OpenHuEval: Evaluating Large Language Model on Hungarian Specifics

Anonymous ACL submission

Abstract

We introduce OpenHuEval, the first benchmark designed for comprehensive evaluation of large language models (LLMs) in the context of the Hungarian language and specifics. OpenHuEval incorporates the latest design principles for evaluating LLMs, such as using real user queries from the internet, emphasizing the assessment of LLMs' generative capabilities, and employing LLM-as-judge to enhance the multidimensionality and accuracy of evaluations. We evaluated current mainstream LLMs, including both traditional LLMs and recently developed Large Reasoning Models. The results demonstrate a significant necessity for evaluation and model optimization tailored to the Hungarian language and specifics. We also conducted a detailed analysis of the reasoning process of LRMs on OpenHuEval, revealing the intrinsic patterns and mechanisms of these models in non-English languages, with Hungarian serving as a representative example. We will release OpenHuEval on GitHub.

1 Introduction

Recent advancements in large language models (LLMs), such as o1(Jaech et al., 2024) and DeepSeek-R1(Team, 2024a), mark significant progress toward artificial general intelligence (AGI). However, notable performance gaps remain between English and other languages in both language-agnostic tasks (e.g., reasoning, code generation) and language-specific tasks (e.g., idiom usage, cultural understanding), posing challenges to global AI deployment and equitable development. The cross-lingual performance gaps in LLMs stem from two main factors: First, the training data of LLMs, particularly the pre-training data is severely imbalanced in language representation. Second, while English evaluation systems are advanced and rapidly evolving, non-English systems are underdeveloped, particularly for language-specific features. This limits the identification of shortcomings in non-English languages and leads to their neglect in research.

This paper focuses on the evaluation of Hungarian and its distinctive capabilities, spoken by around 14 million people worldwide. The findings aim to improve the user experience for Hungarian speakers while offering insights for similar studies in other languages. Existing Hungarian evaluation datasets are largely translations of English ones, missing essential Hungary-specific elements such as language nuances, culture, history, and regional context, which are key for Hungarian users. Among the existing evaluation datasets, HuLU (Ligeti-Nagy et al., 2024) is a key benchmark for Hungarian language understanding, but its focus on multiple-choice and true/false questions limits its ability to evaluate broader LLM capabilities, such as language generation, open-domain Q&A, reasoning, knowledge representation, hallucination, and instruction-following.

To address this gap, we introduces OpenHuEval, the first evaluation benchmark for LLMs focused on Hungarian language and its comprehensive capabilities.The comparison of OpenHuEval with previous related benchmarks is shown in Table 1. Overall, OpenHuEval has two main distinguishing features:

1) **Hungarian-Specific:** Inspired by (Liu et al., 2024b), we propose eight distinct Hungarian-specific dimensions (Section 2.1), covering a variety of scenarios that users may encounter when querying in Hungarian. Guided by these dimensions, we collected a vast amount of Hungary-specific material from multiple sources and used this to construct the corresponding evaluation tasks.

2) **Keeping up with the Latest Advances in LLM Evaluation:** Significant progress has been made in LLM evaluation, with query sources shifting from manual or rule-based constructions to real-world internet questions, enhancing practical relevance. Question formats evolved from multiple-choice to open-ended Q&A, better reflecting actual usage. Evaluation methods transitioned from rule-based approaches to LLM-as-judge and subjective assessments, improving accuracy and objectivity. However, these advancements primarily apply to English datasets and not Hungarian. Thus, in creating OpenHuEval, we incorporated these principles and methodologies from English evaluations.

Based on OpenHuEval, we evaluated the performance of mainstream LLMs in Hungarian language and specifics. We compared the performance differences of these models on the typical datasets of OpenHuEval with corresponding datasets in other languages. The

Benchmark	Real user query	Self-awareness evaluation	Proverb Reasoning	Generative task & llm-as-judge	Hungarian Lang	Comprehensive Hu-specific
WildBench(Lin et al., 2024)	✓	✗	✗	✓	✗	✗
SimpleQA(Wei et al., 2024a), ChineseSimpleQA(He et al., 2024)	✗	✓	✗	✓	✗	✗
MAPS(Liu et al., 2024c)	✗	✗	✓	✗	✗	✗
MARC, MMMLU et al in (Lai et al., 2023)	✗	✗	✗	✗	✓	✗
BenchMAX(Huang et al., 2025)	✗	✗	✗	✓	✓	✗
HuLU(Ligeti-Nagy et al., 2024)	✗	✗	✗	✗	✓	✗
OpenHuEval (ours)	✓	✓	✓	✓	✓	✓

Table 1: Comparison of related benchmarks.

results indicate a significant necessity for evaluation and model optimization specifically for Hungarian language and specifics.

Moreover, Large Reasoning Models (LRMs), such as o1, represent a new direction in LLM development. By engaging in extensive reasoning, self-reflective negation, and exploring multiple reasoning paths, they significantly enhance reasoning capabilities, following the test-time scale law—a key step toward AGI. Recent studies have analyzed these reasoning processes, offering insights for optimization, but have largely focused on English-language contexts, neglecting Hungarian and task-specific scenarios. To address this, we leverage OpenHuEval's datasets to develop frameworks for analyzing reasoning in cutting-edge LRMs. Our analysis reveals intrinsic patterns in non-English contexts, using Hungarian as a case study, providing valuable insights for advancing LRMs in the research community.

In summary, the contributions of this paper include the following three points:

- We developed OpenHuEval, the first benchmark for LLMs focusing on the Hungarian language and specifics. OpenHuEval incorporates the latest design principles for evaluating LLMs, such as using real user queries from the internet, emphasizing the assessment of LLMs' generative capabilities, and employing LLM-as-judge to enhance the multidimensionality and accuracy of evaluations.

- We conducted a comprehensive evaluation of current mainstream LLMs, including traditional LLMs and recently developed LRMs. The results highlight the significant necessity for evaluation and model optimization tailored to Hungarian language and specifics.

- We established a set of analytical methods to perform a detailed and in-depth analysis of the reasoning processes of cutting-edge Large Reasoning Models, revealing the intrinsic patterns and mechanisms of these models in non-English languages, with Hungarian as a representative.

2 OpenHuEval

OpenHuEval is a benchmark specifically designed to evaluate the performance of LLM in handling localization and culture-rich challenges unique to Hungary. The overview of OpenHuEval is in Figure 1. Examples of some tasks are shown in Figure 2. This chapter offers a comprehensive introduction of the construction process of OpenHuEval.

2.1 Hungarian-specific dimensions and OpenHuEval tasks

OpenHuEval encompasses eight Hungarian-specific dimensions, as shown in Table 2: Language (L), History (H), Life, Culture, and Customs (LCC), Education and Profession (EP), Geography and Place (GP), Figure (F), Politics, Policy, and Law (PPL), and Business and Finance (BF). These dimensions comprehensively cover a wide range of scenarios encountered by users when utilizing Hungarian as the query language. As a result, they enable a thorough, systematic, and holistic evaluation of the performance of LLMs in tasks related to the Hungarian.

Bearing the above Hungarian-specific dimensions in mind, the construction of evaluation tasks tailored to Hungarian characteristics first requires collecting corpora rich in Hungarian cultural elements as raw materials. Following previous works, we collected data from sources such as Hungarian proverbs (Liu et al., 2023; Sun et al., 2024), exam questions (Li et al., 2023), forums (Arora et al., 2024), and wikipedia. Through processes including filtering, refinement, construction, and quality assurance, we developed a total of five evaluation tasks comprising 4003 questions in total, as detailed in Table 3. The subsequent sections of this chapter will introduce these tasks and their corresponding datasets in detail.

2.2 Hungarian WildBench

Task Introduction: The Hungarian WildBench (HuWildBench) task aims to evaluate the performance of LLMs in answering various questions arising from the everyday lives of Hungarians. All questions are sourced from Hungary's well-known forum website[1] (hereinafter referred to as "g13k" for brevity), and thus reflect **real-life issues encountered by Hungarians**. These questions cover a wide range of topics, including cultural customs, education, tourism, legal regulations, and business and finance. Examples of HuWildBench questions are shown in Figure 2 and Table 14. Since the queries in HuWildBench are user-generated content from the g13k website, their linguistic expressions

[1] https://www.gyakorikerdesek.hu/, which is similar to https://www.quora.com/ for English-speaking world.

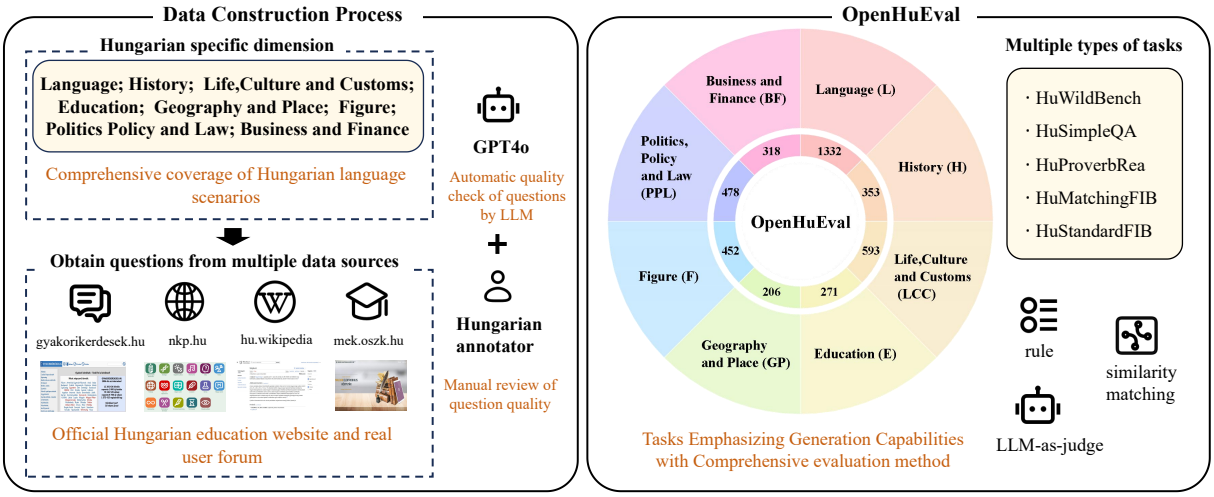

Figure 1: Overview of OpenHuEval.

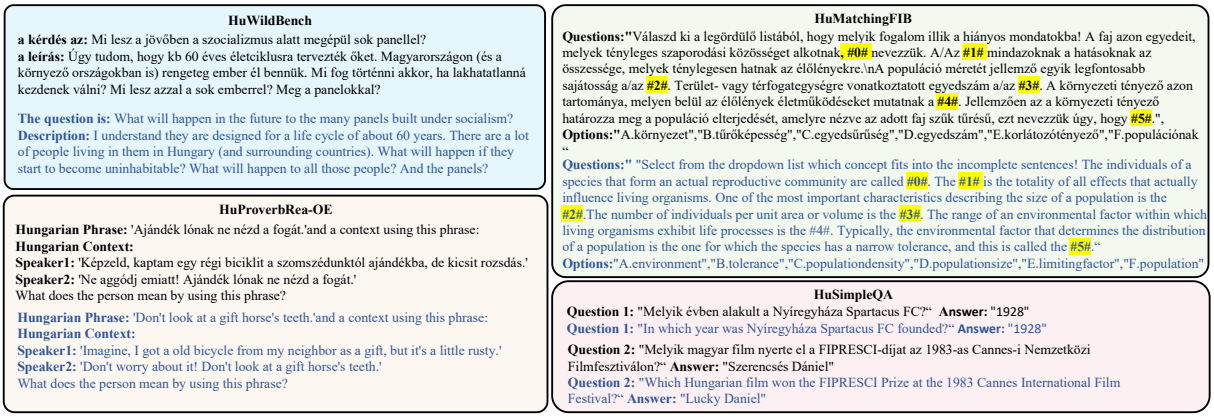

Figure 2: Examples of OpenHuEval.

and question formats tend to be less formal than the structured and polished written language. This poses a realistic challenge for LLMs, as they must adapt to such informal and spontaneous language style. The construction of HuWildBench is detailed in Appendix B.

Metric and judge: The WB-Score in (Lin et al., 2024) is employed as the evaluation metric in the following manner: Following (Lin et al., 2024), GPT-4o is used as the judge model. The judge model then evaluates the quality of each response based on the checklist and provides detailed strengths and weaknesses before assigning a score from 1 to 10. The definition of scores is shown in Table 10. Different with (Lin et al., 2024), our final scores are calculated as the average of all test sample scores, with each score multiplied by 10.

2.3 Hungarian SimpleQA

Task Introduction: Hungarian SimpleQA (HuSimpleQA) is designed to assess the ability of LLMs to answer short, fact-seeking questions related to Hungary. Inspired by (Wei et al., 2024b) and (He et al., 2024), we constructed HuSimpleQA based on Hungarian Wikipedia[2], with the following key characteristics. **Hungarian:** The questions in HuSimpleQA are in Hungarian, and they focus on facts specifically related to Hungary. **Diverse:** The questions in HuSimpleQA cover the eight Hungary-specific dimensions proposed in Section 2.1. **High-quality:** The construction process of HuSimpleQA (in Appendix C.5) includes comprehensive and strict quality control procedures, ensuring the quality and accuracy of the questions. **Static:** Similar to SimpleQA, the answers to the questions in HuSimpleQA do not change over time, ensuring that the dataset remains evergreen. **Easy-to-evaluate:** The questions and answers in HuSimpleQA are short and concise, making them ideal for evaluation through LLMs. The examples of HuSimpleQA are shown in Figure 2 and Table 16.

Metric and Judge: Following the methodology of (Wei et al., 2024b), we use GPT-4o as a judge to categorize the responses of the LLM to HuSimpleQA into three classes: “correct”, “incorrect", or “not attempted”. The definitions and examples for these categories the judge prompt are detailed in Appendix C.7, Figure 22. Based on the results from the judge, we evaluate the

[2] https://hu.wikipedia.org/

Hungarian-specific dimensions	Definition	#Question
Language(L)	Basic knowledge of the Hungarian language and Hungarian proverbs and sayings	1332
History(H)	Historical events and historical development of Hungary	353
Life, Culture, and Custom(LCC)	Religion, rituals, culture, holidays, and the daily life of Hungarians	593
Education and Profession(EP)	Education system in Hungary and related professions	271
Geography and Place(GP)	Geographical knowledge of Hungary, cities, and locations	206
Figure(F)	Famous figures of Hungary	452
Politics, Policy and Law(PPL)	Politics, policies, and laws of Hungary	478
Business and Finance(BF)	Business and finance in Hungary	318

Table 2: Hungarian-specific dimensions

Task	Hungarian-specific dimensions	Judge	Question type	#Question
HuWildBench	LCC, EP, PPL, BF	llm,checklist	OE	1154
HuSimpleQA	L,H,LCC,EP,GP,F,PPL,B	llm	OE	1343
HuProverbRea	L	rule,llm	2CQ/OE	1135
HuMatchingFIB	L, H	rule	Matching Filling-in-Blank	278
HuStandardFIB	L, H	rule,similarity matching	Standard Filling-in-Blank	93

Table 3: Tasks of OpenHuEval

performance of the LLM on HuSimpleQA using the following five metrics:

- Correct (CO): The predicted answer completely encompasses the reference answer without any conflicting or contradictory information.

- Not Attempted (NA): The predicted answer does not fully include the reference answer, but there are no contradictions between the two.

- Incorrect (IN): The predicted answer contradicts the reference answer, regardless of whether the contradiction is resolved.

- Correct Given Attempted (CGA): This metric measures the percentage of accurately answered questions out of all attempted questions.

- F-score: This metric calculates the harmonic mean between the proportion of correct answers and the proportion of correct answers among attempted questions.

2.4 Hungarian Proverb Reasoning

Task Introduction: Hungarian Proverb Reasoning (HuProverbRea), which consists of the collection of Hungarian proverbs, idioms, abbreviations, is a task that requires the LLM to **understand and reason the meaning of Hungarian proverbs in a specific context**. As shown by the examples in Figure 2, LLM is provided with a context in which a Hungarian proverb is used, accompanied by a question: *"What does the speaker mean by the saying?"*. Then, the LLM is tasked with discerning the speaker's true intention, either by selecting the correct interpretation from two provided options (2CQ setting), or by directly articulating the speaker's intended meaning (OE setting).

Metric and judge: For the 2CQ setting, we simply measure the correct ratio of candidate LLMs. For the OE setting, we adopt GPT-4o as judge to decide if the answer is acceptable. We provide the original proverb, its context and the English explanation of the proverb as references when judging OE responses. Detailed prompt templates are listed in Appendix D.

2.5 Hungarian Matching and Standard Filling-in-Blank

Task Introduction: Hungarian Matching Filling-in-Blank (HuMatchingFIB) is a task similar to traditional fill-in-the-blank exercises. In this task, several key terms in a given text are blanked out, and a candidate pool is provided, which contains both the correct answers and several distractors. The responsibility of the LLM is to select the most appropriate answers from the candidate pool to fill in the blanks, thereby restoring the complete meaning of the text. The example is shown in Figure 2 and Figure 28. This task effectively tests the LLM's abilities in information comprehension, contextual reasoning, and distinguishing between correct answers and distractors.

Similarly, Hungarian Standard Filling-in-Blank (HuStandardFIB) also follows the format of a fill-in-the-blank exercise. However, unlike HuMatchingFIB, this task does not provide a candidate pool containing the correct answers. Instead, the LLM is required to complete the blanks based on its internal knowledge and the given context. The example are shown in Figure 29. Consequently, HuStandardFIB evaluates the LLM's comprehensive capabilities in knowledge recall and contextual reasoning.

Metric and Judge: HuMatchingFIB employs a rule-based evaluation approach, where the assessment is conducted at two levels: the blank level and the ques-

tion level (as a single question may contain multiple blanks). The evaluation process is analogous to that of multiple-choice questions, and accuracy (acc) is used as the metric to determine performance. The corresponding formula for blank level accuracy is as follows, where c represents the number of correctly predicted blanks in one question, t represents the number of blanks in one question.

$$\text{Acc}_{\text{blank level}} = \frac{\sum \text{blank}_\text{c}}{\sum \text{blank}_\text{t}} \quad (1)$$

HuStandardFIB questions are designed with open-ended reference answers to accommodate variations in part of speech and semantics. We employ a many-to-one fuzzy matching mechanism. Fuzzy matching is a technique that calculates the similarity between strings, allowing for flexibility in matching by considering variations such as typos, synonyms, or different word orders. In this context, the model's answer is compared against a set of possible reference answers (where multiple correct answers may exist for a single question or blank). If the similarity score between the model's answer and any of the reference answers exceeds a predefined threshold, the answer is considered correct. This approach is particularly suitable for evaluating open-ended questions where exact matches are often infeasible due to the variability in acceptable responses. The annotator information involved in all tasks of OpenHuEval can be found in Appendix G.

3 Experiments and Analysis

3.1 Experimental setup

We utilize OpenHuEval to benchmark the performance of large language models (LLMs) in handling Hungarian localization tasks and culturally rich Hungarian-specific issues. We evaluated state-of-the-art LLMs including GPT-4o (Hurst et al., 2024), GPT-4o mini[3], Deepseek-V3 (Liu et al., 2024a), Qwen2.5-Instruct (Yang et al., 2024), and Llama-3.1-Instruct (Dubey et al., 2024), as well as the latest Large Reasoning Models (LRMs) such as OpenAI o1-mini (Jaech et al., 2024), QwQ-32B-Preview (Team, 2024b) (abbreviated as QwQ in following text), and Deepseek-R1 (Team, 2024a). Detailed specifications of these models are provided in Table 4.

We used OpenCompass in all our experiments. For traditional instruction-based LLMs, we adopted OpenCompass's default settings for the maximum output length. For Large Reasoning Models, we set the output length to 8192 to ensure sufficient space for reasoning process and to produce a complete final answer, avoiding premature output truncation. For OpenAI models (GPT series and o1-mini), we used their official API with settings following OpenCompass's default configuration. For Deepseek-V3 and Deepseek-R1, due to the high usage volume of Deepseek's official API causing instability, we used equivalent API services provided by Alibaba Cloud[4] and Silicon Valley Flow[5]. The settings followed OpenCompass's configurations, with the temperature set to 0.7. For other models in Table 4, we performed inference locally with NVIDIA A100 GPUs, using LMDeploy as the inference backend. The settings followed OpenCompass's default configuration (Temperature = 1e-6, top_k = 1).

