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

Anonymous ACL submission

Abstract

As large language models (LLMs) converge towards similar capabilities, the key to advancing their performance lies in identifying and incorporating valuable new information sources. However, evaluating which text collections are worth the substantial investment required for digitization, preprocessing, and integration into LLM systems remains a significant challenge. We present a novel approach to this challenge: an automated pipeline that evaluates the potential information gain from text collections without requiring model training or fine-tuning. Our method generates multiple choice questions (MCQs) from texts and measures an LLM's performance both with and without access to the source material. The performance gap between these conditions serves as a proxy for the collection's information potential. We validate our approach using three strategically selected datasets: EPFL PhD manuscripts (likely containing novel specialized knowledge), Wikipedia articles (presumably part of training data), and a synthetic baseline dataset. Our results demonstrate that this method effectively identifies collections containing valuable novel information, providing a practical tool for prioritizing data acquisition and integration efforts.

1 Introduction

Recent advances in large language models (LLMs) have revealed a striking phenomenon: as these models scale, they tend to develop remarkably similar internal representations and capabilities regardless of their architecture (Huh et al., 2024). This convergence, coupled with established scaling laws (Kaplan et al., 2020) and the growing recognition that AI development should shift from model-centric to data-centric approaches (Zha et al., 2025), suggests that the key to advancing LLM capabilities lies in identifying and incorporating novel, high-quality information sources rather than architectural innovations. However, identifying valuable text collections for model enhancement presents significant challenges: digitization and preprocessing are costly, and training or fine-tuning models on new data requires substantial computational resources.

This creates a critical need: how can we efficiently evaluate whether a text collection contains information that would meaningfully expand an LLM's knowledge? Current approaches typically require actually training or fine-tuning models on new data to assess its value - an expensive and time-consuming process. While retrieval-augmented generation (RAG) (Lewis et al., 2020) offers a promising approach for helping models access long-tail knowledge without full retraining (Kandpal et al., 2023), this still requires careful curation of knowledge bases and comes with its own computational costs. Moreover, the challenge of identifying valuable text collections remains: digitization and preprocessing are costly, and organizations need ways to evaluate the information potential of document collections before investing in their integration.

We present a novel approach to this challenge: an automated pipeline that evaluates the potential information gain from text collections without requiring model training or fine-tuning. Our method generates and leverages multiple choice questions (MCQs) to systematically probe whether the knowledge contained in a text collection is already accessible to an LLM or represents genuinely novel information.

To rigorously validate our approach, we evaluate it across three strategically selected datasets: (1) EPFL PhD manuscripts, containing specialized academic knowledge likely novel to LLMs, (2) Wikipedia articles that were presumably part of the

LLMs' training data, and (3) a synthetic baseline dataset composed of model-generated texts. Our key contributions are:

1. A novel end-to-end pipeline that uses MCQs to efficiently evaluate the potential information gain from text collections without requiring model training or fine-tuning
2. A self-supervised methodology for ensuring MCQ quality through complementary similarity metrics that address two key challenges: verifying question relevance and grounding in source material, while ensuring distractors are plausible yet unambiguously incorrect
3. A systematic analysis of knowledge gaps across diverse text collections that identifies areas of model uncertainty, quantifies the potential value of different information sources, and guides strategic decisions about data collection and integration

Importantly, our approach is dataset-agnostic and can be applied to any text corpus. This makes it applicable to different domains and research questions. This pipeline could be leveraged to select textual data for model enhancements through fine-tuning or retrieval-augmented generation systems.

2 Related Work

Information Gain. The challenge of efficiently selecting new information sources for LLMs can be viewed through the lens of optimal experiment design, a field pioneered by Fedorov et al. (1972). This framework emphasizes maximizing information gain while being strategic about resource allocation – a particularly relevant consideration given the costs associated with integrating new data into LLM systems. Information gain itself has been conceptualized across various fields: in information theory, it relates to reductions in algorithmic information content (Cover et al., 1989); in machine learning, it quantifies a feature's contribution to model performance (Odhiambo Omuya et al., 2021); and in cognitive science, it represents uncertainty reduction in our experience of the world (Damiano et al., 2021). While these theoretical frameworks provide valuable insights, they have nott been previously applied to the specific challenge of evaluating the potential value of text collections for enhancing LLM knowledge. Our work bridges this gap by proposing a practical, MCQ-based approach that quantifies information gain by measuring an LLM's ability to answer questions about a text collection with and without access to the source material.

