B scheme.* B) FX-B scheme. C) EB-B scheme. D) SX-O-B scheme.	[...] As a consequence, the extremely steep density gradients necessary for the FX-B mode conversion cannot be obtained in TCV plasma equilibria and the only possible Bernstein waves excitation scheme in TCV is the O-SX-B double mode conversion. [...]	Technical Terminology
What technological challenge is associated with reducing power consumption in CMOS circuits? *A) Maintaining acceptable dynamic range in the face of digital noise* B) Reducing the intrinsic capacitance per unit area. C) Ensuring constant voltage swing at all frequencies. D) Achieving higher gain at lower supply voltages.	[...] In downscaled processes with lower supply voltages, the coupling and noise through the substrate is higher, partially because of the limitations of the substrate and well bias [40,41]. Therefore, sometimes noise that is produced by the chip due to the digital blocks may be orders of magnitude above the thermal noise, so to achieve the required dynamic range we require a proportional increase in power. [...]	Technical Terminology
In learning molecule representations directly in the sparse code domain, what is the main constraint imposed on sparse codes? *A) They must be linear combinations of deformed molecules.* B) They must be nonlinear mixtures of entire signal sets. C) They must consist of deactivated elements. D) They must strictly adhere to original signal morphology.	[...] We constrain sparse codes to be linear combinations of a few, possibly deformed, molecules and we design an algorithm that can learn the structure from the codes without transforming them back into the signal domain. [...]	Technical Terminology
Which component significantly contributes to total variance in back-to-back scan-rescan scenarios? A) 2-week-gap variance. *B) Scan-rescan variability* C) Session-dependent offsets. D) Repositioning effects.	[...] The scan-rescan differences in back-to-back scanning scenario significantly contributed to the total variance and represented a significant proportion of between-subject variance for all of the investigated structures. [...] Both repositioning (R2) and 2-week-gap between a rescan (R3) did not significantly contribute to the total variability compared to back-to-back scans and between-subject variability. [...]	Complex Relationship
Which factor most critically affects the measurement noise in an ex-situ detection setup? *A) The frequency at which measurements are taken.* B) The remanence of the magnetic core. C) The sensitivity of the lock-in amplifier. D) The microbead placement precision.	[...] However, this increases the measurement noise, as the measurement is carried out in the 1/f noise frequency range. [...]	Complex Relationship
What condition allows the bond in the RMIB model to be unbreakable under compressive deformation? *A) High hydrostatic compressive stress.* B) High thermal conductivity. C) Low volumetric strain. D) Low thermal resistance.	[...] It is known that the hydrostatic compressive strength is infinite for most materials, which means the bond in RMIB model for these cases cannot be broken under compressive deformation. [...]	Complex Relationship
What is the approach used by RouLette to manage materialization overhead in symmetric joins? *A) Symmetric join pruning of tuples forming outputs.* B) Incremental materialization of queried tuples. C) Partial materialization of all relations. D) Deferred materialization until query execution.	[...] Symmetric joins require that all relations be materialized and hence incur materialization overhead. To reduce the overhead, RouLette materializes only tuples that can form output tuples for their query-set. We call this symmetric join pruning [...]	Unique Mentions
What primary limitation affects the clinical success of MPCs in bone healing? *A) Limited number of available endogenous MPCs.* B) Extensive proliferation in vitro. C) Over-differentiation into non-mesenchymal lineages. D) High heterogeneity in cell populations.	[...] its clinical outcome was rather disappointing 4 . One of reasons for this seems to be the limiting number of available endogenous mesenchymal progenitor cells (MPCs) that can give rise to bone cells. [...]. Hence, there is a clear clinical need for implants that augment the homing/recruitment of endogenous MPCs to fracture sites [...]	Unique Mentions
What leads to the gradual increase in average THC concentration over time during oscillations? *A) Accumulation of carbonates on ceria sites.* B) Thermal degradation of the catalyst. C) Continuous ceria site activation. D) Increasing gas hour space velocity (GHSV).	[...] The higher average THC concentrations levels with time was caused by the gradual accumulation of carbonates on ceria sites during the periodic oscillations. [...]	Unique Mentions

Table 3: Selection of EPFL Dataset MCQs and their Relevant Context Passages Categorized by Question Type where GPT-4o Required Context to Correctly Answer Consistently