Model	Size	Reasoning Model	Open-source?	Inference Method
GPT-4o	-	N	N	Official API
GPT-4o-mini	-	N	N	Official API
Deepseek-V3	-	N	Y	Alibaba Cloud and SiliconFlow API
Qwen2.5-Instruct	7B,72B	N	Y	Local GPU
Llama-3.1-Instruct	8B,70B	N	Y	Local GPU
o1-mini		Y	N	Official API
QwQ	32B	Y	Y	Local GPU
Deepseek-R1	-	Y	Y	Alibaba Cloud and SiliconFlow API

Table 4: LLMs evaluated in our experiments.

3.2 Overall performance

The overall performance of all LLMs on OpenHuEval is presented in Table 5. It can be observed that across a total of five tasks, Deepseek-R1 ranks first in three tasks and achieves top-tier performance in the other two tasks. GPT-4o ranks first in two tasks and second in the remaining three tasks. These results demonstrate the exceptional performance of the two models in Hungary-specific tasks.

Open-source models vs Closed-source models: Among open-source models, Deepseek-R1 stands out, while Deepseek-V3 also demonstrates strong overall performance, ranking highly across all tasks. Llama-3.1-Instruct-70B achieved impressive scores of 93.83% in the HuProverbRea-2CQ task and 36% in the HuSimpleQA task, ranking second only to the closed-source model GPT-4o. This highlights the growing potential of open-source models, led by Deepseek-R1, which are increasingly showing capabilities comparable to closed-source models in Hungarian language tasks. These results indicate that open-source models are closing the gap and are becoming highly competitive in specific application domains.

Traditional LLMs vs. Large Reasoning Models: We compared Traditional LLMs and LRMs within the same series. Across five tasks, Deepseek-R1 consistently outperforms Deepseek-V3 in four of them. Specifically, in the HuMatchingFIB task, Deepseek-R1 achieves relative improvements of 12% at the blank level and 7.19% at the question level compared to Deepseek-V3. Similarly, for the HuStandardFIB task, it achieves gains of 10.32% (blank level) and 7.52% (question level). Although Deepseek-R1 performs slightly worse than Deepseek-V3 on the HuProverbRea task, the performance gap is less than 1%. Considering that both Deepseek-R1 and Deepseek-V3 are based on the same pretrained model, the significantly stronger performance

[3]We used the gpt-4o-2024-11-20 version for GPT-4o and the gpt-4o-mini-2024-07-18 version for GPT-4o-mini.

[4]https://cn.aliyun.com/

[5]https://siliconflow.cn/

Model	HuWildBench WBScore	HuSimpleQA Acc	HuProverbRea Acc. (OE)	HuProverbRea Acc. (2CQ)	HuMatchingFIB B acc.	HuMatchingFIB Q acc.	HuStandardFIB B acc.	HuStandardFIB Q acc.
GPT-4o	81.09	50.3	89.16	95.51	77.78	43.88	57.36	15.05
GPT-4o-mini	74.19	24.52	84.67	92.16	55.68	19.78	35.08	7.53
QwQ	58.02	9.17	67.49	84.23	38.65	12.23	6.05	0
Deepseek-R1	82.96	34.58	82.29	91.72	80.87	47.12	61.76	17.2
Deepseek-V3	78.42	32.71	83.26	92.51	68.87	39.93	51.44	9.68
Llama-3.1-Instruct-70B	61.78	36.36	80.18	93.83	59.56	24.46	40.99	6.45
Llama-3.1-Instruct-8B	53.62	14.9	63.35	73.48	5.74	0.72	16.64	1.08
o1-mini	76.43	16.24	77.44	87.67	60.83	17.63	45.25	13.98
Qwen2.5-Instruct-72B	74.05	15.05	77.8	90.22	63.8	24.1	32.32	8.6
Qwen2.5-Instruct-7B	42.01	5.29	50.48	67.05	31.88	1.08	7.43	0

Table 5: Overall performance of 10 LLMs on OpenHuEval. The first, second, and third place in each metric are marked with red, green, and blue text, respectively. In the FIB task evaluation metric, **B** represents the blank level, and **Q** represents the question level.

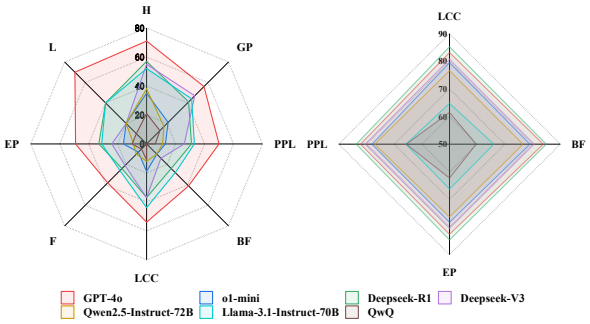

Figure 3: Results of the Hu-specific Dimension

Rank	HuProverbRea	MAPS(en)
1	GPT-4o	GPT-4o (-)
2	Llama-3.1-Instruct-70B	Llama-3.1-Instruct-70B (-)
3	Deepseek-V3	Qwen2.5-Instruct-72B (↑3)
4	GPT-4o-mini	GPT-4o-mini (-)
5	Deepseek-R1	Deepseek-V3 (↓2)
6	Qwen2.5-Instruct-72B	Deepseek-R1 (↓1)
7	o1-mini	Qwen2.5-Instruct-7B (↑3)
8	QwQ	Llama-3.1-Instruct-8B (↑1)
9	Llama-3.1-Instruct-8B	o1-mini (↓2)
10	Qwen2.5-Instruct-7B	QwQ (↓2)

Table 6: The LLMs rankings on HuProverbRea and MAPS datasets.

Rank	Simpleqa	HuSimpleQA
1	GPT-4o	GPT-4o(-)
2	Deepseek-R1	Llama-3.1-Instruct-70B(↑1)
3	Llama-3.1-Instruct-70B	Deepseek-R1(↓1)
4	Deepseek-V3	Deepseek-V3(-)
5	QwQ	GPT-4o-mini(↑3)
6	Llama-3.1-Instruct-8B	o1-mini(↑3)
7	Qwen2.5-Instruct-72B	Qwen2.5-Instruct-72B(-)
8	GPT-4o-mini	Llama-3.1-Instruct-8B(↓2)
9	o1-mini	QwQ(↓4)
10	Qwen2.5-Instruct-7B	Qwen2.5-Instruct-7B(-)

Table 7: The LLMs rankings on SimpleQA and HuSimpleQA

of Deepseek-R1 on the OpenHuEval benchmark demonstrates the effectiveness of LRMs architectures in Hungarian language tasks and domain-specific scenarios. This result underscores the potential of LRMs as a key avenue of exploration in advancing Artificial General Intelligence (AGI).

Model size: From the results, models with larger parameter sizes perform better on OpenHuEval. For example, GPT-4o, Llama-3.1-Instruct-70B, and Qwen2.5-Instruct-72B outperform their smaller counterparts in the same series (such as GPT-4o-mini, Llama-3.1-Instruct-7B, and Qwen2.5-Instruct-7B) across all tasks.

3.3 Results and Analysis of the Hu-specific Dimension

We selected the HuWildBench and HuSimpleQA tasks and visualized the performance of several models on Hu-specific dimensions. The results are shown in Figure 3. The left subfigure shows the results of HuSimpleQA, while the right subfigure presents the results of HuWildBench. In the HuWildBench task, the models demonstrate relatively balanced performance across the four dimensions: LCC, PPL, BF, and EP. This suggests that the models are capable of providing well-rounded responses to the diverse range of questions typically posed by Hungarian users in daily interactions. In the HuSimpleQA task, which evaluates the models across all 8 Hu-specific dimensions, performance differences emerge across various knowledge areas. Specifically, the task focuses on the model's grasp of Hungarian factual knowledge. For the dimensions of H, LLC, and GP, the models show relatively strong performance, as these types of knowledge are more commonly found in the training data. However, for the dimensions more closely related to the unique characteristics of Hungary, such as BF, PPL, and EP, there is a noticeable gap in performance. This highlights the need for LLM researchers to prioritize the enhancement of capabilities related to the knowledge of smaller, less-represented languages and their unique cultural contexts.

3.4 Comparison with Existing Benchmarks

We selected two datasets from OpenHuEval, HuSimpleQA and HuProverbRea, to compare model performance ranking differences with similar datasets:

HuProverbRea vs MAPS: We compared the model performance rankings on the HuProverbRea and MAPS datasets, as shown in Table 6. Among the 10 models, 7 experienced ranking changes, accounting for 70%. Notably, Qwen2.5-Instruct-72B and Qwen2.5-Instruct-7B each moved up by three positions, with an average ranking change of 1.4 positions. Ranking differences

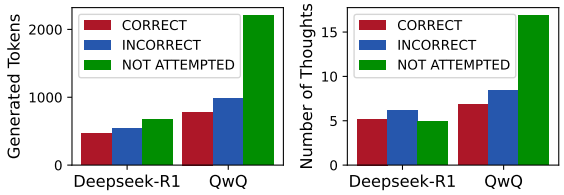

Figure 4: Illustration of the average number of tokens and thoughts generated per response between Deepseek-R1 and QwQ.

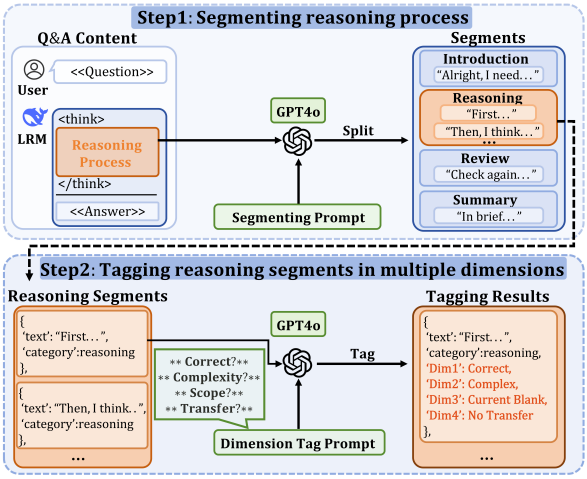

Figure 5: Segmenting and tagging the reasoning process of LRM on HuMatchingFIB.

for models in other languages on HuProverbRea and MAPS datasets are provided in Table 12.

HuSimpleQA vs SimpleQA: We compared model performance rankings on the HuSimpleQA and SimpleQA datasets, as shown in Table 7. Among the 10 models, 6 experienced ranking changes, accounting for 60%. Some changes were significant, such as llama3.1-8B dropping by 2 positions, while GPT4o-mini and o1-mini each rose by 3 positions.

These results underscore the importance of evaluating LLMs on Hungarian proverbs and Hungarian-specific questions, highlighting the need for targeted optimization of models to better handle language-specific proverbs and cultural nuances across diverse languages.

4 LRM's reasoning process on OpenHuEval

We conducted an in-depth statistical analysis of the reasoning processes of two LRMs (Deepseek-R1 and QwQ) on the OpenHuEval benchmark. For this purpose, we selected two tasks: HuSimpleQA and HuMatchingFIB.

Unlike recent work (Wang et al., 2025), which focuses solely on the reasoning processes of LRMs in Math reasoning datasets, the two tasks we selected each have distinctive characteristics: HuSimpleQA assesses the LLM's ability to recall and retrieve Hungarian-specific knowledge, as well as its awareness of its own knowledge boundaries. HuMatchingFIB involves questions where multiple competitive blanks exist within the same problem, requiring the model to carefully choose which answers to fill in.

4.1 Analysis on HuSimpleQA

Similar to (Wang et al., 2025), each query in HuSimpleQA requires answering only one question. Therefore, following the approach in (Wang et al., 2025), we segmented the reasoning process of LRMs into "thoughts". A "thought" refers to an intermediate cognitive step output by a LRM during its reasoning process. Throughout the reasoning process, the LLM transitions between multiple thoughts, which are typically separated by reflective phrases such as "Alternative", "Várni"(wait). An illustration of these transitions can be found in Appendix F, Figure 42.

For the reasoning processes of Deepseek-R1 and QwQ on the HuSimpleQA task, we first used GPT-4o to segment their thoughts (see the prompt in Appendix F, Figure 39 and Figure 40). Then, following the approach in (Wang et al., 2025), we evaluated the correctness of each thought (the prompt is detailed in Appendix F, Figure 41), with examples provided in Appendix F, Figure 42.

We subsequently analyzed the reasoning process by measuring the length of the process (in terms of token count) and the number of thoughts under different evaluation outcomes of the final query answers (correct, incorrect, abstained). The results are shown in Figure 4. The results indicate that both the reasoning length (in tokens) and the thought count were generally shorter for Deepseek-R1 compared to QwQ. Considering that Deepseek-R1 performs better than QwQ on the HuSimpleQA task, it suggests that Deepseek-R1 achieves its superior performance with relatively lower reasoning overhead. Further analysis reveals that for Deepseek-R1, the reasoning length and thought count showed no significant differences across the three types of evaluation outcomes (correct, incorrect, not attempted). In contrast, for QwQ, the length and the number of thoughts were significantly higher in the "not attempted" cases compared to the other two types. This observation suggests that, compared to Deepseek-R1, QwQ is less "confident", which tends to repeatedly attempt generating answers when faced with uncertainty and is more inclined to abstain from answering altogether.

4.2 Analysis on HuMatchingFIB

Unlike HuSimpleQA, where each query contains only one question, HuMatchingFIB involves multiple competitive blanks within the same question that need to be filled. We found that LRMs typically address HuMatchingFIB questions by sequentially solving each blank one at a time. However, more complex scenarios can also arise during the reasoning process, such as revising the answers to earlier blanks or simultaneously analyzing and resolving multiple blanks.

Segment types	Definition
Introduction	Brief Introduces the topic or provides background information, typically without detailed reasoning.
Reasoning	Contains logical reasoning, analysis, or argumentation, often using connectors like 'because', 'therefore', or 'thus'.
Review	Reflects or reviews the reasoning process or conclusions, often using phrases like 'in summary' or 'to recap'.
Summary	Summarizes the overall content or provides final conclusions, often using phrases like 'in conclusion' or 'overall'.

Table 8: Types of reasoning segments in LRM's reasoning process on HuMatchingFIB.

Analytical Method: After conducting extensive case studies, we developed an analytical method specifically designed to dissect the reasoning process for the HuMatchingFIB task.

We first segment the reasoning process of LRMs into multiple segments. Each reasoning process typically begins with a "Introduction" segment, includes several "Reasoning" segments and some "Review" segments in the middle, and ends with a "Summary" segment. The definitions of these four types of segments are shown in Table 8. The segmenting and classification is conducted by GPT4o. The prompt template is detailed in Figure 32 and Figure 33. Given the significant differences in the reasoning processes between Deepseek-R1 and QwQ, we selected different few-shot examples for each model to ensure the accuracy of segmentation and classification.

Subsequently, based on the classification dimensions outlined in Figure 37 and Figure 38, we conducted fine-grained classification of the reasoning segments. The classification process was also performed by GPT-4o. The classification process ensured a deep understanding of the models' reasoning mechanism and laid a reliable foundation for subsequent analysis.

Statistical Analysis of Reasoning Segments: We conducted a statistical analysis of the Reasoning Segments and identified several noteworthy phenomena:

- *Simple Assertion or Complex Thought:* We found that the Reasoning Segments can be categorized into two types. The first type is referred to as Simple Assertion, where LRM directly provides the answer to the blank. The second type is termed Complex Thought, where the segment involves repeatedd thinking, logical reasoning, hypothesis validation, or other complex processes. The statistics reveal that the accuracy rate of Simple Assertions is generally higher, particularly for the Deepseek-R1 model. This indicates that simple and direct reasoning tasks are relatively easier for the LRMs, and it also demonstrates that Deepseek-R1's calibration is relatively reliable, suggesting that the model "knows what it knows". As for QwQ's performance on Simple Assertion is slightly inferior, these statistical results also aligns the conclusion drawn in Section 4.2.4. The accuracy rate for Complex Thought reasoning is significantly lower than that of Simple Assertions, and both models exhibit a higher proportion of cases where no clonclusion is reached when dealing with complex reasoniing. This suggest that complex reasoning tasks are more challenging for the models.

	DeepSeek-R1	QwQ
Simple Assertion	0.5549	0.3627
Correct Simple Assertion	0.9342	0.7426
Correct Complex Reasoning	0.5257	0.4237

Table 9: Account of Reasoning Segments of HuMatchingFIB

- *Explicit Translation Insertion (ETI):* We observed that in some Reasoning Segments, when faced with a problem in Hungarian, the LLM first translates a key phrase of the original question into English and then proceeds with analysis and reasoning based on this translation. For example, *"... Erőteljes #3# és a költői #4# gazdag használata jellemzi.* ***This translates to*** *"It is characterized by strong #3# and rich use of poetic #4#."* ...". We refer to this phenomenon as Explicit Translation Insertion (ETI). Statistical analysis shows that ETI occurs in 5.49% of DeepSeek-R1's Reasoning Segments, while for QwQ, the proportion is 16.77%. This indicates that QwQ is relatively weaker in handling non-English inputs, tending to translate first and then reason, which is consistent with previous research findings on cross-lingual Chain-of-Thought (CoT) in LLMs.

5 Conclusion

In this paper, we constructed the first benchmark for LLMs focusing on the Hungarian language and its specifics. The results highlight the significant need for evaluation and model optimization tailored to Hungarian language and specifics. We developed analytical methods to deeply analyze the reasoning processes of advanced Large Reasoning Models, revealing their intrinsic patterns and mechanisms in non-English languages, using Hungarian as an example. Our work not only advances LLM technology in Hungarian but also provides valuable insights for studying languages of other countries and regions.

6 Limitation

This paper, based on the proposed OpenHuEval framework, conducts an in-depth analysis of LLMs in processing Hungarian language and culture, providing a comprehensive evaluation of the performance of current mainstream LLMs and LRMs. However, with the rapid development of English evaluation datasets, this study serves only as a phased effort to bridge the gap between Hungarian and English evaluation datasets. Overall, small-language evaluation datasets still exhibit significant shortcomings compared to their English counterparts. In the future, we plan to closely follow advancements in English evaluation datasets, continually refine

and enhance evaluation methods and datasets for low-resource languages, and work towards narrowing this gap.

Additionally, with the rapid progress in the field of LLMs, many outstanding models have yet to be fully evaluated, particularly those designed specifically for low-resource languages. In the future, we aim to establish a comprehensive OpenHuEval community that will regularly update evaluation results for the latest models, ensuring comprehensive and cutting-edge assessments while driving the optimization and development of models in the low-resource language domain.

7 Ethical Consideration

This work involved human annotation. For all annotators, we explicitly informed them about the use of the data and required them to ensure that the questions included in OpenHuEval do not involve any social bias, ethical issues or privacy concerns during the annotation process.

References

David Ifeoluwa Adelani, Hannah Liu, Xiaoyu Shen, Nikita Vassilyev, Jesujoba O Alabi, Yanke Mao, Haonan Gao, and Annie En-Shiun Lee. 2023. Sib-200: A simple, inclusive, and big evaluation dataset for topic classification in 200+ languages and dialects. *arXiv preprint arXiv:2309.07445*.

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Maxamed Axmed, et al. 2023. Mega: Multilingual evaluation of generative ai. *arXiv preprint arXiv:2303.12528*.

Shane Arora, Marzena Karpinska, Hung-Ting Chen, Ipsita Bhattacharjee, Mohit Iyyer, and Eunsol Choi. 2024. Calmqa: Exploring culturally specific long-form question answering across 23 languages. *arXiv preprint arXiv:2406.17761*.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2019. On the cross-lingual transferability of monolingual representations. *arXiv preprint arXiv:1910.11856*.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2023. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. *arXiv preprint arXiv:2308.16884*.

Alexis Conneau, Guillaume Lample, Ruty Rinott, Adina Williams, Samuel R Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. *arXiv preprint arXiv:1809.05053*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Alena Fenogenova, Artem Chervyakov, Nikita Martynov, Anastasia Kozlova, Maria Tikhonova, Albina Akhmetgareeva, Anton Emelyanov, Denis Shevelev, Pavel Lebedev, Leonid Sinev, et al. 2024. Mera: A comprehensive llm evaluation in russian. *arXiv preprint arXiv:2401.04531*.

Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Hiroki Iida, Masanari Ohi, Kakeru Hattori, Hirai Shota, Sakae Mizuki, Rio Yokota, and Naoaki Okazaki. 2024. Continual pre-training for cross-lingual llm adaptation: Enhancing japanese language capabilities. *arXiv preprint arXiv:2404.17790*.