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

3 Methods

Multiple Choice Questions (MCQs) are a well-established tool for knowledge assessment, supported by research in cognitive science and educational psychology (Thomas M. Haladyna and Rodriguez, 2002). Their four-option format, consisting of one correct answer and three distractors, offers an optimal balance between assessment reliability and cognitive load (Vyas and Supe, 2007). MCQs are particularly valuable for automated evaluation as they provide objective correctness measures while efficiently testing understanding across diverse topics (Oc and Hassen, 2024). When working with Large Language Models (LLMs), MCQs offer an additional advantage: they constrain the output space to a finite set of options, eliminating the ambiguity and variability inherent in open-ended responses and enabling precise evaluation of model knowledge.

Generating high-quality MCQs presents unique challenges, particularly due to the absence of ground truth signals typically available in human-curated educational assessments. Without explicit supervision on what constitutes a good question or appropriate distractors, we must derive reliable metrics to filter and validate the generated MCQs. This self-supervised setting raises two critical chal-

lenges: ensuring questions are both relevant and non-trivial, and guaranteeing that distractors are plausible yet unambiguously incorrect. The latter is particularly crucial when evaluating LLMs – if distractors are too simple, models might succeed through elimination rather than true knowledge; if too similar to the correct answer, they might confuse models even when provided with context. To address these challenges, we propose similarity-based filtering metrics that serve as proxy signals for question quality and difficulty, enabling automated quality control without human supervision.

3.1 Pipeline

As illustrated in Figure 1, our pipeline consists of three key stages: (1) MCQ generation from text using LLMs, (2) quality-focused filtering using similarity metrics, and (3) position-debiased evaluation to assess model knowledge both with and without context.

3.2 MCQ Generation and Filtering

Question Generation Process. The input text is first divided into manageable chunks of 2000 words to ensure consistent context length across questions. These chunks serve as the basis for LLM-generated MCQs, where each question is crafted to test understanding of specific information within the chunk. The exact prompts used for generation can be found in table 2 of the Supplementary Methods.

Two-Stage Quality Filtering. To address the dual challenges of ensuring question relevance and appropriate distractor difficulty without human supervision, we employ two complementary filtering mechanisms:

1. *Context-Answer Alignment Filter:* To ensure questions are both relevant and grounded in the source material, we verify that the correct answer is more strongly aligned with the source context than any distractor:

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

where similarity is measured using both Jaccard index and ROUGE-L score (Lin, 2004). This helps eliminate misleading or incorrect MCQs while ensuring questions test information actually present in the context.

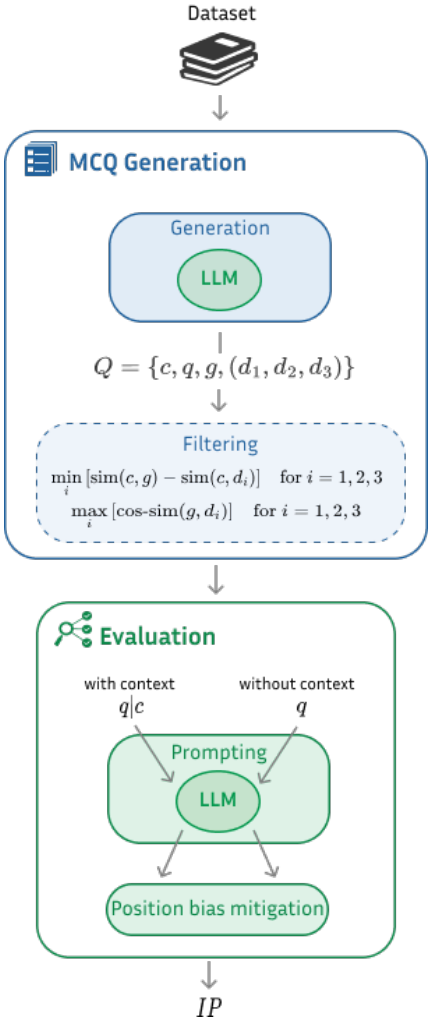


Figure 1: **Overview of our knowledge evaluation pipeline.** The framework consists of three main components: (1) MCQ Generation using LLMs to create questions from input datasets, (2) Quality-focused filtering using similarity metrics, and (3) Position-debiased evaluation to assess model knowledge with and without context. (c: context, q: question, g: ground truth, d_i: distractors)

2. *Distractor Plausibility Filter:* To prevent trivial questions while maintaining unambiguous correctness, we ensure distractors are semantically similar to the correct answer but not identical:

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

using NVIDIA's state-of-the-art NV-Embed-v2 model (de Souza P. Moreira et al., 2024; Lee et al., 2024) for text embeddings. This creates challenging questions where distractors are plausible enough to require true understanding while remaining distinctly incorrect.