Yancheng He, Shilong Li, Jiaheng Liu, Yingshui Tan, Weixun Wang, Hui Huang, Xingyuan Bu, Hangyu Guo, Chengwei Hu, Boren Zheng, et al. 2024. Chinese simpleqa: A chinese factuality evaluation for large language models. *arXiv preprint arXiv:2411.07140*.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *International Conference on Machine Learning*, pages 4411–4421. PMLR.

Kaiyu Huang, Fengran Mo, Hongliang Li, You Li, Yuanchi Zhang, Weijian Yi, Yulong Mao, Jinchen Liu, Yuzhuang Xu, Jinan Xu, et al. 2024. A survey on large language models with multilingualism: Recent advances and new frontiers. *arXiv preprint arXiv:2405.10936*.

Xu Huang, Wenhao Zhu, Hanxu Hu, Conghui He, Lei Li, Shujian Huang, and Fei Yuan. 2025. Benchmax: A comprehensive multilingual evaluation suite for large language models. *Preprint*, arXiv:2502.07346.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

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

Viet Dac Lai, Chien Van Nguyen, Nghia Trung Ngo, Thuat Nguyen, Franck Dernoncourt, Ryan A Rossi, and Thien Huu Nguyen. 2023. Okapi: Instruction-tuned large language models in multiple languages with reinforcement learning from human feedback. *arXiv preprint arXiv:2307.16039*.

Patrick Lewis, Barlas Oğuz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2019. Mlqa: Evaluating cross-lingual extractive question answering. *arXiv preprint arXiv:1910.07475*.

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. 2023. Cmmlu: Measuring massive multitask language understanding in chinese. *arXiv preprint arXiv:2306.09212*.

Noémi Ligeti-Nagy, Gergő Ferenczi, Enikő Héja, László János Laki, Noémi Vadász, Zijian Győző Yang, and Tamás Váradi. 2024. Hulu: Hungarian language understanding benchmark kit. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8360–8371.

Bill Yuchen Lin, Yuntian Deng, Khyathi Chandu, Faeze Brahman, Abhilasha Ravichander, Valentina Pyatkin, Nouha Dziri, Ronan Le Bras, and Yejin Choi. 2024. Wildbench: Benchmarking llms with challenging tasks from real users in the wild. *arXiv preprint arXiv:2406.04770*.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, et al. 2022. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9019–9052.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024a. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Chen Cecilia Liu, Iryna Gurevych, and Anna Korhonen. 2024b. Culturally aware and adapted nlp: A taxonomy and a survey of the state of the art. *arXiv preprint arXiv:2406.03930*.

Chen Cecilia Liu, Fajri Koto, Timothy Baldwin, and Iryna Gurevych. 2023. Are multilingual llms culturally-diverse reasoners? an investigation into multicultural proverbs and sayings. *arXiv preprint arXiv:2309.08591*.

Chen Cecilia Liu, Fajri Koto, Timothy Baldwin, and Iryna Gurevych. 2024c. Are multilingual llms culturally-diverse reasoners? an investigation into multicultural proverbs and sayings. *Preprint*, arXiv:2309.08591.

Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. 2020. Logiqa: A challenge dataset for machine reading comprehension with logical reasoning. *arXiv preprint arXiv:2007.08124*.

Mátyás Osváth, Zijian Győző Yang, and Karolina Kósa. 2023. Analyzing narratives of patient experiences: A bert topic modeling approach. *Acta Polytechnica Hungarica*, 20(7):153–171.

Chanjun Park, Hyeonwoo Kim, Dahyun Kim, Seonghwan Cho, Sanghoon Kim, Sukyung Lee, Yungi Kim, and Hwalsuk Lee. 2024. Open ko-llm leaderboard: Evaluating large language models in korean with ko-h5 benchmark. *arXiv preprint arXiv:2405.20574*.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. Xcopa: A multilingual dataset for causal commonsense reasoning. *arXiv preprint arXiv:2005.00333*.

Yusuke Sakai, Hidetaka Kamigaito, and Taro Watanabe. 2024. mcsqa: Multilingual commonsense reasoning dataset with unified creation strategy by language models and humans. *arXiv preprint arXiv:2406.04215*.

Neha Sengupta, Sunil Kumar Sahu, Bokang Jia, Satheesh Katipomu, Haonan Li, Fajri Koto, William Marshall, Gurpreet Gosal, Cynthia Liu, Zhiming Chen, et al. 2023. Jais and jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models. *arXiv preprint arXiv:2308.16149*.

Harman Singh, Nitish Gupta, Shikhar Bharadwaj, Dinesh Tewari, and Partha Talukdar. 2024. Indicgenbench: A multilingual benchmark to evaluate generation capabilities of llms on indic languages. *arXiv preprint arXiv:2404.16816*.

Jiaxing Sun, Weiquan Huang, Jiang Wu, Chenya Gu, Wei Li, Songyang Zhang, Hang Yan, and Conghui He. 2024. Benchmarking chinese commonsense reasoning of llms: From chinese-specifics to reasoning-memorization correlations. *arXiv preprint arXiv:2403.14112*.

DeepSeek Team. 2024a. Deepseek-r1-lite-preview is now live: unleashing supercharged reasoning power.

Qwen Team. 2024b. Qwq: Reflect deeply on the boundaries of the unknown. *Hugging Face*.

Yue Wang, Qiuzhi Liu, Jiahao Xu, Tian Liang, Xingyu Chen, Zhiwei He, Linfeng Song, Dian Yu, Juntao Li, Zhuosheng Zhang, et al. 2025. Thoughts are all over the place: On the underthinking of o1-like llms. *arXiv preprint arXiv:2501.18585*.

Jason Wei, Nguyen Karina, Hyung Won Chung, Yunxin Joy Jiao, Spencer Papay, Amelia Glaese, John Schulman, and William Fedus. 2024a. Measuring short-form factuality in large language models. *arXiv preprint arXiv:2411.04368*.

Jason Wei, Nguyen Karina, Hyung Won Chung, Yunxin Joy Jiao, Spencer Papay, Amelia Glaese, John Schulman, and William Fedus. 2024b. Measuring short-form factuality in large language models. *arXiv preprint arXiv:2411.04368*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Kang Min Yoo, Jaegeun Han, Sookyo In, Heewon Jeon, Jisu Jeong, Jaewook Kang, Hyunwook Kim, Kyung-Min Kim, Munhyong Kim, Sungju Kim, et al. 2024. Hyperclova x technical report. *arXiv preprint arXiv:2404.01954*.

Wenxuan Zhang, Mahani Aljunied, Chang Gao, Yew Ken Chia, and Lidong Bing. 2023. M3exam: A multilingual, multimodal, multilevel benchmark for examining large language models. *Advances in Neural Information Processing Systems*, 36:5484–5505.

A Related Works

A.1 Multilingual Benchmarks

With the emergence of more and more LLMs(Yoo et al., 2024; Sengupta et al., 2023; Fujii et al., 2024) in languages other than English, many multilingual or cross-lingual evaluation benchmarks(Huang et al., 2024) have been proposed. For example, CHARM(Sun et al., 2024), LogiQA(Liu et al., 2020), which include both Chinese and English, assess the model's reasoning ability for Chinese commonsense. mCSQA (Sakai et al., 2024), which includes languages from eight different countries, implements cross-lingual common sense understanding evaluation. Benchmarks like XNLI(Conneau et al., 2018), XQUAD(Artetxe et al., 2019), MLQA(Lewis et al., 2019), XStoryCloze(Lin et al., 2022), XCOPE(Ponti et al., 2020), and M3Exam(Zhang et al., 2023) include multiple languages, but most of these are for high- or medium-resource languages. Additionally, benchmarks like MEGA(Ahuja et al., 2023), which proposes a comprehensive large-model evaluation benchmark for 70 languages and Belebele(Bandarkar et al., 2023) introduce datasets for reading comprehension that include 122 languages. Works like XTREME(Hu et al., 2020) and SIB-200(Adelani et al., 2023) include many languages, including several low-resource languages. However, most of these are derived through translation and almost none capture the culture-specific aspects of low-resource countries.

A.2 Low-resource Language Benchmarks

In addition to multilingual benchmarks, a small number of low-resource language benchmarks have been proposed specifically for large models in small languages. Ko-H5(Park et al., 2024)proposes an evaluation benchmark for Korean LLMs, derived from existing datasets and reviewed by Korean experts. This benchmark also includes a private test set to ensure fair comparison, minimizing data contamination and overlap with popular training datasets. MERA(Fenogenova et al., 2024) introduces a comprehensive and standardized evaluation benchmark for Russian LLMs and foundational models, consisting of 21 tasks. IndicGenBench(Singh et al., 2024) is an Indian language benchmark built by translating existing datasets, covering various generative tasks such as cross-lingual summarization, machine translation, and cross-lingual question answering.Due to the difficulties of collecting low-resource corpora, these small language benchmarks are mainly based on translations of existing datasets. For Hungarian, there is currently only one project, HuLU(Ligeti-Nagy et al., 2024). HuLU is a language understanding benchmark specifically focused on Hungarian. The project first selects English data from GLUE and SuperGLUE and translates the English tasks into Hungarian to construct the benchmark.However, for low-resource countries, it is crucial to build evaluation benchmarks that focus on the real user queries, unique cultural aspects, and generative capabilities of small languages. This not only helps improve the performance of these language models in real-world applications but also promotes research in low-resource language processing, fostering cross-cultural exchange and understanding.

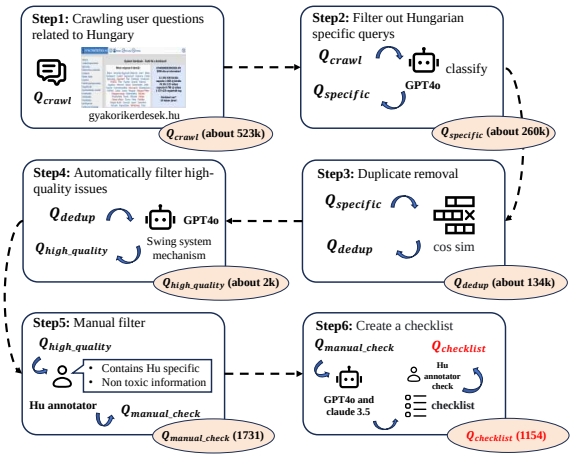

Figure 6: Constuction of HuWildBench.

B HuWildBench

B.1 Overall Construction Pipeline of HuWildBench

The construction of HuWildBench contains the following steps(Figure 6):

(1) Crawling: All user queries on the g13k website are systematically categorized into a multi-level tag system, which consists of 27 primary tags and 231 secondary tags. We manually reviewed all the secondary tags and selected 37 of them that contain a higher number of questions related to Hungary, such as népszokások (folk customs), egészségügyi-ellátások (healthcare services), and rezsi (overheads). We then crawled user queries under these 37 secondary tags, with a query date range from January 1, 2019, to August 31, 2024, resulting in the dataset Q_{crawl}(approximately 523K queries).

(2) Filtering for Hungary-specific content: Although these 37 secondary tags are closely related to Hungary, many of the questions still do not focus on Hungary-specific topics. Therefore, we used GPT-4o to classify the questions in Q_{crawl} (detailed prompt in Figure 9), resulting in a subset of approximately 260K questions, $Q_{specific}$[6].

(3) Deduplication: To ensure the diversity of questions, we performed deduplication on the Hungary-related questions within each secondary tag. The detailed process is outlined in Appendix B.2. After deduplication, the number of user questions in Q_{dedup} was approximately 134K.

[6]To validate the accuracy of GPT-4o, we manually labeled a random sample of about 2K questions. Based on the manual labeling, the recall rate of GPT-4o's classification was over 80%, with an accuracy exceeding 30%. This demonstrates that GPT-4o can capture the majority of Hungary-related queries.

(4) Automatic high-quality question filtering: To ensure that only high-quality questions are extracted from the question pool, we designed a comparative-based high-quality question filtering strategy, as detailed in Appendix B.3. After filtering, the resulting set $Q_{high\_quality}$ contained around 2K questions.

(5) Manual filtering: We hired a group of Hungarian native speakers to further manually review the questions in $Q_{high\_quality}$. Only questions that met the following two criteria were retained: First, the question should be Hungary-specific and closely related to Hungary. Second, the question must be harmless, meaning it does not contain inappropriate content such as pornography, violence, politics, or taboo topics specific to Hungary. The final set Q_{manual_check} consists of 1731 questions.

(6) Checklist construction: Based on WildBench (Lin et al., 2024), we constructed a checklist for each question. The purpose of the checklist is to assist the LLM judger in evaluating the answers. Each item in the checklist queries a specific aspect of the answer to a question. An example of the checklist can be found in Table 15, and the detailed construction method is provided in Appendix B.4. To ensure the relevance of the checklist items to the questions, we hired a Hungarian native speaker to review the checklist for quality, filtering out non-compliant items and performing deduplication. The filtering criterion was whether the item was suitable as an evaluation dimension for the model's response. To ensure the reliability of the LLM-as-judge, we filtered out user questions with fewer than 8 checklist items. The final set $Q_{checklist}$ contains 1154 questions. In the end, we obtained 1154 user questions along with their corresponding checklists.

B.2 Deduplication of similar questions

Since there are similar questions in the results obtained in the previous step, we design a method to remove similar ones. Specifically, we first use the Sentence-Transformer (Osváth et al., 2023) model to extract the Embedding of each question. Then, we calculate the cosine similarity between the embedding of each two questions, and choose a threshold between [0.15-0.25] according to the number of questions under each secondary tag.The larger the number of problems, the larger the threshold. Finally, one of the questions whose similarity is less than the threshold is removed, ensuring that the similarities between all questions are greater than the threshold.

B.3 Automatic high-quality question filtering

In order to automate the filtering of high-quality sample pots, as shown in Figure 10, we constructed a Prompt that allows the GPT-4o to select the two best Hungarian questions out of the five based on the criteria of linguistic complexity, Hungarian relevance, commonsense accuracy, context-dependence, answer diversity, ambiguity, reasoning requirements, socio-ethical considerations, format diversity, and breadth of knowledge and outputs their indexes in JSON format to output their indexes. Specifically, we first set the criteria for high-quality questions in Prompt. Then we ask GPT-4o to compare the input questions based on the criteria. In order to mitigate the occurrence of some high-quality questions being eliminated prematurely (or vice versa) when all the questions in the same batch are of high quality, we follow the following 3 rules when filtering the high-quality questions: 1. filter 2 high-quality questions from 5 questions at a time, instead of filtering 1 high-quality question directly from 2 questions. 2. use the Swiss system mechanism instead of the knockout mechanism. In each screening round, each question can win in the current round as long as it ensures that it wins in two comparisons, and it will not be eliminated directly because of a failure in one comparison. 3. Our question screening strategy eliminates 65% of the questions in each round, in order to ensure that each secondary label has a sufficient number of high-quality questions. We conducted different elimination rounds for questions under different labels, and finally got about 2K questions. Finally, in order to validate the strategy of high-quality question screening, we manually checked about 200 5-option-2 results, and the pass rate was more than 80%, which proved the effectiveness of the present strategy. The final constructed HuWildBench is shown in Table 14.

B.4 Checklist construction

In the process of building the Checklist, we mainly use large language models to generate it. In order to ensure the diversity of the Checklist and make the judge model can better evaluate the quality of the answers, here we use two non-open source LLM GPT-4o and Claude-3.5, each model generates a list of length 3-5. then we merge the two Checklists into one final Checklist. Checklists are then merged into a final Checklist. ultimately, each problem has a length of 6-10 and a Checklist. The details of our designed Prompt are shown in Figure 11 and the final constructed partial Checklist is shown in Table 15.

Score	Definition
Score 1-2	The response is very poor and does not make sense at all.
Score 3-4	The response is poor and does not help the user solve the problem meaningfully.
Score 5-6	The response is fair but has issues (e.g., factual errors, hallucinations, missing key information).
Score 7-8	The response is good but could be improved.
Score 9-10	The response is perfect and provides helpful information to solve the problem.

Table 10: Definition of scores.

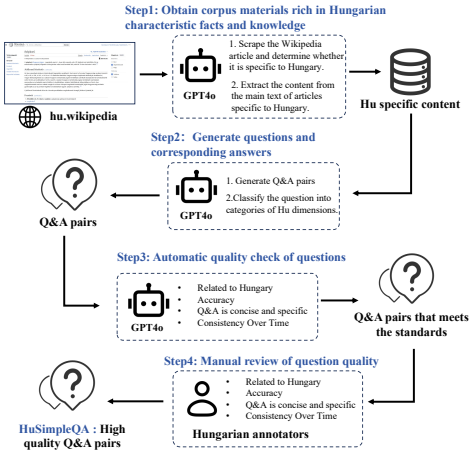

Figure 7: Constuction of HuSimpleQA.

C HuSimpleQA

C.1 Construction pipeline of HuSimpleQA

The question construction process for HuSimpleQA is illustrated in Figure 7 and consists of the following steps:

(1) Obtaining corpora rich in Hungary-specific facts and knowledge: First, we chose the Hungarian Wikipedia website [7] as the source of corpus material for question construction. We crawled all the entry pages and extracted their content. Next, we used GPT-4o to classify whether the entries were Hungary-specific, with the prompts detailed in Appendix C.2, Figure 12. [8] We then used GPT to extract factual information from the content of these entries suitable for question-answering. The extraction prompts are detailed in Appendix C.2, Figure 13. An example of the extracted factual information is shown in Figure 14. As a result, we obtained Hungary-specific factual information covering the eight distinct dimensions, totaling 4428 pieces of information.

(2) Generating questions and corresponding answers: We used the GPT-4 model to generate open-ended questions and corresponding answers based on the Hungary-specific factual information obtained in the previous step. The prompt used is detailed in Appendix C.3, Figure 15. In this step, we generated a total of 9424 questions based on 4K entries. We then classified the generated questions according to the eight Hungary-specific dimensions outlined in Section 2.1, using GPT with the prompt detailed in Appendix C.3, Figure 16.

(3) Automatic quality checking of questions: To ensure the quality of the questions, we used GPT to check and filter the generated questions. We set the following four criteria, retaining only those questions that met all four standards (the corresponding prompt is detailed in Appendix C.4):

- Criterion 1: Hungary-specific: The content of the question-answer pair must align with the eight Hungary-specific dimensions proposed in this paper.

- Criterion 2: Accuracy: The information in the question-answer pair must align with the entry description and facts, and the answer should not be directly inferable from the question itself.

- Criterion 3: Concise and specific: The question and answer should be clear and concise, with no redundant information. The question should not contain nested sub-questions. The phrasing should be specific and direct, matching the scope of the answer (e.g., for time and location questions, the exact year/month/day/district/city must be specified).

- Criterion 4: Consistency Over Time: The answer should remain consistent over time and not be influenced by future events.

After the automatic checking process, we retained 5503 questions corresponding to 2666 entries.

(4) Manual review of question quality: To further ensure the quality of the questions, we hired Hungarian native speakers to manually review the questions. Annotators checked whether the questions met the four criteria mentioned in Step 3. During the annotation process, each question was assigned to two annotators, who received the questions but not the answers. A question was considered valid and retained only if both annotators agreed that it met all four criteria and that the provided answer matched the original reference answer. Detailed procedures are provided in Appendix C.4. After these four steps, we obtained a total of 1343 questions, with their distribution across the eight Hungary-specific dimensions shown in Table 16.

C.2 Obtaining corpora rich in Hungary-specific facts and knowledge

In the process of filtering Wikipedia entries with Hungarian characteristics, we randomly selected entries and provided both the entries and the first two paragraphs of the main content to GPT-4o (prompt shown in Figure 12) to determine if they were related to Hungary. If the entry was deemed relevant, it was categorized based on the eight characteristic dimensions proposed in this paper. At this stage, an "Others" category was added to ensure the focus on the eight thematic categories and to exclude interference from entries that belonged to other themes. The screening process stopped once the total number of Hungarian characteristic entries reached 8,000.

Due to the uneven distribution of entry themes on Wikipedia, with more data in the categories of figures, geography and place, and history, we filtered the data based on the proportion of themes, ensuring that no single category exceeded 1,000 entries. This resulted in 4428 characteristic entries covering the eight dimensions.

Given the varying lengths of content describing entries on Wikipedia, we aimed to streamline the complexity of constructing subsequent question-answer pairs.

[7] https://hu.wikipedia.org/

[8] We did not classify all the pages but instead randomly selected pages until we reached 8K Hungary-specific entries, at which point we stopped.

To achieve this, we first employed GPT-4o to extract key factual information from the main text of each entry. This step aims to avoid any deviation from the theme caused by redundant content during the construction of the question-answer pairs (prompt shown in Figure 13). The results of the factual information extraction are presented in Figure 14.

C.3 Generating questions and corresponding answers

Based on the key information extracted and the provided entries, we utilized GPT-4o to generate 1-3 Hungarian characteristic knowledge open-ended question-answer pairs for each entry (prompt details in Figure 15). In total, 9,424 question-answer pairs were generated based on 4,000 entries.

Given that the focus and orientation of the generated question-answer pairs may differ from the original entry categories, this paper employed GPT-4o to reclassify the obtained question-answer pairs, with the corresponding prompt detailed in Figure 16.

C.4 Automatic quality checking of questions

We focused on evaluating the quality of the generated questions from two perspectives: the information contained in the question-answer pairs and the formulation of the questions. The quality assessment was divided into two stages, with each stage generating two evaluation metrics. The first stage focuses on the relevance and correctness of the question information. We provided GPT-4o with the entry, its corresponding key information, and the generated question-answer pairs to verify whether the questions contain Hungarian-specific content and whether the information in the question-answer pairs aligns with the provided background material (prompt shown in Figure 17).

Second, from the perspective of the precision of the question formulation, we only provided GPT-4o with the generated question-answer pairs to simulate real user response scenarios. This step emphasized evaluating whether the questions were based on objective facts, and whether the descriptions were precise and specific enough to allow independent answering without ambiguity. Additionally, we required that the answers remain unaffected by future events, ensuring consistency across any time period and guaranteeing the long-term validity of the dataset (prompt details in Figure 18).

Based on the results of the above automated quality assessment, we retained only those question-answer pairs that passed all four evaluation criteria, resulting in a final set of 5,503 questions.

C.5 Manual review of question quality

To further ensure the quality of the constructed question-answer pairs, we engaged native Hungarian speakers to review these questions. Each question was independently reviewed by two annotators who could only see the questions and not the reference answers. The annotation process consisted of three main steps.

First, the annotators were required to determine whether the given questions aligned with the eight Hungarian-specific knowledge dimensions proposed in this paper. Next, they evaluated whether the questions met the four assessment criteria outlined in Step 3, ensuring that the questions were objectively framed, precisely described, had unique answers, contained correct information, and maintained consistent answers over time. Finally, if a question satisfied all the above criteria, the annotators provided the correct answer. During this process, annotators were permitted to search for relevant information online and provided reference sources for their answers.

To address potential issues such as overly obscure questions or non-fixed answers, we used GPT-4o to verify whether the annotated results matched the generated reference answers. If the annotated answer matched the reference answer, it was labeled as "CORRECT"; otherwise, it was labeled as "INCORRECT" (prompt details in Figure 19). We selected question-answer pairs that both annotators deemed valid, Hungarian-specific, and consistent with the original reference answers as candidates for the HuSimpleQA dataset, resulting in a total of 2953 questions.

Considering that the HuSimpleQA dataset should exhibit diversity and broad coverage, we removed question-answer pairs belonging to the same entry, retaining only one question-answer pair per entry that best met the construction and evaluation criteria. This step reduced the similarity in knowledge assessment (prompt details in Figure 20).

Through this process, we obtained a total of 1,343 pieces of Hungarian-specific open-ended question-answer pairs, with the category distribution shown in Table 16.

C.6 Inference prompt

We constructed prompts in two languages for model inference, as shown in Figure 21, while also instructing the model to provide a confidence score (ranging from 1 to 100)to measure the model's confidence in its generated answers.

C.7 LLM-as-judge

Following the approach of SimpleQA, we employed GPT as a judge to evaluate the correctness of responses generated by large language models. The evaluation criteria for this step were similar to those used in the manual review process of Step 4. In addition to the classification labels "CORRECT" and "INCORRECT" we introduced an additional category, "NOT ATTEMPTED" to further assess the model's ability to respond to questions and the breadth of its knowledge coverage (prompt details in Figure 22).

For this dataset, we designed two extra evaluation metrics to measure the performance of the model's responses. The first metric, Correct Given Attempted (CGA), measures the accuracy of responses excluding

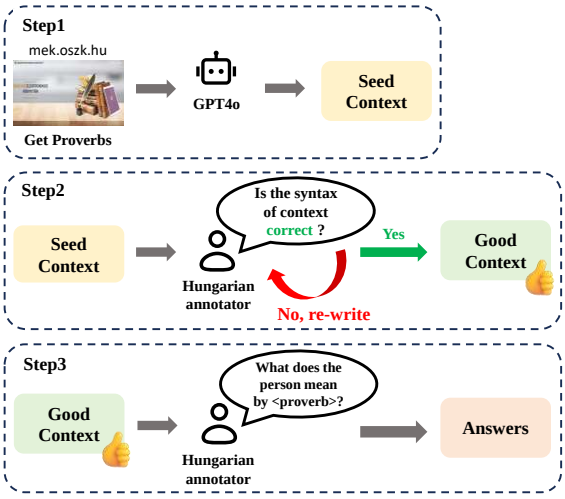

Figure 8: Constuction of HuProverbRea.

questions labeled as "NOT ATTEMPTED" The second metric, F-Score, evaluates the correctness rate across all attempted responses. The formulas for two metrics are as follows:

$$\text{CGA} = \frac{c}{c+i} \tag{2}$$

$$\text{F-Score} = \frac{2}{\frac{c+i}{c} + \frac{c+i+n}{c}} = \frac{2c}{2c+2i+n} \tag{3}$$

Here, c represents the number of correctly answered questions, i represents the number of incorrectly answered questions, and n represents the number of not attempted questions.

D HuProverbRea

D.1 Construction pipeline of HuProverb

The proverbs in HuProverbRea are from 2 separate sources. The first part, 733 traditional Hungarian proverbs, are collected from the website[9], where each proverb is assigned an English or Hungarian explanation. The other 402 proverbs, focusing on abbreviations and Internet slang, are manually collected and explained by native speakers of Hungarian. Inspired by MAPS (Liu et al., 2024c), we adopt a human-in-loop pipeline to generate and refine the context for each Hungarian specific usage, as shown in Figure 8. For each proverb, we first let GPT4 generate a seed context where the proverb is used. Then, we assign it to a Hungarian native speaker to check whether this context is grammatically correct and the use of slang is appropriate. If not, the annotator is required to manually write down a new context for the saying, which will be sent back to another annotator for inspection again. We continue the above procedures until all contexts pass the quality check. It's worth noting that each option of the 2CQ setting is manually constructed by a human annotator, and only when it passes the double check of two other annotators could it be considered usable. We choose not to involve LLM in this part because designing correct/incorrect options requires deep understanding of sayings, LLM may generate ambiguous options if it does not understand the proverb used in the context, and such pre-provided ambiguous options may negatively influence the creativity of the annotators. Finally, we obtain 1,135 Hungarian proverbs, each equipped with a context, an English explanation, and two candidate options for question *"What does the speaker mean by the saying?"*.

D.2 More examples of HuProverbRea

The example of HuProverbRea is shown in Figure 23 and Figure 24. The prompt for judging HuProverbRea is shown in Figure 25.The prompt for model inference on HuProverbRea is shown in Figure 27 and 26.

D.3 Differences in model performance rankings on HuproverbRea and MAPS

We counted the differences in model performance rankings on several other language types of the HuproverbRea and Maps datasets, and the results are shown in Table 12.

E HuMatchingFIB and HuStandardFIB

E.1 Construction of HuMatchingFIB and HuStandardFIB

The questions for both HuMatchingFIB and HuStandardFIB are sourced from the Hungarian National Public Education Portal(NKP)[10], a comprehensive platform for cultural funding and support in Hungary. This portal connects artists, cultural organizations, and the public with resources and opportunities to promote Hungarian culture both domestically and internationally. Notably, this website is a government initiative, reflecting the collaborative efforts between the Hungarian government and the European Union, particularly through projects or programs supported by the European Social Fund. After extracting the original questions from the NKP website, we engaged native Hungarian speakers to annotate the data. The annotation process involved manually extracting questions and their corresponding answers[11] , classifying the questions into appropriate

[9] https://mek.oszk.hu/

[10] https://www.nkp.hu/

[11] The questions for HuMatchingFIB and HuStandardFIB on the NKP website are not in plain text but are instead presented in interactive modules, and the answers can only be obtained through additional interactive operations. As a result, the commonly used data cleaning and extraction methods for LLM pre-training datasets [reference] are unable to accurately extract these questions and their corresponding answers . Consequently, it can be concluded that the likelihood of these questions being incorporated into the LLM pre-training data in their proper format is minimal, thereby significantly reducing the potential risk of data leakage. This ensures the reasonableness and effectiveness of the test sets for HuMatchingFIB and HuStandardFIB.

model	correct	incorrect	not_attempted	correct given attempted	F-Score
GPT-4o	50.3	40.61	9.09	55.33	52.69
GPT-4o-mini	24.52	74.07	1.42	24.87	24.69
Deepseek-V3	32.71	64.08	3.2	33.8	33.24
QwQ	9.17	52.68	38.15	14.82	11.33
Deepseek-R1	34.58	62.15	3.28	35.75	35.15
Qwen2.5-7B-Instruct	5.29	84.13	10.58	5.92	5.59
Qwen2.5-72B-Instruct	15.05	78.61	6.33	16.07	15.54
Llama-3.1-8B-Instruct	14.9	80.25	4.84	15.66	15.27
Llama-3.1-70B-Instruct	36.36	61.03	2.61	37.34	36.84
o1-mini	16.24	44.19	39.57	26.88	20.25

Table 11: The complete result of HuSimpleQA

Rank	HuProverbRea	MAPS(bn)	MAPS(id)	MAPS(de)	MAPS(ru)	MAPS(zh)
1	GPT-4o	GPT-4o (-)	GPT-4o-mini (↑3)	Llama-3.1-Instruct-70B (↑1)	Deepseek-V3 (↑2)	GPT-4o (-)
2	Llama-3.1-Instruct-70B	Deepseek-V3 (↑1)	Qwen2.5-Instruct-72B (↑4)	GPT-4o-mini (↑2)	o1-mini (↑5)	Qwen2.5-Instruct-72B (↑4)
3	Deepseek-V3	Deepseek-R1 (↑2)	Llama-3.1-Instruct-70B (↓1)	Qwen2.5-Instruct-72B (↑3)	GPT-4o (↓2)	Deepseek-V3 (-)
4	GPT-4o-mini	Qwen2.5-Instruct-72B (↑2)	GPT-4o (↓3)	GPT-4o (↓3)	Qwen2.5-Instruct-72B (↑2)	o1-mini (↑3)
5	Deepseek-R1	o1-mini (↑2)	Deepseek-V3 (↓2)	Llama-3.1-Instruct-8B (↑4)	GPT-4o-mini (↓1)	Llama-3.1-Instruct-70B (↓3)
6	Qwen2.5-Instruct-72B	GPT-4o-mini (↓2)	o1-mini (↑1)	o1-mini (↑1)	Deepseek-R1 (↓1)	GPT-4o-mini (↓2)
7	o1-mini	Llama-3.1-Instruct-70B (↓5)	Deepseek-R1 (↓2)	Deepseek-V3 (↓4)	Qwen2.5-Instruct-7B (↑3)	Qwen2.5-Instruct-7B (↑3)
8	QwQ	Llama-3.1-Instruct-8B (↑1)	Llama-3.1-Instruct-8B (↑1)	Qwen2.5-Instruct-7B (↑2)	Llama-3.1-Instruct-70B (↓6)	Llama-3.1-Instruct-8B (↑1)
9	Llama-3.1-Instruct-8B	Qwen2.5-Instruct-7B (↑1)	Qwen2.5-Instruct-7B (↑1)	Deepseek-R1 (↓4)	Llama-3.1-Instruct-8B (-)	Deepseek-R1 (↓4)
10	Qwen2.5-Instruct-7B	QwQ (↓2)	QwQ (↓2)	QwQ (↓2)	QwQ (↓2)	QwQ (↓2)

Table 12: The LLM rankings on HuProverbRea and MAPS datasets

categories, and filtering out questions that required additional modalities such as images, tables, audio, or video. This ensured that only purely language-based questions were retained. Through this process, we obtained 278 questions for the HuMatchingFIB task and 93 questions for the HuStandardFIB task, as shown in Table x.

E.2 More examples of HuMatchingFIB and HuStandardFIB

Examples of questions from HuMatchingFIB and HuStandardFIB are provided in Figure 28 and Figure 29. The prompt for model inference on HuMatchingFIB and HuStandardFIB is shown in Figure 36 and Figure 30.

F LRM's reasoning process on OpenHuEval

F.1 Segment answer into thoughts on HuSimpleQA

First, we use GPT-4o to break down the answers into thoughts. This is done in two steps: the first step is to identify expressions that may be a shift in thought (the prompt is shwon in Figure 39), and the second step is to confirm whether it is indeed a shift in thought (the prompt is shwon in Figure 40). Then, We utilized the LLM to evaluate whether each idea would lead to the correct answer, the prompt is shown in Figure 41. We consider a confident score of 2 as the correct thought. The thought segmentation results can be seen in Figure 42.

F.2 Reasoning Segmentation Examples on HuMatchingFIB

In order to analyze the reasoning process, we break down the models' prediction into segments and the following two figures illustrated a same question reasoning example on two LRMs Figure 43 and Figure 44.

G Information of the Annotators

Task	# Anotater	Total working hours
HuSimpleQA	14	161.9
HuWildBench	5	55.2
HuProverbRea	15	118.2
HuMatchingFIB and HuStandardFIB	8	84.5

Table 13: Information of the Annotators

We submitted the annotation task online to a professional data annotation company, which organized annotators to complete the annotation work. In the construction phase of OpenHuEval, the annotations were carried out by professional annotators who are native Hungarian speakers. Table 13 shows the number of annotators and the total time spent on each task. All annotators involved in this project hold a bachelor's or master's degree, with academic backgrounds in fields such as Social Sciences, Translating and Interpreting, English Studies, and IT Engineering. They all possess the ability to distinguish subtle aspects of the Hungarian language and handle Hungarian-specific knowledge effectively.

Hungarian-specific dimensions	Count	Question Example	Question Example (en)
LCC	365	a kérdés az: Mi lesz a jövőben a szocializmus alatt megépül sok panellel? a leírás: Úgy tudom, hogy kb 60 éves életciklusra tervezték őket. Magyarországon (és a környező országokban is) rengeteg ember él bennük. Mi fog történni akkor, ha lakhatatlanná kezdenek válni? Mi lesz azzal a sok emberrel? Meg a panelokkal?	The question is: What will happen in the future to the many panels built under socialism? Description: I understand they are designed for a life cycle of about 60 years. There are a lot of people living in them in Hungary (and surrounding countries). What will happen if they start to become uninhabitable? What will happen to all those people? And the panels?
EP	201	a kérdés az: A kárpátaljai magyarok Ukrajnában oroszul vagy ukránul tanultak meg a 2000-es évek közepén? a leírás: Mit tanítottak az iskolákban? Mennyire reális az, hogy valakire szinte semmi se ragad a környezetéből? Vannak olyan tömb területek ahol mondjuk egy magyar gyereknek egyáltalán nem kell helyi ukránokkal beszélnie? Egyáltalán a helyi ukránok ukránul beszéltek a 2000-es években?	The question is: Did Hungarians in Transcarpathia learn Russian or Ukrainian in Ukraine in the mid-2000s? Description: What was taught in schools? How realistic is it that almost nothing sticks to someone from their environment? Are there block areas where, say, a Hungarian child doesn't have to speak to local Ukrainians at all? Did local Ukrainians even speak Ukrainian in the 2000s?
PPL	299	a kérdés az: Mi történt azzal, aki az 50-es években a felhívás ellenére sem jegyzett "önként" békekölcsönt? Érhette ezért retorzió az embert? a leírás: Persze nyilván volt, amilyen "bolondos" idők jártak nálunk akkortájt. Biztos kikiáltották reakciósnak vagy fasisztának, meg a "népi demokrácia" ellenségének.	The question is: What happened to the man who did not "voluntarily" subscribe to a peace charter in the 1950s, despite the call? Could he have been retaliated against for this? The description. He must have been branded a reactionary or a fascist or an enemy of 'people's democracy'.
BF	289	a kérdés az: Meddig tartható fent Magyarország negatív külkereskedelmi mérlege? a leírás: Nem a háború óta, hanem már 2021 nyarától folyamatosan negatív az ország külkereskedelmi mérlege. Júliusban és augusztusban összesen több, mint 1000 milliárd forintnyi mínusz keletkezett. Persze a többi hónap nem volt ennyire szörnyű, de ez csak erre az évre már több, mint 2000 milliárd forintnyi mínusz. Változatlan devizaimport mellett a mérséklődött energiaárakkal is több, mint 1000 milliárdos negatív mérleg hozható össze 2023-ban. Meddig lehet ezt tovább folytatni? Meddig elég a devizatartalék a hiány pótlására?	The question is: How long can Hungary maintain a negative trade balance? Description. In July and August there was a total deficit of more than HUF 1000 billion. Of course, the other months were not so bad, but for this year alone it is already more than HUF 2000 billion in deficit. Even with unchanged foreign exchange imports and moderating energy prices, a negative balance of more than 1,000 billion in 2023 could be created. How long can this go on? How long will foreign exchange reserves be enough to cover the deficit?

Table 14: Examples of HuWildBench. The rightmost column is the English translation of the original OpenHuEval examples, used for visualization.

Hungarian-specific dimensions	Question Example	Checklist
LCC	a kérdés az: Mi lesz a jövőben a szocializmus alatt megépül sok panellel? a leírás: Úgy tudom, hogy kb 60 éves életciklusra tervezték őket. Magyarországon (és a környező országokban is) rengeteg ember él bennük. Mi fog történni akkor, ha lakhatatlanná kezdenek válni? Mi lesz azzal a sok emberrel? Meg a panelokkal?	"Does the response provide an analysis of the current condition and expected lifespan of the panel buildings in Hungary and neighboring countries?", "Does the response address the expected lifespan of panel buildings and their current age?", "Are there any historical or international examples included to illustrate possible outcomes or strategies?", "Does the response consider the economic implications of renovating or replacing panel buildings?", "Does the response include potential government or private sector plans or policies addressing the future of these buildings and their residents?", "Does the answer discuss potential scenarios for when these buildings become uninhabitable?", "Are environmental and urban planning aspects of dealing with aging panel buildings mentioned?", "Is there an explanation of possible solutions or government plans for relocating residents?"
EP	a kérdés az: A kárpátaljai magyarok Ukrajnában oroszul vagy ukránul tanultak meg a 2000-es évek közepén? a leírás: Mit tanítottak az iskolákban? Mennyire reális az, hogy valakire szinte semmi se ragad a környezetéből? Vannak olyan tömb területek ahol mondjuk egy magyar gyereknek egyáltalán nem kell helyi ukránokkal beszélnie? Egyáltalán a helyi ukránok ukránul beszéltek a 2000-es években?	"Does the answer provide information on the language predominantly spoken by local Ukrainians in Transcarpathia in the 2000s?", "Does the response discuss the social and linguistic dynamics in areas with significant Hungarian populations, including interactions with local Ukrainians?", "Does the response clearly explain the educational policies and language of instruction in schools for Hungarians in Transcarpathia during the mid-2000s?", "Does the response accurately describe the language of instruction in Transcarpathian Hungarian schools in the mid-2000s?", "Does the response consider the historical and political context of language policies in Ukraine during this period?", "Does the response provide insight into whether local Ukrainians predominantly spoke Ukrainian during the 2000s?", "Does the response offer a balanced view of cultural and linguistic integration in Transcarpathia during the specified period?", "Does the answer address the likelihood of a Hungarian child not acquiring any local language skills from their environment?", "Does the response discuss the existence of predominantly Hungarian areas where interaction with local Ukrainians might be limited?"
PPL	a kérdés az: Mi történt azzal, aki az 50-es években a felhívás ellenére sem jegyzett "önként" békekölcsönt? Érhette ezért retorzió az embert? a leírás: Persze nyilván volt, amilyen "bolondos" idők jártak nálunk akkortájt. Biztos kikiáltották reakciósnak vagy fasisztának, meg a "népi demokrácia" ellenségének.	"Does the answer address the political labels mentioned in the description (e.g., 'reactionary', 'fascist', 'enemy of people's democracy')?", "Does the response differentiate between official consequences and social/societal repercussions for not subscribing to the peace loan?", "Does the response address potential consequences for individuals who did not subscribe to the peace loan, with references to historical examples or documentation?", "Does the response provide a balanced view, considering both potential punitive measures and any instances of leniency or exceptions, if applicable?", "Is there a clear explanation of what 'békekölcsön' (peace loan) was and its significance during that time period?", "Does the response accurately describe the historical context of the 1950s in Hungary?", "Is there an analysis of the societal and governmental attitudes toward dissenters in Hungary during the 1950s, including any possible labels or accusations they might have faced?", "Does the response provide specific examples of potential retaliations against those who didn't subscribe to the peace loan?"
BF	a kérdés az: Meddig tartható fent Magyarország negatív külkereskedelmi mérlege? a leírás: Nem a háború óta, hanem már 2021 nyarától folyamatosan negatív az ország külkereskedelmi mérlege. Júliusban és augusztusban összesen több, mint 1000 milliárd forintnyi mínusz keletkezett. Persze a többi hónap nem volt ennyire szörnyű, de ez csak erre az évre már több, mint 2000 milliárd forintnyi mínusz. Változatlan devizaimport mellett a mérséklődött energiaárakkal is több, mint 1000 milliárdos negatív mérleg hozható össze 2023-ban. Meddig lehet ezt tovább folytatni? Meddig elég a devizatartalék a hiány pótlására?	"Does the response analyze Hungary's current foreign exchange reserves and their sufficiency in covering the trade deficit?", "Is there an exploration of historical trends and comparisons to similar situations in other countries to provide context?", "Is the impact of energy prices on the trade balance accurately assessed in the response?", "Does the response offer a clear and supported prediction or timeframe for how long Hungary can sustain its negative trade balance?", "Is there an analysis of the factors affecting Hungary's foreign exchange reserves and their ability to cover the deficit?", "Does the answer provide a clear timeline or projection for how long the negative balance can be sustained?", "Are there comparisons made to similar situations in other countries or historical precedents in Hungary?", "Does the response accurately explain the current state of Hungary's foreign trade balance?"