3.3 Evaluation

We evaluate the information potential (IP) of a context c for a given LLM f with Eq 3 as the improvement in correctness when incorporating context in the prompt. Formally, it is defined as:

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

where:

$$C_{\text{context}} = \sum_{q,c \in \mathcal{Q}} \mathbb{1}(f(q|c) = g), \quad C_{\text{direct}} = \sum_{q \in \mathcal{Q}} \mathbb{1}(f(q) = g),$$

$$I_{\text{context}} = \sum_{q,c \in \mathcal{Q}} \mathbb{1}(f(q|c) \neq g), \quad I_{\text{direct}} = \sum_{q \in \mathcal{Q}} \mathbb{1}(f(q) \neq g).$$

Here, $\mathbb{1}(\cdot)$ is an indicator function equal to 1 if the condition is true, and 0 otherwise. The denominator ensures that we focus on questions t rather than letting widespread model failures distort the results.

Position Bias Mitigation. To address the known issue of positional bias in LLMs (Shi et al., 2024a), we evaluate each MCQ four times, rotating the correct answer through all possible positions (A, B, C, D) while randomly arranging the distractors. This effect can be seen in Figures 4, 6, and 5. This ensures that model performance reflects true knowledge rather than position-based preferences.

3.4 Data

To evaluate our pipeline's efficacy, we conduct experiments across three strategically selected datasets representing distinct knowledge domains: (1) 177 EPFL PhD manuscripts containing specialized academic research, (2) Wikipedia articles, and (3) synthetic LLM-generated text. This dataset composition enables systematic assessment of knowledge novelty. We hypothesize that PhD manuscripts will yield the highest knowledge gaps, containing recent research contributions likely absent from training data. The synthetic dataset, comprising model-generated content, serves as a lower-bound baseline for knowledge novelty. Wikipedia articles, being a primary source for LLM training (Brown et al., 2020; Touvron et al., 2023), function as a control group representing knowledge presumably well-embedded in the models' parameters.

Data collection. The Wikipedia dataset is created by leveraging the Wikipedia API. With the help of the PhD manuscript titles, related articles are fetched and combined into similar-sized texts. Similarly, the manuscript titles are used to generate the baseline dataset. To overcome the LLMs' generation limits, in a first step, subtopics are generated surrounding the manuscript titles. Then, chunks of around 600 words are generated, using the subtopics as overarching themes, and concatenated.

All data used in our experiments (including chunks, synthetic and baseline datasets) as well as the complete code of our evaluation pipeline are made available open-source with this submission under creative commons license.

4 Results

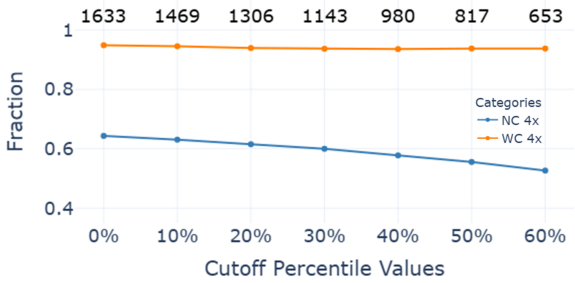
Our analysis reveals three key findings: (1) similarity thresholding effectively increases question difficulty while maintaining answerable questions, (2) PhD manuscripts contain significantly more novel information compared to Wikipedia and synthetic datasets, and (3) larger language models show consistently lower information potential across all datasets, suggesting better knowledge retention during pre-training. We present detailed evidence for each of these findings below.

4.1 Effectiveness of Similarity Thresholding

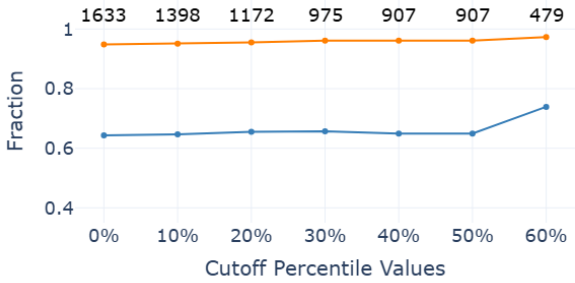
Effect of Cosine Similarity Thresholding. Figure 2a displays the effect of cosine thresholding. It shows a strong decrease of around 10% between the absence of thresholding and the 50th percentile cutoff in the performance of the model with no context (NC 4x), while that of the model with context (WC 4x) remains steady. Therefore, the cosine similarity has the expected effect of increasing the MCQ difficulty when answering without context while maintaining the context-based performance.