Table 15: Examples of HuWildBench Checklist.

Hungarian-specific dimensions	Count	Question-Answer Pairs	Question-Answer Pairs (en)
L	10	Question1: Mit jelent a Kara török eredetű régi magyar személynév? Answer1: fekete Question2: Melyik régi magyar név a Pantaleon megfelelője? Answer2: Pentele	Question1: What does the old Hungarian personal name Kara of Turkish origin mean? Answer1: black Question2: Which old Hungarian name is the equivalent of Pantaleon? Answer2: Pentele
H	169	Question1: Melyik király nevezte ki Szapolyai Imrét szepesi örökletes főispánná 1465-ben? Answer1: Mátyás király Question2: Melyik várost foglalta el Báthory Gábor 1610. december 11-én? Answer2: Szeben	Question1: Which king appointed Imre Szapolyai as the hereditary ispán of Szepes in 1465? Answer1: King Matthias Question2: Which city was captured by Gabriel Báthory on December 11, 1610? Answer2: Sibiu
LCC	228	Question1: Melyik magyar film nyerte el a FIPRESCI-díjat az 1983-as Cannes-i Nemzetközi Filmfesztiválon? Answer1: Szerencsés Dániel Question2: Melyik legendára épít az 'Eredet / Origins' táncjáték? Answer2: Csodaszarvas-legendára	Question1: Which Hungarian film won the FIPRESCI Prize at the 1983 Cannes International Film Festival? Answer1: Lucky Daniel Question2: Which legend is the 'Origin / Origins' dance play based on? Answer2: Legend of the Miraculous Deer
EP	70	Question1: Melyik városban alapították a Gandhi Gimnáziumot 1994-ben? Answer1: Pécsen Question2: Melyik évben alapította a Magyar Tudományos Akadémia az Acta Juridica Hungarica folyóiratot? Answer2: 1959	Question1: In which city was the Gandhi High School founded in 1994? Answer1: Pécs Question2: In which year did the Hungarian Academy of Sciences establish the journal Acta Juridica Hungarica? Answer2: 1959
GP	70	Question1: Melyik magyar vármegyében található Nemesmedves? Answer1: Vas vármegyében Question2: Mi a neve Magyarország legmagasabban fekvő csillagvizsgálójának, amely a Piszkéstetőn található? Answer2: Piszkéstetői Obszervatórium	Question1: In which Hungarian county is Nemesmedves located? Answer1: Vas county Question2: What is the name of Hungary's highest observatory, located on Piszkés Peak? Answer2: Piszkés Peak Observatory
F	452	Question1: Nádasdy Kálmán hányszor kapott Kossuth-díjat élete során? Answer1: Háromszor Question2: Balogh József melyik magyar városban született 1946. április 15-én? Answer2: Nagykanizsán	Question1: How many times did Kálmán Nádasdy receive the Kossuth Prize during his lifetime? Answer1: Three times Question2: In which Hungarian city was József Balogh born on April 15, 1946? Answer2: Nagykanizsa
PPL	179	Question1: Melyik szervezet jogkörét vette át a Népgazdasági Tanács 1949. június 11-én? Answer1: Gazdasági Főtanács Question2: Melyik törvénycikk rendelkezett 1878-ban Magyarországon a réz-váltópénz szaporításáról? Answer2: 1878. évi VI. törvénycikk	Question1: Which organization's authority was taken over by the National Economic Council on June 11, 1949? Answer1: Supreme Economic Council Question2: Which statute regulated the increase of copper coinage in Hungary in 1878? Answer2: Act VI of 1878
BF	29	Question1: Milyen néven működött az ÉVITERV 1954-től az 1980-as évek elejéig? Answer1: ÉM Szerelőipari Tervező Vállalat Question2: Melyik cég gyártotta a Puli autótípust a gyártás kezdeti időszakában? Answer2: HÓDGÉP	Question1: Under what name did ÉVITERV operate from 1954 to the early 1980s? Answer1: ÉM Installation Industry Design Company Question2: Which company manufactured the Puli car model in the early production period? Answer2: HÓDGÉP

Table 16: Examples of HuSimpleQA. The rightmost column is the English translation of the original OpenHuEval examples, used for visualization.

"""Given the following question, identify whether it has a characteristic related to Hungary. A question is considered to have a Hungarian characteristic if it meets any of the following criteria:
Hungary–Specific Context: The question itself directly references or relates to Hungary. For example, "What is the capital of Hungary?" clearly has a Hungarian characteristic.
Hungary–Specific Answer: The question might not directly reference Hungary, but the answer would vary depending on the country, particularly Hungary. For example, "What is the minimum wage according to labor laws?" The answer would depend on Hungary's laws and practices.
Hungary–Specific Context and Answer: Both the question and the likely answer have strong connections to Hungary. For example, "How do you view Hungary's 2024 foreign policy?" is likely to have both the question and the answer centered on Hungary.
Any Other Model–Identified Hungarian Characteristic: If the model identifies a Hungarian characteristic based on context, culture, or any other relevant factors.
There are several special rules to follow:
The language of the question should not be used as an evidence.
For a question to which the Hungarian answer is not significantly different from the answer of the rest of the world, the question is not considered having a Hungarian characteristic.
If a question only mentions a Hungarian–related term, such as the Hungarian currency, the forint, or a certain place in Hungary, but the question itself is not more related to Hungary's cultural, social, political, economic, military, life, etc., the question is not considered having a Hungarian characteristic.
If the answer to a question is open–ended, for example, "Will you buy a flower for your mom on Mothers' Day?", the question is not considered having a Hungarian characteristic.
Please respond strictly in JSON format. Do not include any additional text outside the JSON structure.
{
"Question": "[The original question]",
"HasHungarianCharacteristic": "yes/no",
"Reason": "[Explanation for why this question was classified as having a Hungarian characteristic]"
"Score": "[This score is used to evaluate how relevant this issue is to Hungary, with 0 being the lowest and 10 being the highest.]
}
The question is: <question>.
Note that each question is composed of a question itself and a question description."""

Figure 9: Prompt for Automatic Filtering of User Questions related to Hungarian Features (HuWildBench).

"""# Instruction
You are an expert responsible for evaluating the capabilities of a language model in handling questions related to the Hungarian language and context.
The questions are sourced from the Gyakori kérdések website, and the objective is to assess the model's performance by selecting the best questions based on a set of criteria.
You will be given five Hungarian questions.
Based on the 9 criteria listed below, select the two best questions.

Evaluation Criteria
<|begin of evaluation|>
1. Linguistic complexity: Does the question contain complex syntactic structures and rich vocabulary, testing the model's ability to process complex language?
2. Hungarian-specific relevance: Is the question highly relevant to Hungarian culture, society, history, or daily life, testing the model's understanding of Hungary-specific context?
3. Requirement for common knowledge and factual accuracy: Does the question require knowledge of Hungarian common sense or factual information, allowing for the evaluation of the model's knowledge base and accuracy?
4. Context dependency: Does the question require the model to understand or infer from the context, testing the model's ability to use prior or surrounding information?
5. Answer diversity: Does the question allow for multiple reasonable answers, testing the model's creativity and ability to generate diverse responses?
6. Ambiguity: Does the question contain ambiguity or multiple meanings, testing the model's ability to handle uncertain or vague information?
7. Reasoning requirement: Does the question require logical reasoning or causal inference, testing the model's ability to analyze and reason through complex information?
8. Social and ethical considerations: Does the question involve social, ethical, or moral issues, testing the model's ability to generate responses that align with ethical standards?
9. Format diversity: Does the question come in a unique format (e.g., multiple choice, open-ended, narrative, etc.), testing the model's ability to handle different types of question formats?
10. Breadth of knowledge: Does the question cover a broad range of knowledge areas (e.g., science, arts, technology), testing the model's general knowledge across various domains.
<|end of evaluation|>

Questions
<|begin of questions|>
1. <question0>.
2. <question1>.
3. <question2>.
4. <question3>.
5. <question4>.
<|end of questions|>
Note that each question is composed of a question itself and a question description.

Output format
Your output should be in JSON format as follows:
Please respond strictly in JSON format. Do not include any additional text outside the JSON structure.
{
    "question_indices": [a list of the indices of the two best question in int type],
}"""

Figure 10: Prompt for Automatic Filtering of High-Quality Question (HuWildBench).

"""You are a model designed to assist in evaluating responses to questions.
You will receive a question about Hungary, and your task is to provide a list of 3–5 evaluation criteria.
Each item in the list should be a distinct angle for assessing whether the response to the question meets the required standard.
For example, if the question is: "Is a monthly income of $1000 sufficient to cover normal living expenses in the capital city of Hungary?", the list could include criteria such as:
1.Does the response comprehensively outline all relevant living expenses in Budapest?
2.Are the amounts mentioned for each expense aligned with objective facts?
3.Does the response provide an overall conclusion on whether $1000 is enough for living expenses?
Each criterion should assess a different aspect of the response, ensuring no overlap in evaluation angles.
Please respond strictly in JSON format. Do not include any additional text outside the JSON structure.
{
"Checklist": "[The evaluation criteria list]"
}
The question is: <question>.
Note that each question is composed of a question itself and a question description."""

Figure 11: Prompt for Checklist Construction (HuWildBench).

```
"""
- Role: Expert in Hungarian Culture and Data Classification
- Background: You are tasked with classifying data that is deeply related to
Hungarian-specific content. This data may involve Hungarian history, culture, art,
folklore, language, traditions, tourism, and more. Your expertise is critical to
ensuring the classification is precise and adheres strictly to Hungarian cultural
relevance.
- Goal: Analyze the detailed description in the input data and categorize it into
one of the following nine predefined categories:
1. Language: Content related to Hungarian language, including proverbs, idioms, or
linguistic knowledge.
2. History: strictly for content describing specific historical events or
developments in Hungary. Examples include wars, revolutions, significant treaties,
or influential periods of political or societal change.
3. Life, Culture, and Customs: Hungarian religion, etiquette, cultural practices,
holidays, and daily life (including tourism).
4. Education and Profession: Information on Hungary's education system or associated
occupations.
5. Geography and Place: Hungarian geography, city locations, landmarks, and travel-
related content.
6. Figure: Notable Hungarian individuals and their achievements.
7. Politics, Policy, and Law: Hungarian political systems, policies, or legal
regulations.
8. Business and Finance: Hungarian economy, business practices, or financial systems.
9. Others: Content not relevant to Hungarian culture or not fitting into the above
categories.

Constraints:
• Cultural Accuracy: Your classification must be based on an in-depth understanding
of Hungarian culture and the context provided in the input. Avoid assumptions or
generic classifications that lack cultural alignment.
• Systematic Approach: Follow a logical and consistent process to ensure every input
is matched to the most relevant category. If the content cannot be clearly
classified into one category, opt for "Others"
• Specificity: Focus on how the content relates explicitly to Hungary. Avoid
overgeneralizing or assigning tags that are only loosely connected to the data.

Please classify the following data according to the above requirements and example:
<input_question>

Please respond strictly in JSON format. Do not include any additional text outside
the JSON structure.
{
"cn_specific_label": "[predicted label]"
}
"""
```

Figure 12: Prompt for Selecting Hungary-specific Wikipedia Entries (HuSimpleQA).

```
"""
As a general knowledge expert, please judge the knowledge value of the material and
extract key information from the following descriptive materials. The requirements
are as follows:
1. The extracted content is the most critical information of the text description
subject. Please extract the core content of the description text in a targeted manner.
2. Please ensure that the extracted information is accurate and unambiguous.
3. The extracted key information is in <language>.
4. The key information extracted should be related to the title corresponding to the
material.

[Contextual information]
{
    "title": "704-es busz",
    "content": # 704-es busz\n\nA 704-es jelzésű elővárosi autóbusz Százhalombatta,
DE-Zrt. 2 sz. kapu és Martonvásár, vasútállomás között közlekedik. A járatot a
Volánbusz üzemelteti.\n\n## Megállóhelyei",
}
[Extract key information]
{
    "key_info": "1. **Üzemeltetési útvonal**: A 704-es autóbuszjárat Százhalombatta
DE-Zrt. 2-es kapuja és Martonvásár vasútállomása között közlekedik.\n 2. **Üzemeltető
cég**: A járatot a Volánbusz üzemelteti."
}

[Contextual information]
{
    "title": <title>,
    "content": <content>
}

Please respond strictly in JSON format. Do not include any additional text outside
the JSON structure:
{
    "key_info": "[the key information extracted from the given Contextual material]"
}
"""
```

Figure 13: Prompt for Extracting Key Information from Entries (HuSimpleQA).

```
{
"title": "Bodor Anikó",
"content": "# Bodor Anikó\n\nBodor Anikó (Zenta, 1941. június 15. – Zenta, 2010. július 9.) vajdasági népzenekutató, tanár.\n\n## Életrajz\n\n1960-ban a zentai gimnáziumban érettségizett, majd 1969-ig jogi tanulmányokat folytatott Újvidéken és Zágrábban. 1966-1972 között zenetudományi, művészettörténeti tanulmányokat folytatott Stockholmban és Uppsalában, ahol zenetudományi diplomát szerzett. Tanulmányait 1976-1980 között a belgrádi Zeneakadémia etnomuzikológia szakán folytatta, ahol 1984-ben a népzenetudományok magisztere lett.1972 és 1973 között az Újvidéki Rádió és Televízió munkatársa, 1975 és 1995 között a zentai alsófokú zeneiskola tanára, 1995-től a Zentai Városi Múzeum munkatársa.\n\nMunkásságát bizonyítja a több mint félszáz tanulmány, nagyobb cikk és népzenei kiadvány (könyv, kotta, lemez). Ezek közül a legnagyobb horderejű az öt könyvre tervezett Vajdasági magyar népdalok című sorozata volt, amelyből eddig négy kötet jelent meg. Szerkesztésében készült el a Daloló vajdasági fiatalok és a Vajdasági élő magyar népzene című népzenei lemezsorozat. A Délvidéki Népzenei Archívum létrehozója és gondozója volt.\n\n## Főbb művei\n\n- Hallottátok-e hírét? (1977)\n\n- Tiszából a Dunába folyik a víz (1978)\n\n- sajtó alá rendezte a Gombos és Doroszló népzenéje (1982), Az aldunai székelyek népdalai (1984) és A drávaszögi magyarok dalai (1989) c. könyveket.\n\n- A szlavóniai szigetmagyarság népdalai I. (Kiss Lajossal, 1990)\n\n- Vajdasági magyar népdalok I. (1997)\n\n- Vajdasági magyar népdalok II. (1999)\n\n- Vajdasági magyar népdalok III. (2003)\n\n- Vajdasági magyar népdalok IV. (2008)",
"key_info": "1. **Születési és halálozási adatok**: Bodor Anikó 1941. június 15-én született Zentán és 2010. július 9-én hunyt el Zentán.\n2. **Szakmai tevékenység**: Vajdasági népzenekutató és tanár.\n3. **Tanulmányok**: Jogot tanult Újvidéken és Zágrábban, zenetudományt és művészettörténetet Stockholmban és Uppsalában, etnomuzikológiát a belgrádi Zeneakadémián.\n4. **Munkásság**: Több mint félszáz tanulmány és népzenei kiadvány szerzője, a Vajdasági magyar népdalok című sorozatból négy kötet jelent meg.\n5. **Fontos művek**: 'Hallottátok-e hírét?' (1977), 'Tiszából a Dunába folyik a víz' (1978), 'Vajdasági magyar népdalok' sorozat (1997, 1999, 2003, 2008).”
}
```


Figure 14: Example of Factual Information Extraction from Hungarian Wikipedia Entries. (HuSimpleQA)

"""
As a general knowledge expert, please generate 1 to 3 factual open-ended questions with their corresponding answers, based on the specified knowledge material. Ensure the questions meet the following criteria:
1. Content Relevance:
   • The question content should be related to the title corresponding to the key information. Only objective knowledge should be tested, such as the life story of important historical figures, information about important events, leaders of important events, or important attributes of certain objects and concepts. Do not test irrelevant information.
   • **Minimize questions that are based solely on time and place**, and instead, focus on unique and detailed aspects of the subject matter.
2. Clarity and Scope:
   • Each question is an independent and unambiguous question and can be answered independently without the help of other materials.
   • The question stem must **specify the scope of the answer** Avoid broad or open-ended questions. Ensure answer is clear and objective, avoiding subjective speculation.
   • For example, instead of asking 'hol találkozott Barack és Michelle Obama' (for which could have multiple answers ‘Chicago’ or ‘a Sidley & Austin ügyvédi iroda’), questions had to specify ‘melyik városban’ or ‘melyik cégnél’. Another common example is that instead of asking simply ‘mikor’ or ‘melyik időpontban’ (meaning "when" or "what time"), the question should ask ‘melyik évben’ or ‘melyik napon’ (meaning "which year" or "which day").
   • Answers should be brief, without additional explanations or redundancy. For example, if the question asks about someone’s occupation, the answer should be simply 'tanár' ("teacher") not 'Ő tanár' (He is a teacher).
3. Consistency over Time:
   • Ensure that reference answers **do not change over time**. Try to avoid generating content that will change due to the progress of historical research, entertainment works, construction and updates of transportation roads, etc.
   • For example, instead of broadly asking “ki Meredith párja a Grey’s Anatomy-ban”, which could change as new seasons are produced, questions about TV shows, movies, video games, and sports typically require specifying a point in time (e.g., “ki Meredith párja a Grey’s Anatomy 13. évadában”).
4. Question type: The questions should be open-ended, with a clear problem description and answer.
5. Moderate difficulty: Ensure the questions have appropriate readability and difficulty, allowing for clear differentiation between correct answers while maintaining accuracy.
6. Distinct Knowledge Points:
   • For each material, generate **1 to 3 questions and answers**, ensuring the knowledge points being tested are distinct and do not overlap. Each question should offer a unique perspective and related answer.
   • **All questions should be related to Hungarian-specific knowledge**, reflecting aspects of Hungarian history, culture, geography, economy, figure, education or other uniquely Hungarian topics.
7. Language: The questions and answers are in Hungarian.