Effect of Jaccard and ROUGE-L Thresholding. A different trend is observed when applying the Jaccard and ROUGE-L thresholding, as seen in Figure 2b. Here, a slight increase in performance with context (WC 4x) of 2% can be observed between no thresholding and the 50th percentile cutoff, while performance without context (NC 4x) remains stable.

Comparative Analysis of Thresholding Methods. Our results demonstrate that cosine similarity thresholding is more effective at increasing ques-



(a) Only Thresholding with Cosine Similarity



(b) Only Thresholding with Jaccard and Rouge-L

Figure 2: **Llama 3 70B Performance on EPFL Dataset Across Different Cutoff Percentages.** (with Separated Thresholding and Number of MCQs for the Corresponding Percentiles). **NC 4x:** Evaluation with no context correct for all 4 bias mitigation evaluations of each question; and **WC 4x:** Evaluation with context correct for all 4 bias mitigation evaluations of each question.

tion difficulty than Jaccard and ROUGE-L methods. This difference likely stems from cosine similarity's ability to capture semantic relationships in the embedded space, while Jaccard and ROUGE-L focus on surface-level textual overlap. The gradual reduction in MCQ count under cosine thresholding (from 1633 to 653) suggests a more controlled filtering process compared to the sharp drop-off observed with Jaccard and ROUGE-L (907 to 479 MCQs between 50th and 60th percentiles).

4.2 Information Potential (IP) Across Different Datasets

EPFL PhD Manuscripts (IP: 0.229). The EPFL dataset yields the highest information potential among all tested collections. We observe a significant performance gap between context-free (73.4%) and context-provided (98%) conditions, highlighting the substantial amount of novel information in these documents. The questions generated from this dataset predominantly focus on specialized technical concepts and novel research findings. These performance patterns strongly indicate that PhD manuscripts contain substantial novel information not present in the model's pre-training data.

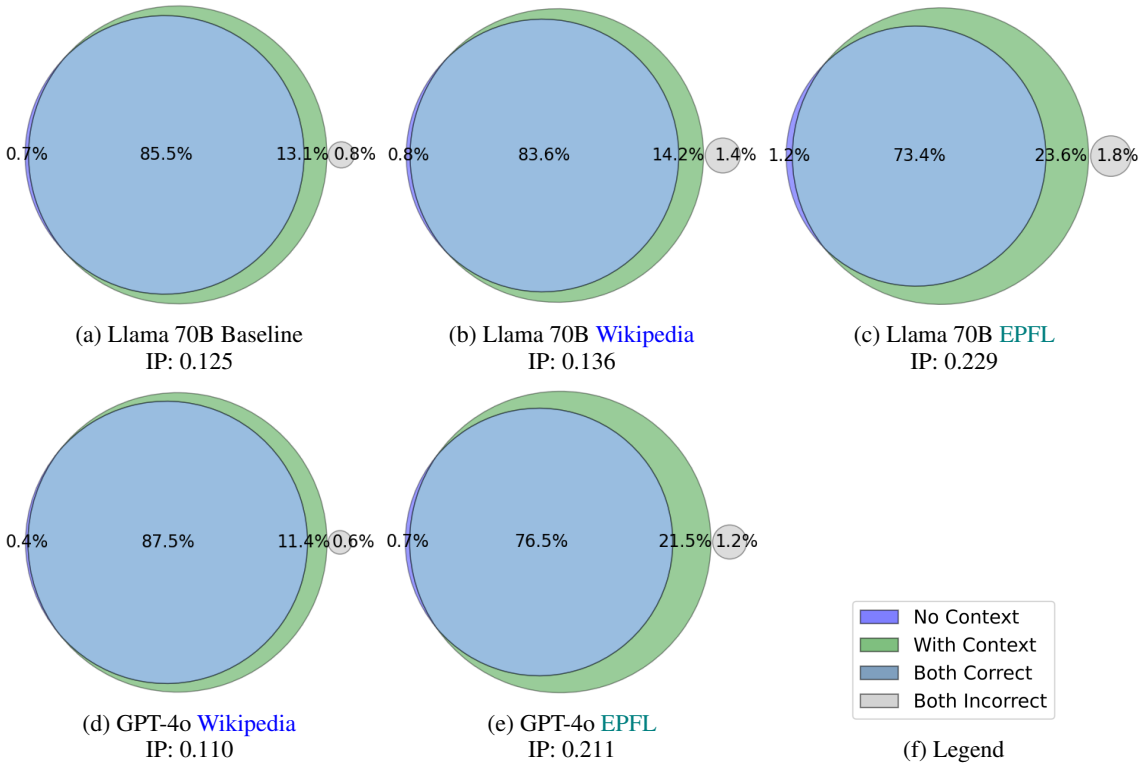
Wikipedia Dataset (IP: 0.136). The Wikipedia dataset demonstrates moderate information potential, with a notably smaller gap between context-free and context-provided performance (83.6% vs. 97.8%). Questions from this dataset primarily cover general knowledge and well-documented facts. The performance patterns suggest partial inclusion of this content in the model's training data, which aligns with our expectations given that these articles were selected from versions predating the model's training cutoff date.