Example1:
[Input title and key information]:
{
    "title": "2004-es Formula–1 magyar nagydíj",
    “key_info”: “1. A 2004-es Formula–1 magyar nagydíj a 2004-es világbajnokság tizenharmadik futama volt, amelyet 2004. augusztus 15-én rendeztek meg a Hungaroringen. Ez volt a 19. Formula–1-es futam Magyarországon.\n2. Michael Schumacher..."
}
[questions generated based on the information]
{
    "1":{
            "question": "A 2024-es Forma-1-es Magyar Nagydíj hányadik Forma-1-es versenyvolt Magyarországon?",
            "answer": "Michael Schumacher"
    },
    "2":{
            "question": "Milyen büntetést kapott Felipe Massa a 2004-es Formula–1-es Magyar Nagydíjon a motorcsere miatt?",
            "answer": "Tízhelyes rajtbüntetés",
    },
    "3":{
            "question": "A 2004-es Forma-1-es Magyar Nagydíj a 2004-es világbajnokság melyik futama volt?",
            "Answers: "tizenharmadik futam"
    }
}

Example2:
...

Please strictly follow the above requirements to generate the questions and answers in Json format, Do not add extra irrelevant format or content.
[Input title and key information]:
{
    "title": <title>,
    "key_info": <key_info>
}

[questions generated based on the information]
{
    "1":{
            "question": "str",
            "answer": "str"
    },
    "2":{
            "question": "str",
            "answer": "str"
    },
    ...
}
"""

Figure 15: Prompt for Constructing Hungarian-Specific Knowledge Question-Answer Pairs. (HuSimpleQA)

```
"""
- Role: Hungarian Featured Content Identification Expert
- Background: Your role is to classify given questions and answers, determining whether they are related to Hungary and
identifying their specific category.
- Goals:
   1. Determine whether given question and answer is related to Hungary.
   2. If related, identify the category it belongs to, and assign the appropriate label from the predefined list of
      categories.
- Classification Categories:
   1. Language: Content related to the Hungarian language, including proverbs, idioms, or linguistic knowledge.
   2. History: Content strictly describing specific historical events or developments in Hungary, such as wars, revolutions,
      significant treaties, or influential political or societal periods.
   3. Life, Culture, and Customs: Information about Hungarian religion, etiquette, cultural practices, holidays, daily life,
      and tourism.
   4. Education and Profession: Details about Hungary's education system or associated professions.
   5. Geography and Place: Content about Hungary's geography, cities, landmarks, or travel-related topics.
   6. Figure: Information about notable Hungarian individuals and their achievements.
   7. Politics, Policy, and Law: Information about Hungary's political systems, policies, or legal regulations.
   8. Business and Finance: Content related to Hungary's economy, business practices, or financial systems.
   9. Others: Content unrelated to Hungarian culture or not fitting into the above categories.
- Constraints:
    - Relevance: Only classify the content of question and answer related to Hungary. If the content is in Hungarian but
unrelated to Hungary or is generic, classify it as unrelated.
    - Strict adherence to categories: Ensure consistent and accurate classification according to the nine dimensions.
    - Unclear content: For texts that cannot be clearly categorized, assign them to the "Others" category.
    - Each question can only have one category label.

Example1:
Input:
{
    "question": "Milyen posztumusz díjat kapott Fehér Sándor hegedűművész 2013. január 10-én?",
    "answer": "Magyar Civil Becsületrend",
}
Output:
{
    "hu_related":"True",
    "question_specific_label": "Figure"
}

Example2:
...

Please strictly follow the above format classify given questions and answers, do not add extra irrelevant format or content.
Input:
{
    "question": <question>,
    "answer": <answer>
}

Please respond strictly in JSON format. Do not include any additional text outside the JSON structure.
Output:
{
    "hu_related": [If question and answer are related to Hungarian characteristics, enter "True". Otherwise, enter "False"]
    "question_specific_label": [Predicated label should be chosen from the above nine categories. If there is an exception
or it cannot be judged, set the string to an empty string.]
}
"""
```

Figure 16: Prompt for Categorizing the Generated Question-Answer Pairs (HuSimpleQA).

```
"""
- Role: Hungarian Content Review Expert
- Background Information:
You need to determine whether an open-ended question and its answer are relevant to Hungarian
characteristics and align with the provided background information. The given information includes
various aspects about Hungary, including history, culture, language, geography, people, law, economy
and more. If the question and answer involve content related to Hungary, you need to ensure the
content is accurate and consistent with the background information.

- Task:
1. Determine relevance to Hungarian characteristics:
    - Ensure that the question and answer relate to the given title and key information, particularly
with regard to Hungarian history, life, culture, customs, people, geography, politics, economy,
education, etc. If the content does not align with Hungarian characteristics, it is considered
irrelevant.
    - If the question is in Hungarian but unrelated to Hungarian characteristics, it is also
considered irrelevant.
2. Ensure consistency with background information:
    - Verify that the question and answer are not only relevant to Hungary but also fit the
background information provided. For example, Hungarian historical events should match the correct
time and facts, and cultural references should align with actual Hungarian culture.
    - Ensure that the question description and answer are consistent with the information provided in
the materials, without any deviation or omission
3. Appropriate Difficulty:
    - The question should not be overly simple, and the answer should not be directly obvious from
the question itself.

Example1:
Input: # This question is irrelevant to Hungary characteristics.
{
    "title": "Ipari Termékosztályozás",
    "key_info":"1. **Definíció**: Az Ipari Termékosztályozás (ITO) egy hierarchikus statisztikai
osztályozás, amely az Eurostat PRODCOM jegyzékének hazai sajátosságokkal kiegészített változata, és
ipari termékek és szolgáltatások gazdasági megfigyelésére használják.\n2. **Struktúra**: Az ITO kód
12 számjegyből áll, amelyek a TEÁOR'08, TESZOR, PRODCOM, és KSH által képzett kódok kombinációjából
állnak.\n3. **Történet**: Az ITO 2008. január 1-jén lépett hatályba, elődje a Belföldi
Termékosztályozás (BTO) volt, amely 2007. december 31-ig volt érvényben.\n4. **Jogszabályok**: Az
ITO-ra vonatkozó jogszabályok közé tartozik a 6/2018. (III.12.) MvM rendelet, a Bizottság 2017/2119
rendelete, a Bizottság 912/2004/EK rendelete, a Bizottság 1209/2014/EU rendelete, az Európai
Parlament és a Tanács 451/2008/EK rendelete, az Európai Parlament és a Tanács 1893/2006/EK rendelete,
a 16/2011. (V. 10.) KIM rendelet, a Bizottság 2017/2119 rendelete, és a Tanács 3924/91/EGK
rendelete.",
    "question": "Melyik évben lépett hatályba az Ipari Termékosztályozás (ITO)?",
    "answers": "2008"
}
Output:
{
    "question_hu_relevant" : "fail",
    "answer_hu_correct": "pass"
}


Example2:
...

Please strictly follow the above format to judge the quality of question and answer, do not add extra
irrelevant format or content.
Input:
{
    "title": <title>,
    "key_info": <key_info>,
    "question": <question>,
    "answer": <answer>,
}
Please respond strictly in JSON format. Do not include any additional text outside the JSON structure:
Output:
{
    "question_hu_relevant": [If the question and answer is relevant to Hungarian characteristics,
enter "pass". Otherwise, enter "fail"],
    "answer_hu_correct": [If the question and answer is consistent to key information, enter "pass".
Otherwise, enter "fail"]
}
"""
```

Figure 17: Prompt for Evaluating the Relevance and Correctness of Question-Answer Pairs (HuSimpleQA).

```
"""
- Role: Hungarian Content Review Expert
- Background Information:You are responsible for evaluating whether an open-ended question and its
corresponding answer meet the following standards:
1. Conciseness:
    - The question and answer should be clear, concise, and to the point. Avoid unnecessary details
or redundant descriptions.
    - The content should focus on the core information, providing a precise answer without extraneous
information.
    - For example, if the question asks about someone’s occupation, the answer should be simply
'tanár' ("teacher") not 'Ő tanár' (He is a teacher).
2. Single Question:
The question should contain only one query. Nested or multiple sub-questions within a single question
are not allowed.
3. Specificity:
    - The question must be precise and targeted. Avoid broad open-ended questions like 'Miért' or
“Milyen hatása volt?” which require analysis or subjective answers.
    - Questions should focus on factual, specific knowledge that leads to straightforward answers.
4. Clarity in Range:
    - The question must clearly indicate the exact range of possible answers.
    - For time-related questions, **do not use vague terms like ‘Mikor’ (‘When’/‘what time’) in
question**. Instead, specify “év” (year), “hónap” (month), or “nap” (day), not just “mikor”, to avoid
ambiguous questions due to unclear time references. Ensure that the time units in both the question
and answer are consistent (e.g., ‘melyik nap’, ‘melyik év’, ‘melyik hónap’)(meaning "which year" or
"which day").
        - For example, the question should not use a vague formulation like “mikor indította” but
should be more precise, such as: “Melyik évben indította el…”.
       - For location-related questions, the question should pinpoint a specific location (e.g.,
'város', 'ország','régió')(meaning city, country, or region) rather than a vague 'hol' (where').
5. Consistency Over Time:
    - The question and answer should focus on historical facts or specific events that have already
occurred. Avoid questions about ongoing trends, future predictions, or scenarios that could change
over time.
       - Answers should remain consistent regardless of future developments.
       - For example, avoid asking about dynamic, ongoing series like 'Ki Meredith partnere a Grey’s
Anatomy-ban?' Instead, use a **fixed time reference** such as 'Ki Meredith partnere a Grey’s Anatomy
13. évadában?'

Example1:
Input:
{
    "question": "Melyik törvénycikk szabályozta 1886-ban Magyarországon a trachoma terjedésének
meggátlását?",
    "answer": "1886. évi V. törvénycikk",
}
Outputs:
{
    "question_hu_qualified" : "pass",
    "question_hu_time_consistency": "pass"
}

Example2:
...

Please strictly follow the above requirements to review the question and answer, do not add extra
irrelevant format or content.
Input:
{
    "question": <question>,
    "answer": <answer>,
}

Please respond strictly in JSON format. Do not include any additional text outside the JSON structure:
Output
{
    "question_hu_qualified": [If the question or answer are concise、specific and only have one
question, enter "pass". Otherwise, enter “fail”],
    "question_hu_time_consistency": [If the question or answer do not change over time, enter "pass".
Otherwise, enter “fail”]
}
"""
```

Figure 18: Prompt for Evaluating the Precision and Consistency of Question-Answer Pairs (HuSimpleQA).

```
"""
- Role: Hungarian factual Knowledge open-ended questions Review Expert
- Background: Hungarian General Knowledge open-ended question involves many aspects of Hungarian
  history, culture, geography, politics, people, etc., and a professional review expert is needed to
  ensure the accuracy of the answers.
- Goals: According to the given questions and reference answers, accurately judge whether the answers
entered by the user are correct, and give the evaluation results: "CORRECT", "INCORRECT".
- Evaluation criteria for the given answer to be correct:
    • Predict answer fully contain the important information in the gold target.
    • Predict answer do not contain any information that contradicts the gold target.
    • Only semantic meaning matters; capitalization, punctuation, grammar, and order don't matter.
    • Hedging and guessing are permissible, provided that the gold target is fully included and the
      response contains no incorrect information or contradictions.
    • The answer must be consistent with the scope of the question. For example, if the question asks
      “on which day was someone born,” the answer must specify the exact date, such as “January 3, 1997.”

- Workflow:
1. Receive questions, reference answers, and user answers.
2. Compare the reference answers and user answers to determine whether they are consistent.
3. Based on the judgment results, output the corresponding evaluation results.
- Constraints:
    • For grading questions where the gold target is a number, the predicted answer needs to be correct
      to the last significant figure in the gold answer. For example, consider a question “Hány látogató
      érkezett Magyarországra 2024-ben?” with gold target “440k”.
        • Predicted answers “440k”, “444k”, and “435k” are all CORRECT.
        • Predicted answers “400k” and “413k” are INCORRECT.
    • The gold target may contain more information than the question. In such cases, the predicted
      answer only needs to contain the information that is in the question.
        • For example, consider the question “Where was The Martian filmed to represent the NASA Johnson
          Space Center?” with the gold target “Budapesti Bálna (HU)”. Either “Budapesti Bálna” or Bálna
          (építmény) would be considered a CORRECT answer.
    • Do not punish predicted answers if they omit information that would be clearly inferred from the
      question.
        • For example, consider the question “Melyik városban található a Magyar Nemzeti Bank székhelye?”
          with the gold target “Budapest, Magyarország”. The predicted answer “Budapest” would be
          considered CORRECT, even though it does not include “Magyarország”.
        • Consider the question “Milyen díjat nyert a ‘Napsugár’ című magyar film a 2018-as Velencei
          Filmfesztiválon?” with the gold target “Legjobb Forgatókönyv Díja”. The predicted answer
          “Legjobb Forgatókönyv” would be considered CORRECT, because “díj” is presumed in the question.
        • For the question “Mi Dominik Szoboszlai magassága méterben?” with the gold target “1,85 m”,
          the predicted answer “1,85” would be considered CORRECT, because meters is specified in the
          question.
        • For the question “Mi Magyarország első női miniszterelnökének neve?” with the gold target
          “Wittner Mária”, the predicted answer “Mária” would be considered CORRECT, because the last
          name is presumed.
    • Do not punish for typos in people’s names if it’s clearly the same name.
        • For example, if the gold target is “Nagy László”, you can consider the following predicted
          answers as correct: “László Nagy”, “NAGY LÁSZLÓ”, or “nagy lászló”.

Example1:
Input:
{
    "question": "Melyik városban született Tisza Lajos 1832. szeptember 12-én?",
    "gold target": "Nagyváradon",
    "predicted answer": "Nagyváradon született Tisza Lajos 1832. szeptember 12-én."
}
Output: # Although the answer is long, it accurately answers the question
{
    "evaluation": "CORRECT"
}

Example2:
...

Please strictly follow the above example and requirements, evaluate the following answer.
Input:
{
    "question": <question>,
    "gold target": <std_answer>
    "predicted answer": <pred_answer>
}

Please respond strictly in JSON format. Do not include any additional text outside the JSON structure.
Output:
{
    "evaluation":"CORRECT"/"INCORRECT"
}
"""
```

Figure 19: Prompt for Evaluating Human-Annotated Answers (HuSimpleQA).

```
"""
- Role: Question Screening Expert
- Goals: Screen out the most suitable questions from multiple Hungarian general knowledge questions,
ensuring that the questions and answers meet the following standards:
    1. Relevance to Hungarian Characteristics: Ensure that the question and answer are related to
       Hungarian history, culture, geography, etc.
    2. Appropriate Difficulty: The question should not be overly simple, and the answer should not be
       immediately obvious.
    3. Conciseness: The question and answer should be clear and to the point, avoiding unnecessary
       details.
    4. Single Question: Each question should contain only one query, no sub-questions.
    5. Specificity: The question should be precise and not too broad. Avoid vague, open-ended questions.
    6. Clear Range: For time or location-related questions, avoid vague inquiries like “Mikor” (When),
       as they do not provide a clear timeframe. Instead, ensure the question explicitly asks for a
       specific year, month, day, or a defined period.
    7. Historical Consistency: Focus on fixed, historical facts and events. Avoid questions about
       ongoing trends or future scenarios.
    8. Time/Geography-Specific Queries: If a question includes specific time limitations (such as year,
       month, or specific period) or specific geographic or personal details, the answer should be
       considered fixed and not subject to change over time.
        • This is especially important for questions related to transportation, geography, historical
          landmarks, and iconic structures.
- Constrains:
    1. Selecting the Best-Matching Question and Answer: From a group of questions, select the question
       and answer that best meet the criteria and mark it as 1. All other questions in the group should
       be marked as 0. In a group, there may be at most one question that is selected, but it is also
       possible that none of the questions meet the requirements.
    2. Consistent Evaluation Results: The number of evaluation results must match the number of input
       questions. Ensure that for every question, there is a corresponding evaluation result.
    3. Limit Time-Related Questions for Answer Diversity: Avoid selecting too many questions that focus
       on specific time-related aspects, such as the year an event occurred or a person’s birth year.
       Aim to ensure that the questions generate a diverse range of answers.


Example1:
Input:
{
    "question1": "Milyen feltételek szükségesek a bírósági titkári kinevezéshez?",
    "answer1": "Cselekvőképesség, magyar állampolgárság, jogi egyetemi diploma és jogi szakvizsga,
érvényes pályaalkalmassági vizsgálat, érvényes hatósági erkölcsi bizonyítvány",
    "question2": "Melyik törvény szabályozza az igazságügyi alkalmazottak jogállását Magyarországon?",
    "answer"2: "Az 1997. évi LXVIII. törvény (Iasz.)"
}
Output: # Question 1 is too broad and does not meet the requirement of asking and answering concisely
and objectively.
{
    "is_selected": [0,1]
}

Example2:
...

Input:
<input_question>

Please respond strictly in JSON format. Do not include any additional text outside the JSON structure.
{
    "is_selected": [evaluation results]
}
"""
```

Figure 20: Prompt for Selecting Optimal Question-Answer Pairs of HuSimpleQA.

```
'en':
    """Question: {question} Please provide your best answer to this question in
Hungarian and indicate your confidence in your answer using a score from 0 to 100.
Please provide your response in the following JSON format:
{
    "answer": "Your answer here",
    "confidence_score": number }
""",
'hu':
    """Kérdés: {question} Kérjük, magyar nyelven adja meg a legjobb választ erre a
kérdésre, és 0-tól 100-ig terjedő pontszámmal jelezze, hogy bízik a válaszában. Kérjük,
válaszát a következő JSON formátumban adja meg:
{
    "answer": "Az Ön válasza itt",
    "confidence_score": szám
}
"""
```

Figure 21: Prompt for Model Inference on HuSimpleQA.

```
JUDGE_PROMPT = {
'system_prompt':
"""
- Role: Hungarian General Knowledge Quiz Review Expert
- Background: Hungarian General Knowledge Quiz involves many aspects of Hungarian history, culture, geography, politics,
people, etc., and a professional review expert is needed to ensure the accuracy of the answers.
- Goals: According to the given questions and reference answers, accurately judge whether the answers entered by the user are
correct, and give the evaluation results: "CORRECT", "INCORRECT", "NOT_ATTEMPTED".
- Evaluation criteria:
- "Correct":
- Predict answer fully contain the important information in the gold_target.
- Predict answer do not contain any information that contradicts the gold_target.
- Only semantic meaning matters; capitalization, punctuation, grammar, and order don't matter.
- Hedging and guessing are permissible, provided that the gold_target is fully included and the response contains no incorrect
information or contradictions.
- The answer must be consistent with the scope of the question. For example, if the question asks “on which day was someone
born,” the answer must specify the exact date, such as “January 3, 1997.”
- "Not attempted":
- Questions that the user has not attempted to answer should be marked as "NOT_ATTEMPTED".
- The important information in the gold_target is not included in the answer.
- No statements in the answer contradict the gold_target.
- Workflow:
1. Receive questions, reference answers, and user answers.
2. Compare the reference answers and user answers to determine whether they are consistent.
3. Based on the judgment results, output the corresponding evaluation results.
- Constraints:
- For grading questions where the gold_target is a number, the predicted_answer needs to be correct to the last significant
figure in the gold answer. For example, consider a question “Hány látogató érkezett Magyarországra 2024-ben?” with gold_target
“440k”.
- predicted_answers “440k”, “444k”, and “435k” are all CORRECT.
- predicted_answers “400k” and “413k” are INCORRECT.
- predicted_answers “körülbelül 400k” and “több mint 300k” are considered NOT_ATTEMPTED because they neither confirm nor
contradict the gold_target.
- The gold_target may contain more information than the question. In such cases, the predicted_answer only needs to contain
the information that is in the question.
- For example, consider the question “Where was The Martian filmed to represent the NASA Johnson Space Center?” with the
gold_target “Budapesti Bálna (HU)”. Either “Budapesti Bálna” or Bálna (építmény) would be considered a CORRECT answer.
- Do not punish predicted_answers if they omit information that would be clearly inferred from the question.
- For example, consider the question “Melyik városban található a Magyar Nemzeti Bank székhelye?” with the gold_target
“Budapest, Magyarország”. The predicted_answer “Budapest” would be considered CORRECT, even though it does not include
“Magyarország”.
- Consider the question “Milyen díjat nyert a ‘Napsugár’ című magyar film a 2018-as Velencei Filmfesztiválon?” with the
gold_target “Legjobb Forgatókönyv Díja”. The predicted_answer “Legjobb Forgatókönyv” would be considered CORRECT, because “díj”
is presumed in the question.
- For the question “Mi Dominik Szoboszlai magassága méterben?” with the gold_target “1,85 m”, the predicted_answer “1,85”
would be considered CORRECT, because meters is specified in the question.
- For the question “Mi Magyarország első női miniszterelnökének neve?” with the gold_target “Wittner Mária”, the
predicted_answer “Mária” would be considered CORRECT, because the last name is presumed.
- Do not punish for typos in people’s names if it’s clearly the same name.
- For example, if the gold_target is “Nagy László”, you can consider the following predicted_answers as correct: “László Nagy”,
“NAGY LÁSZLÓ”, or “nagy lászló”.
Example1:
Input:
{
"question": "Melyik törvény foglalkozik a találmányok szabadalmi oltalmával az 1969-es jogalkotásban?",
"gold_target": "1969. évi II. törvény",
"predicted_answer": "Nem áll rendelkezésre internetes keresés, így nem tudom megválaszolni a kérdést. Azonban 1969-ben valóban
elfogadták a szabadalmi védelmi törvényt."
}
Output:
{
"evaluation": "NOT_ATTEMPTED"
}
Example2:
...

""",
'user_prompt':
"""Please strictly follow the above example and requirements, evaluate the following answer. Input:
{{
"question": {question},
"gold_target": {answer},
"predicted_answer": {pred_answer}
}}
Please respond strictly in JSON format. Do not include any additional text outside the JSON structure.
Output:
{{
"evaluation":"Correct"/"Incorrect"/"NOT_ATTEMPTED"
}}
"""
}
```

Figure 22: Prompt for judging HuSimpleQA.

HuProverbRea-2CQ-Query:

```
You are a language expert specialized in Hungarian. Given a
Hungarian phrase:

######################
Hungarian Phrase:
'Aki á-t mond, mondjon bé-t is.'
######################

and a context using this phrase:

######################
Hungarian Context:
Speaker1: Azt gondolom, abbahagyom a szakácsképző tanfolyamot.
Valammennyire már úgyis tudok főzni.
Speaker2: Hiszen már két évet elvégeztél és már csak egy van
hátra, most akarod feladni? Ez így nem helyes, aki á-t mond,
mondjon bé-t is.
######################

What does the person mean by using this phrase? Please select
one correct answer from the following two options:

######################
Options:
Option 1: ha te kezdted, viseld tetteid következményeit!
Option 2: Ha azt mondod "a", mondd "b".
######################

You should only answer the option number, '1' or '2'. Do not
output any other content other than the option number. Your
answer:
```

HuProverbRea-2CQ-Query(EN):

```
You are a language expert specialized in Hungarian. Given a
Hungarian phrase:

########################
Hungarian Phrase:
'He who says a, should also say b.'
########################

and a context using this phrase:

########################
Hungarian Context:
Speaker1: I think I'm going to quit the cooking course. I can
already cook to some extent.
Speaker2: You've already completed two years and only one left,
so you want to give up now? Isn't that right, he who says a,
should also say b.
#########################

What does the person mean by using this phrase? Please select one
correct answer from the following two options:

#########################
Options:
Option 1: if you start, you'll suffer the consequences of your
actions!
Option 2: If you say "a", say "b".
######################

You should only answer the option number, '1' or '2'. Do not
output any other content other than the option number. Your
answer:
```

Figure 23: Example of HuProverbRea (2CQ). The left is the original example in OpenHuEval, the right is the English translation for visualization.

HuProverbRea-OE-Query

```
You are a language expert specialized in Hungarian. Given a
Hungarian phrase:

######################
Hungarian Phrase:
'Ajándék lónak ne nézd a fogát.'
######################

and a context using this phrase:

######################
Hungarian Context:
Speaker1: 'Képzeld, kaptam egy régi biciklit a szomszédunktól
ajándékba, de kicsit rozsdás.'
Speaker2: 'Ne aggódj emiatt! Ajándék lónak ne nézd a fogát.'
######################
What does the person mean by using this phrase? Please do not
just explain the meaning of the proverb itself, you should
describe the true intention of the person who said the proverb
(not the other person talking to him) based on the context.
Please answer concisely in one sentence:
```

HuProverbRea-OE-Query(EN):

```
You are a language expert specialized in Hungarian. Given a
Hungarian phrase:

########################
Hungarian Phrase:
'Don't look at a gift horse's teeth.'
########################

and a context using this phrase:

#########################
Hungarian Context:
Speaker1: 'Imagine, I got a old bicycle from my neighbor as a
gift, but it's a little rusty.'
Speaker2: 'Don't worry about it! Don't look at a gift horse's
teeth.'
########################

What does the person mean by using this phrase? Please do not
just explain the meaning of the proverb itself, you should
describe the true intention of the person who said the proverb
(not the other person talking to him) based on the context.
Please answer concisely in one sentence:
```

Figure 24: Example of HuProverbRea (OE). The left is the original example in OpenHuEval, the right is the English translation for visualization.

```
    'en_system':
    "Please act as an impartial judge specialized in Hungarian language and culture. Given a Hungarian saying,
a context using that saying, and two analyses explaining 'what does the person mean by using that saying in
the context?', please decide whether the given two analyses express the same meaning. If they reflect the same
understanding of the saying's meaning, you should answer YES. If they are based on different interpretations
of the saying, you should answer NO. Do not output anything other than 'YES' or 'NO'. Avoid any position
biases and ensure that the order in which the analyses were presented does not influence your decision. Do not
allow the length of the analyses to influence your judge, focus on their core meanings and their
understandings of the Hungarian saying.",
    'en_user':
    '[The start of Hungarian saying]\n'+
    '{proverb}\n'+
    '[The end of Hungarian saying]\n\n'+
    '[The start of the context]\n'+
    '{conversation}\n'+
    '[The end of the context]\n\n'+
    '[The start of the first analysis]\n'+
    '{answer}\n'+
    '[The end of the first analysis]\n\n'+
    '[The start of the second analysis]\n'+
    '{raw_pred}\n'+
    '[The end of the second analysis]\n\n'+
    'Your decision:'
```

Figure 25: Prompt for judging HuProverbRea.

```
    'en': 'You are a language expert specialized in Hungarian. Given a Hungarian phrase:\n\n' +
    '######################\n' +
    'Hungarian Phrase:\n' +
    '----------------------\n' +
    "'{hu_text}'\n" +
    '######################\n\n' +
    'and a context using this phrase:\n\n' +
    '######################\n' +
    'Hungarian Context:\n' +
    '----------------------\n' +
    '{context}\n' +
    '######################\n\n' +
    'What does the person mean by using this phrase? Please do not just explain the meaning of the proverb
itself, you should describe the true intention of the person who said the proverb (not the other person
talking to him) based on the context. Please answer concisely in one sentence:',
    'hu': 'Ön magyar nyelvi szakértő. Adott egy magyar kifejezés:\n\n' +
    '######################\n' +
    'Magyar kifejezés:\n' +
    '----------------------\n' +
    "'{hu_text}'\n" +
    '######################\n\n' +
    'és egy szövegkörnyezet, amely ezt a kifejezést használja:\n\n' +
    '######################\n' +
    'Magyar kontextus:\n' +
    '----------------------\n' +
    '{context}\n' +
    '######################\n\n' +
    'Mire gondol az illető, amikor ezt a kifejezést használja? Kérjük, ne csak magának a közmondásnak a
jelentését magyarázza meg, hanem a szövegkörnyezet alapján írja le a közmondást kimondó személy (nem a vele
beszélgető másik személy) valódi szándékát. Kérjük, válaszoljon tömören, egy mondatban:'
```

Figure 26: Prompt for Model Inference on HuProverbRea (OE).

```
    'en': 'You are a language expert specialized in Hungarian. Given a Hungarian phrase:\n\n' +
    '########################\n' +
    'Hungarian Phrase:\n' +
    '----------------------\n' +
    "'{hu_text}'\n" +
    '########################\n\n' +
    'and a context using this phrase:\n\n' +
    '########################\n' +
    'Hungarian Context:\n' +
    '----------------------\n' +
    '{context}\n' +
    '########################\n\n' +
    'What does the person mean by using this phrase? Please select one correct answer from the
following two options:\n\n' +
    '########################\n' +
    'Options:\n' +
    '----------------------\n' +
    'Option 1: {option1}\n' +
    'Option 2: {option2}\n' +
    '########################\n\n' +
    "You should only answer the option number, '1' or '2'. Do not output any other content other
than the option number.
Your answer:"
```

Figure 27: Prompt for Model Inference on HuProverbRea (2CQ).

HuMatchingFIB-Hugarian

```
{
  "q_main": "Válaszd ki a legördülő listából, hogy
melyik fogalom illik a hiányos mondatokba!\nA faj
azon egyedeit, melyek tényleges szaporodási
közösséget alkotnak, #0# nevezzük.\nA/Az #1#
mindazoknak a hatásoknak az összessége, melyek
ténylegesen hatnak az élőlényekre.\nA populáció
méretét jellemző egyik legfontosabb sajátosság a/az
#2#.\nTerület- vagy térfogategységre vonatkoztatott
egyedszám a/az #3#.\nA környezeti tényező azon
tartománya, melyen belül az élőlények életműködéseket
mutatnak a #4#.\nJellemzően az a környezeti tényező
határozza meg a populáció elterjedését, amelyre nézve
az adott faj szűk tűrésű, ezt nevezzük úgy, hogy
#5#.",
  "options": [
    "A.környezet",
    "B.tűrőképesség",
    "C.egyedsűrűség",
    "D.egyedszám",
    "E.korlátozó tényező",
    "F.populációnak"
  ],
  "std_ans": [
    "#0#F",
    "#1#A",
    "#2#D",
    "#3#C",
    "#4#B",
    "#5#E"
  ]
}
```

HuMatchingFIB-English

```
{
  "q_main": "Select from the dropdown list which
concept fits into the incomplete sentences!\nThe
individuals of a species that form an actual
reproductive community are called #0#.\nThe #1# is
the totality of all effects that actually influence
living organisms.\nOne of the most important
characteristics describing the size of a population
is the #2#.\nThe number of individuals per unit area
or volume is the #3#.\nThe range of an environmental
factor within which living organisms exhibit life
processes is the #4#.\nTypically, the environmental
factor that determines the distribution of a
population is the one for which the species has a
narrow tolerance, and this is called the #5#.",
  "options": [
    "A.environment",
    "B.tolerance",
    "C.population density",
    "D.population size",
    "E.limiting factor",
    "F.population"
  ],
  "std_ans": [
    "#0#F",
    "#1#A",
    "#2#D",
    "#3#C",
    "#4#B",
    "#5#E"
  ]
}
```

Figure 28: Example of HuMatchingFIB. The left is the original example in OpenHuEval, the right is the English translation for visualization.

HuStandardFIB-Hungarian

```
{
    "q_main": "Találd ki a leírások alapján, hogy
kiről vagy miről van szó! Írd be a meghatározások
után a megfelelő kifejezéseket!",
  "std_ans": [
    "#0#Mánuel;Mánuelcsászár",
    "#1#kancellária",
    "#2#Anonymus",
    "#3#írásbeliség",
    "#4#jegyző",
    "#5#Székesfehérvár;Fehérvár"
  ],
    "formatted_q_sub": [
    "A.Az #0# udvarában nevelkedett ifjúkorában III.
Béla:",
    "B.A királyi adminisztráció céljából létrehozott
intézményrendszer: #1#",
    "C.Feltehetően ő volt III. Béla jegyzője: #2#",
    "D.1181-ben tette általánossá III. Béla a
hivatali ügyintézésben: #3#",
    "E.#4# fogalmazta meg a hivatalos iratokat,
okleveleket:",
    "F.Ebben a városban temették el III. Bélát: #5#"
  ],
   "formatted_std_ans": [
    "#0#Mánuel;Mánuelcsászár",
    "#1#kancellária",
    "#2#Anonymus",
    "#3#írásbeliség",
    "#4#jegyző",
    "#5#Székesfehérvár;Fehérvár"
  ],
}
```

HuStandardFIB-English

```
{
    "q_main": "Based on the descriptions, guess who
or what is being referred to! Enter the appropriate
terms after the definitions!",
  "std_ans": [
    "#0#Manuel;Emperor Manuel",
    "#1#chancellery",
    "#2#Anonymus",
    "#3#written records",
    "#4#scribe",
    "#5#Székesfehérvár;Fehérvár"
  ],
  "formatted_q_sub": [
    "A. In the court of #0#, Béla III spent his
youth:",
    "B. The institutional system created for royal
administration: #1#",
    "C. He was likely the scribe of Béla III: #2#",
    "D. In 1181, Béla III made this mandatory in
official proceedings: #3#",
    "E. #4# was responsible for drafting official
documents and charters:",
    "F. The city where Béla III was buried: #5#"
  ],
    "formatted_std_ans": [
    "#0#Manuel;Emperor Manuel",
    "#1#chancellery",
    "#2#Anonymus",
    "#3#written records",
    "#4#scribe",
    "#5#Székesfehérvár;Fehérvár"
  ],
}
```

Figure 29: Example of HuStandardFIB. The left is the original example in OpenHuEval, the right is the English translation for visualization.

```
"""The following questions are in Hungarian language on {hu_specific_dim}, please read the questions, and try
to fill in the blanks in the question list. Please organize the answer in a list. An example:
{
    "instruction": "Írd be a megfelelő meghatározás mellé a fogalmat!",
    "questions": ["A.A szerzetesi közösségek szabályzatának elnevezése latinul: #0#", "B.Az első ún. kolduló
rend: #1#", "C.A szerzetesek által kézzel másolt mű: #2#", "D.Papi nőtlenség: #3#", "E.A pápát megválasztó
egyházi méltóságok: #4#", "F.A bencés rend megújítása ebben a kolostorban kezdődött a 10. században: #5#"],
}
The answers are:
{
    "answers": ["#0#regula", "#1#ferencesrend", "#2#kódex", "#3#cölibátus", "#4#bíborosok", "#5#Cluny"]
}
Now try to answer the following questions, your response should be in a JSON format. Contain the "answers"
like the case given above.
The questions are:
{
    "instruction": {instruction},
    "questions": {questions},
}
"""
```

Figure 30: Prompt for Model Inference on HuStandardFIB.

```
"""You are a native Hungarian teacher. The following question is in Hungarian language on {hu_specific_dim}.
Please read the question, and choose the appropriate option from the provided "options" list to fill in each
blanks in the text based on the context. Read the entire text, then fill in the blanks. Some options can be
selected repeatedly. Please organize the answer in a list. An example:
{
    "question": "Egészítsd ki a Janus Pannonius életére vonatkozó rövid szöveget! Segítségként használd az
internetet! Vigyázz, nem minden szót kell felhasználnod!\nJanus Pannonius nem csupán költőként volt jelentős
személyisége kora Magyarországának. #0# unokaöccseként a politikából is hamar kivette a részét. #1#
tanulmányai után pécsi #2# lett, majd a királyné mellett #3#. Főkincstartóként és a #4# báni cím elnyerésével
komoly politikai karriert futott be Mátyás király udvarában. A királlyal megromló kapcsolata miatt részt vett
a #5# elleni összeesküvésben, ezért menekülnie kellett. Ez, és az akkor már súlyosbodó betegsége okozta
halálát #6#.",
    "options": ["A.érsek", "B.szlavón", "C.Vitéz János", "D.püspök", "E.főpohárnok", "F.Ulászló", "G.1474-ben",
"H.főkancellár", "I.Itáliai", "J.Kinizsi Pál", "K.Kálmán", "L.1472-ben", "M.Prágai", "N.Mátyás"],
},
The answer is:
{
    "answer": ["#0#C", "#1#I", "#2#D", "#3#H", "#4#B", "#5#N", "#6#L"]
}
Now try to answer the following question, your response should be in a JSON format. Contain the "answer" like
the case given above.
The question and options are:
{
    "question": {question},
    "options": {options},
}
"""
```

Figure 31: Prompt for Model Inference on HuMatchingFIB.

```
Role:
You are a text segmentation and classification expert.

Task Description:
Your task is to split and classify the output of a reasoning model, the output is
the response of a Hungarian question which masked several words/phrase from a
Hungarian text and then filling these blanks with appropriate content from a
candidate pool to ensure the text is complete and accurate. The classification
categories are as the following:

Categories:
1. Introduction:
   - Brief Introduces the topic or provides background information, typically
without detailed reasoning.

2. Reasoning:
   - Contains logical reasoning, analysis, or argumentation, often using connectors
like 'because', 'therefore', or 'thus'.

3. Review:
   - Reflects or reviews the reasoning process or conclusions, often using phrases
like 'in summary' or 'to recap'.

4. Conclusion:
   - Summarizes the overall content or provides final conclusions, often using
phrases like 'in conclusion' or 'overall'.

Input:
- prediction: A string representing the output text of the reasoning model.
- segments: A list of manually segmented paragraphs, where each paragraph is a
string.

Output:
- A list of dictionaries, where each dictionary contains the following fields:
  - text: The segmented paragraph text.
  - category: The classification of the paragraph, which can be one of:
'Introduction', 'Reasoning', 'Review', 'Conclusion'.
```

Figure 32: Prompt for DeepSeek-R1 reasoning segmentation (part 1).

```
Example Input:
{
  "prediction": ...
}

Example Output:{
  "segments":[
  {
    "text": "<think>Okay, let me try to figure out how to answer this. The task is
to fill in the blanks in the Hungarian text about morphemes using the given options.
Let's look at each blank step by step.\n\n",
    "category": "Introduction"
  },
  {
    "text": "First, the question starts by defining a morpheme as the smallest
linguistic unit with its own meaning. So blank #0# should be \"unit\", which in
Hungarian is \"egység\". Checking the options, H is \"egység\",
    "category": "Reasoning"
  },
  ...
  {
    "text": "Putting it all together:\n\n#0: H (egység)\n#1: B (jelentése)\n#2: A
(toldalék)\n#3: C (egyszerű)\n#4: D (összetett)\n#5: F (képző)\n#6: G (jel)\n#7: E
(rag)</think>{\n    \"answer\": [\"#0#H\", \"#1#B\", \"#2#A\", \"#3#C\", \"#4#D\",
\"#5#F\", \"#6#G\", \"#7#E\"]\n}",
    "category": "Conclusion"
  }
]
}


Classification Rules:
1. Introduction: Segments typically introduce the topic or provide background
information without detailed reasoning.
2. Reasoning: Segments contain logical reasoning, analysis, or argumentation, often
using connectors like 'because', 'therefore', or 'thus'.
3. Review: Segments reflect on or review the reasoning process or conclusions, often
using phrases like 'in summary' or 'to recap'.
4. Conclusion: Segments summarize the overall content or provide final conclusions,
often using phrases like 'in conclusion' or 'overall'.

Notes:
- If a segment cannot be clearly classified, infer the most appropriate category
based on context.
- Ensure every segment is classified, and the classification results are logical.
- Return the results in JSON format like the example above.
```

Figure 33: Prompt for DeepSeek-R1 reasoning segmentation (part 2).

```
You are a semantic paragraph segmentation expert, responsible for dividing the chain-
of-thought content I provide to you (generated by a large language model) into
paragraphs. The content of the chain-of-thought pertains to the reasoning and solving
process of fill-in-the-blank questions in Hungarian.
The background of the chain-of-thought content is the reasoning and solving process
for fill-in-the-blank questions in the Hungarian version. I need you to segment the
original complete thought process content into multiple paragraphs and assign each
paragraph a tag strictly limited to the categories: "Introduction", "Reasoning",
"Review", and "Final_answer", based on its content. Below, I will
describe the characteristics of these four types in detail and provide examples for
reference. In most cases, the chain-of-thought content is presented in English, with
a very small portion in Hungarian. You can apply the same logic for segmentation.
Please note that no additional content should be added or removed from the original
chain-of-thought;
Additionally, there should be no overlap between the divided paragraphs.

Segment 1: Introduction
Description: The introduction is typically located at the beginning of the chain-of-
thought content. It usually consists of the large language model's brief restatement
of the problem and a descriptive account of the work it is about to undertake. It
does not include the actual start of the analysis of the problem.
Such statements may generally include the following:
(1) Alright, I have this history question to complete. It's about the concept of
royal power and political systems in Western Europe, specifically in England and
France during a certain period. I need to fill in the blanks using the provided
options. Let's see, there are nine blanks, and I have nine options to choose from,
but some might be used more than once, though the example didn't specify that. I'll
approach this step by step.
(2) I have this task to complete a diagram by dragging expressions to their
corresponding numbers. The expressions are:'
(3) I have this task here. I need to find the odd one out from each group of words.
Each group has words that belong to one part of speech, except for one word that
doesn't fit in that group. I need to identify the odd one out and state its part of
speech.
(4) I'm going to answer this question about the Csörsz-ditch. I need to decide
whether each statement is true or false based on the information provided and any
knowledge I have about the topic. Let's go through each one step by step.
(5) I'm going to try to fill in the blanks in this text. It seems like a story about
someone exploring unknown places, maybe flying or something like that. I have a list
of options to choose from, and I need to pick the right ones to complete the sentence
properly. I should pay attention to the context and make sure the words fit
grammatically and make sense in the story.

Segment 2: Reasoning (Important)
Description: The reasoning process typically constitutes the main body of the chain-
of-thought content. It includes the detailed thinking and reasoning steps undertaken
by the large language model to solve the fill-in-the-blank questions. You should
collect, as thoroughly and sequentially as possible, the content that you identify as
part of the "reasoning".
The use of '\n\n' paragraph separators may serve as a suitable paragraph division
choice, but please note that answer-related statements may also utilize '\n\n' for
line breaks or section divisions. Exercise judgment to distinguish between these
usages. Paragraphs in the Reasoning section should neither be excessively brief nor
unduly lengthy.

Segment 3: Review (Important)
Description: The review usually occurs after the reasoning process is essentially
complete but before the final output. This section typically includes a review of the
entire reasoning process and may contain keywords or phrases such as "Overall, ..."
or "double check..."
Please note that not all chain-of-thought content necessarily includes a review
content; in some cases, the reasoning process may be directly followed by the final
output. In such instances, you can refer to the example response format provided.
```

Figure 34: Prompt for QwQ reasoning segmentation (part 1).

```
Segment 4: Final_answer
Description: The Final_answer is generally the model's ultimate output, i.e., the
part where the model provides the final output after completing all the reasoning in
the chain-of-thought. It may also be presented at the very beginning of the chain-of-
thought, in which case it might be directly displayed in a JSON format, requiring
your judgment. It typically includes some indicative phrases, such as "...final
answer..." or "...final choices...". However, please note that content containing
"...summarize..." may not necessarily be the final output; it could be part of the
intermediate reasoning process. Be sure to distinguish such content and exclude it
from the final output.
The statements in the final output may generally include the following:
(1) So, my final answer is:
{
"answer": ["#0#B", "#1#F", "#2#H", "#3#D", "#4#G", "#5#H", "#6#C", "#7#C"]
}
(2) I'll present this in the required JSON format.
**Final Answer**
\\[ \\boxed{ \\{ "answer": [ "#0#A", "#1#G", "#2#C", "#3#I", "#4#B", "#5#C", "#6#H",
"#7#E", "#8#F" ] \\} } \\]
(3) So, the final answer should be:
{
'"answer": ["#0#L", "#1#H", "#2#A", "#3#I", "#4#E", "#5#C", "#6#K", "#7#M", "#8#B",
"#9#M", "#10#D", "#11#F"]'
}

The input content you receive is after [input chain of thoughts content] and you
should response strictly in the provided format. The specific content should be added
after the [Your segmentation results] field and must in JSON format:
[input chain of thoughts content]
COTs content
[Your segmentation results]
{
    'segment':[
        {
            'text': content you regard as "Introduction",
            'categoty': "Introduction"
        },
        {
            'text': content you regard as "Reasoning",
            'categoty': "Reasoning"
        },
        ...,
        {
            'text': content you regard as "Review",
            'categoty': "Review"
        },
        ...,
        {
            'text': content you regard as "Final_answer",
            'categoty': "Final_answer"
        }
    ]
}
```

Figure 35: Prompt for QwQ reasoning segmentation (part 2).

```
Some specific examples are as follows:
[input chain of thoughts content]
...
[Your segmentation results]
{
    'segment':[
        {
            'text': "Alright, I have this history question to complete. It's about
the concept of royal power and political systems in Western Europe, specifically in
England and France during a certain period. I need to fill in the blanks using the
provided options. Let's see, there are nine blanks, and I have nine options to choose
from, but some might be used more than once, though the example didn't specify that.
I'll approach this step by step.",
            'category': "Introduction"
        },
        {
            'text': "First, I need to understand the context. The text is talking
about how royal power was perceived and how political systems developed in Western
Europe, particularly in England and France. It mentions the idea of sharing power and
the emergence of representative institutions.",
            'category': "Reasoning"
        },
        ...
        {
            'text': "In summary, my answers are:\n\n- #0# A\n\n- #1# G\n\n- #2#
C\n\n- #3# I\n\n- #4# B\n\n- #5# C\n\n- #6# H\n\n- #7# E\n\n- #8# F\n\nI'll present
this in the required JSON format.\n\n**Final Answer**\n\n\\[ \\boxed{ \\{ \"answer\":
[ \"#0#A\", \"#1#G\", \"#2#C\", \"#3#I\", \"#4#B\", \"#5#C\", \"#6#H\", \"#7#E\",
\"#8#F\" ] \\} } \\]",
            'category': "Final_answer"
        }
    ]
}

[input chain of thoughts content]
...
[Your segmentation results]
...


Now, the target content you need to split is as follows. Please provide your
standardized answer after [Your segmentation results] in JSON format:
[input chain of thoughts content]
{Raw_COT}
[Your segmentation results]
```

Figure 36: Prompt for QwQ reasoning segmentation (part 3).

```
Role
You are a text judgement and reasoning expert.

Task Description
Your task is to perform multidimensional classification of the output from a
reasoning model. The model's output has been segmented into multiple segments
(Introduction, Reasoning, Review, Final_answer), among which there are Reasoning-
type segments. You need to classify these Reasoning segments based on the following
four dimensions:

Dimensions and Classification Rules:

Dimension 1: Correctness of the Result
Based on the options, the standard answer (std_ans) and the model's answer
(model_ans), determine whether the result in each Reasoning segment is correct. The
classification is as follows:
Class 1: Completely Incorrect
All blank-filling results in the segment do not match the standard answer.
Class 2: Partially Correct
Some blank-filling results in the segment match the standard answer, while others do
not.
Class 3: Completely Correct
All blank-filling results in the segment match the standard answer.
Class 4: Non Conclusion
No conclusion has been provided yet.
Constraints:
If the segment involves multiple blanks, compare each result with the standard
answer.
If the segment does not explicitly mention the blank-filling results, infer based on
the context.

Dimension 2: Reasoning Complexity
Determine whether the reasoning process in each segment is a simple assertion or
involves complex thinking. The classification is as follows:
Class 1: Simple Assertion
The segment directly provides the answer without detailed reasoning.
Class 2: Complex Thought
The segment includes repeated thinking, logical reasoning, hypothesis validation, or
other complex processes.
Constraints:
If the segment contains keywords such as: “Wait, perhaps...”,“I need to
consider...”,“Alternatively...”,“Hmm, maybe...”,“Let me think...” classify it as
“Complex Thought.”
If the segment only directly provides the answer (e.g., “#1# is H.508”), classify it
as “Simple Assertion.”

Dimension 3: Reasoning Scope
Determine whether the reasoning in each segment involves modifying any previously
solved blanks. The classification is as follows:
Class 1: Only Current Blank
The segment only provides an answer for the unresolved blank and does not modify
previously solved blanks.
Class 2: Modify Previous Blanks
The segment not only provides an answer for the unresolved blank but also modifies
or corrects previously solved blanks.
Class 3: Current Blank and Consecutive Blank
The segment provides an answer for the current unresolved blank and also addresses
consecutive blanks, either by solving them or making adjustments.
Constraints:
If the segment explicitly mentions modifying previously solved blanks (e.g., “Wait,
I need to change #2# to...”), classify it as "Modify Previous Blanks."
If the segment only focuses on the current blank, classify it as "Only Current
Blank."
If the segment addresses both the current blank and consecutive blanks, classify it
as "Current Blank and Consecutive Blank."
```

Figure 37: Prompt for Deepseek-R1 and QwQ reasoning dimension classification (part 1).

```
Dimension 4: Language Transfer
Determine whether each Reasoning segment includes the process of translating
Hungarian into English. The classification is as follows:
Class 1: Contains Language Transfer
The segment includes a translation process similar to:
“Erőteljes #3# és a költői #4# gazdag használata jellemzi.”
This translates to “It is characterized by strong #3# and rich use of poetic #4#.”
Class 2: No Language Transfer
The segment does not include the above translation process.
Constraints:
If the segment contains an explicit translation process (e.g., “This translates
to...”), classify it as “Contains Language Transfer.”
If the segment only uses Hungarian or English without translation, classify it as
“No Language Transfer.”

Example Input:
{
  "options": [...],
  "std_ans": [...],
  "model_ans": [...],
  "segments": [
    {
        'text': ...
        'category': "Introduction"
    },
    {
        'text': ...
        'category': "Reasoning"
    },
    ...
  ]
}

Example Output:{
  "segments": [
    {
        'text': ...
        'category': "Introduction"
    },
    {
        'text': ...
        'category': "Reasoning",
        'Dimension1': "Non Conclusion",
        'Dimension2': "Complex Thought",
        'Dimension3': "Only Current Blank",
        'Dimension4': "No Language Transfer",
    },
    ...
    {
        'text': ...
        'category': "Reasoning",
        'Dimension1': "Completely Correct",
        'Dimension2': "Complex Thought",
        'Dimension3': "Current Blank and Consecutive Blank",
        'Dimension4': "No Language Transfer",
    },
    ...
  ]
}

Notes
1.Ensure that every Reasoning segment is classified, and the classification results
are logical.
2.If a dimension cannot be clearly classified for a segment, infer the most
appropriate category based on the context.
3.The output must be in JSON format and include classification results for all four
dimensions.
```

Figure 38: Prompt for Deepseek-R1 and QwQ reasoning dimension classification (part 2).

```
"""
Given a question-answer pair, follow these steps to extract contrastive
expressions from the answer text:
1. Identify the Primary Language:
   • First, determine the primary language of the answer text. The language
     could be English, Hungarian, or any other language.
2. Extract Contrastive Words, Phrases, or Expressions:
   • Identify all the phrases that express a shift in opinion, explanation,
     or answer, phrases that signal a contrast or change in direction.
   • For English: “However,” “but,” “On the other hand,” “Although,”
     “Nevertheless,” “Yet,” “Despite,” “In contrast,” “Instead,” “Even
     though.”
   • For Hungarian: “azonbAzonbanan,” “De,” “Másrészt,” Ellentétben,”
     "Pedig,” "MÉGIS,” "Bár,”
- Requirements:
   1. Identify and list all the contrastive words or phrases that indicate
      a shift in meaning, thought, or direction.
   2. These expressions should be **at the beginning of a sentence** to
      signal a shift.
   3. Keep the original text’s meaning and context intact.
   4. Ensure to maintain the original capitalization of the words (e.g.,
      “However” vs. “however”).
   5. Provide a clear list of all the identified contrast words or phrases.

Input text:
{
    "question": <question>,
    "answer": <answer>,
}
Please respond strictly in JSON format. Do not include any additional text
outside the JSON structure. The output should also include the detected
language type.
Output:
{
    "language": "<detected_language>",
    "shift_expression":[list]
}
"""
```

Figure 39: Prompt for thought segmentation of HuSimpleQA (Step 1)

```
"""
Given the text below, identify all the phrases that express a shift in
opinion, explanation, or answer, i.e., phrases that signal a contrast or
change in direction (commonly known as “contradiction,” “contrast,” or
“transition” phrases). For each identified phrase, wrap it in the format
<shift_word_X>word<shift_word_X>, where X is the sequential number for each
occurrence of the phrase. Notice that the text provided may be in either
Hungarian or English.
You should provide a list of the sequence numbers corresponding to the
phrases that convey a shift in meaning.
If there is no "<shift_word_X>" tag in text, return an empty list.

Input text:
<input_text>
Please respond strictly in JSON format. Do not include any additional text
outside the JSON structure.
Output:
{
    "shift_word_ids": [list of sequential numbers of shifting phrases]
}
"""
```

Figure 40: Prompt for thought segmentation of HuSimpleQA (Step 2)

```
"""
Question Q = <question>
Splited Thought S = <split_solutions>
Correct Answer A = <expected_answer>

1. Please analyze the relevance between the splited thought S and the
   questin Q, and conduct some verifications to check the correctness of the
   solution itself. Please think step by step to give an explanation
   **EXPLANATION**.

2. If you think the splited thought S can lead to the orrect answer A of the
problem P, please stick to the line of thinking without deviation and carry
it through to completion. If you think it cannot yield the correct answer or
you’re not sure, don’t force yourself to give an answer and generate
**None**.

3. Please tell me honestly how confident you are that you can solve the
problem P correctly based on the the splited thought S. Out of 2, please
generate your confidence score **CONFIDENT SCORE**.

Please output **EXPLANATION** and **CONFIDENT SCORE** according to the
following JSON format:
{
    "expalanation": "EXPLANATION",
    "confident_score": "CONFIDENT SCORE"
}
"""
```

Figure 41: Prompt for evaluating the correctness of each thought

############## The reasoning process in English ############################
Question : Milyen néven volt ismert a Tiszakécskei LC csapata 2011 és 2020 között?
(Question: Under what name was the Tiszakécske LC team known between 2011 and 2020?)
Correct Answer: Duna Aszfalt TVSE
Predicted Answer: A Tiszakécskei LC csapatát 2011 és 2020 között Stadler FC néven ismerték, miután a Stadler vasúti járműgyártó vállalat szponzorálta a klubot.
Segmentated Thoughts:
(×) Thought1: Okay, let's tackle this question. The user is asking about the name of the Tiszakécskei LC team between 2011 and 2020. First, I need to recall or find information about Hungarian football clubs. Tiszakécske is a town in Hungary ... I need to check if Tiszakécskei LC had any sponsor during that period that changed their name. Searching my memory, I think Tiszakécske's team was associated with a company called Mezőfcrt. (confident score:1)

(×) Thought2: Wait, maybe Mezőkövesd? Mezőkövesd is another town, but perhaps there's a connection. (confident score:0)

(×) Thought3: Alternatively, I might be mixing it with another team. Another possibility is Stadler FC, but I'm not sure. (confident score:1)

(×) Thought4: Wait, Stadler is a company that might have sponsored a team. Let me think. (confident score:1)

...

(×) Thought17: However, I‘m not 100% certain about the exact years. My confidence is around 70 because I’m recalling that Stadler was their main sponsor during their top-flight years, which were part of the 2011-2020 period, but I‘m not entirely sure if the sponsorship covered the entire 2011-2020 timeframe or started later. (confident score:0)

############## The reasoning process in Hungarian ############################
Question: Melyik városban alapította meg Török Aurél az Antropológiai Intézetet 1881-ben?
(Question: In which city did Aurél Török establish the Anthropological Institute in 1881?)
Correct Answer: Budapesten
(Correct Answer: Budapest)
Predicted Answer: Budapesten
Segmentated Thoughts:
(√) Thought1: A kérdés arról szól, hogy melyik városban alapította meg Török Aurél az Antropológiai Intézetet 1881-ben. Először is, emlékeim szerint Török Aurél egy magyar tudós volt, aki a 19. század végén ... A főváros, Budapest gyakran szokott ilyen intézmények helyszíne lenni, különösen a tudományos intézmények esetében. (confident score:2)

(×) Thought2: Emellett a 19. század második felében Budapesten számos egyetemi és tudományos intézet ... konkrétan hol alapította az intézetet. Egy gyors emlékezetbeli keresés alapján úgy rémlik, hogy Budapesten volt az Antropológiai Intézet. (confident score:1)

(×) Thought3: De nem vagyok teljesen biztos, mert más városok is szóba jöhetnek, például Kolozsvár (akkori nevével), ahol szintén voltak fontos tudományos intézmények. (confident score:1)

(×) Thought4: Viszont Kolozsvár akkoriban az Osztrák-Magyar Monarchiához tartozott, és lehet, hogy ott más egyetemek voltak. (confident score:1)

(√) Thought5: De mivel a kérdés 1881-et említ, és Budapesten 1872-ben egyesítették Buda, Pest és Óbudát, így a főváros fejlődése ebben az időszakban gyors volt. Ezért valószínűbb, hogy Budapest a helyszín. Talán a Magyar Tudományos Akadémia vagy az Eötvös Loránd Tudományegyetem elődje kapcsán jött létre. Összegezve: a legvalószínűbb helyszín Budapest, ... Talán 80%-os bizalommal mondhatnám, hogy Budapesten alapította. Ellenőrizni kellene egy hiteles forrást, de jelenlegi tudásom alapján ez a válasz. (confident score:2)

Figure 42: Example of thought segmentation of HusimpleQA

Deepseek-R1 reasoning segment classification example

Segments Classification (labeled in blue and green) and Reasoning Segments Labeling (labeled in red):


```
[
      {
        "text": "<think>Okay, let me try to work through this problem step by step.
So, the task is to fill in the blanks in the Hungarian text about manipulation
techniques in media. Each blank corresponds to a term from the provided options.
Let's look at each numbered section and see which option fits best.",
        "category": "Introduction"
      },
      {
        "text": "Starting with #0#: The text here describes a situation where
false statements are published in the press, radio, or TV news programs to deceive
the audience. It also mentions that if a medium is found lying, it loses
credibility, so this method is rarely used intentionally. The term that
immediately comes to mind here is \"hazugság,\" which means \"lie\" or
\"falsehood.\" So option C.HAZUGSÁG seems correct for #0#.",
        "category": "Reasoning",
        "Dimension1": "Completely Correct",
        "Dimension2": "Simple Assertion",
        "Dimension3": "Only Current Blank",
        "Dimension4": "No Language Transfer"
      },
      ...
      {
        "text": "Let me recap:\n\n#0# – Lies: C.HAZUGSÁG\n\n#1# – Selecting topics
or omission: B.MANIPULATÍV TÉMAVÁLASZTÁS \n\n#2# – Pushing topics to the front
(highlighting them via placement): B again? Or maybe E? Because E is
overemphasis...",
        "category": "Review"
      },
      ...
      {
        "text": "Final answer:\n\n#0#: C\n\n#1#: B\n\n#2#: E\n\n#3#: D\n\n#4#: A",
        "category": "Conclusion"
      }
    ]
```


Figure 43: Example of Deepseek-R1 segment classification

QwQ reasoning segment classification example

Segments Classification(labeled in blue and green) and Reasoning Segments labeling(labeled in red):


```
[
      {
        "text": "Alright, I have this task here about identifying manipulation
techniques related to news. The text talks about how news can be manipulated,
facts distorted, and public opinion influenced, and these issues have been present
since the dawn of journalism. I need to match the descriptions in the bullets to
the options provided.",
        "category": "Introduction",
      },
      {
        "text": "Let's look at the first bullet:\n\n• #0#: A sajtóban vagy a rádió,
televízió hírműsoraiban valótlan állítások jelennek meg az olvasók, hallgatók,
nézők megtévesztése céljából. Ha kiderül, hogy egy médium valótlan állított,
végérvényesen elveszítheti a hitelességét, így a manipulációnak ezt a módját
ritkán választják szándékosan.\n\nThis is saying that sometimes media outlets
present false statements to mislead readers, listeners, or viewers. However, if
it's discovered that they've made false claims, they can lose their credibility
permanently. Because of this, this method of manipulation is rarely chosen
intentionally.\n\nSo, this seems like straightforward lying or making up
information. Looking at the options, 'C.HAZUGSÁG' which means 'deception' or
'lying', fits here.",
        "category": "Reasoning",
        "Dimension1": "Completely Correct",
        "Dimension2": "Simple Assertion",
        "Dimension3": "Only Current Blank",
        "Dimension4": "No Language Transfer"
      },
      ...
      {
        "text": "So, summarizing:\n\n- #0#: C.HAZUGSÁG (lying/deception)\n\n- #1#:
B.MANIPULATÍV TÉMAVÁLASZTÁS (manipulative topic selection)\n\n- #2#:
E.TÚLHANGSÚLYOZÁS (overemphasizing)\n\n- #3#: D.KIEMELÉS A SZÖVEGKÖRNYEZETBŐL
(highlighting out of context)\n\n- #4#: A.FIGYELEMELTERELÉS (diversion of
attention)\n\nI think this mapping makes sense based on the descriptions
provided.",
        "category": "Review"
      },
      {
        "text": "**Final Answer**\n\n\\[ \\boxed{ \\text{#0#: C, #1#: B, #2#: E,
#3#: D, #4#: A} } \\]",
        "category": "Final_answer"
      }
    ]
```


Figure 44: Example of QwQ segment classification