Synthetic Baseline Dataset (IP: 0.125). Among the three datasets, the synthetic baseline yields the lowest information potential. We observe a minimal performance gap between conditions (85.5% vs. 98.8%), with high performance in the context-free setting reflecting the model's inherent familiarity with self-generated content. This dataset effectively serves as a lower bound for the information potential metric, providing a useful reference point for evaluating other collections.

4.3 Comparison between open and closed source model

Despite their architectural and size differences, both Llama 70B and GPT-4o exhibit similar patterns in information potential across datasets, though with notable variations in magnitude. GPT-4o shows slightly lower information potential scores compared to Llama 70B, suggesting better baseline knowledge retention in the larger model. Interestingly, the relative gap between datasets remains consistent across both models - EPFL manuscripts consistently show approximately double the information potential compared to Wikipedia. This consistency across different model architectures and training approaches strengthens the validity of our information potential metric. Additionally, both models maintain near-perfect performance (>97%) when provided with context, regardless of the dataset, indicating robust comprehension capabilities when given access to relevant information.

While a more comprehensive analysis comparing a wider range of LLMs could provide additional



(a) Llama 70B Baseline
IP: 0.125

(b) Llama 70B Wikipedia
IP: 0.136

(c) Llama 70B EPFL
IP: 0.229

(d) GPT-4o Wikipedia
IP: 0.110

(e) GPT-4o EPFL
IP: 0.211

(f) Legend

Figure 3: **Information Potential Analysis Across Datasets and Models.** Venn diagrams showing correct response overlap between context-free and context-provided conditions. Higher Information Potential (IP) scores indicate greater novel information content, with PhD manuscripts (EPFL) showing consistently higher IP (0.211-0.229) compared to Wikipedia (0.110-0.136) and synthetic baseline (0.125). Both open and closed-source models exhibit similar patterns despite architectural differences.

insights into the relationship between model architecture, size, and information potential, such an investigation falls outside the scope of this study, which focuses on establishing the validity of our evaluation methodology.

4.4 Qualitative Analysis of High-Value Information

The MCQs presented in Table 1 exemplify three key patterns in identifying valuable information:

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

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

Table 1: **Examples of technical terminology, novel findings, and complex relationships in EPFL PhD manuscripts.** MCQs requiring context for correct model responses

ing ability becomes blurred when concepts are interconnected in complex ways. The model's performance on such questions suggests that even sophisticated reasoning capabilities cannot compensate for missing foundational knowledge.

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

5 Conclusion and Limitations

5.1 Summary of Contributions

This work introduces a novel and efficient approach for evaluating the information potential of text collections for large language models. Our key contributions include:

1. **Efficient Evaluation Pipeline:** We present a systematic approach combining automated MCQ generation, sophisticated filtering mechanisms, and comparative evaluation to assess information potential without requiring model training or fine-tuning.
2. **Empirical Validation:** Our results validate the method's effectiveness by demonstrating alignment with intuitive expectations across both dataset types and model scales. The information potential increases from synthetic baseline (0.125) to Wikipedia (0.136) to EPFL manuscripts (0.229), while larger models (GPT-4o) consistently show lower information potential than smaller ones (Llama 70B), suggesting better knowledge retention.
3. **Qualitative Framework:** We propose a taxonomy of high-value information types (technical terminology, novel research findings, and complex relationships), providing deeper insights into the nature of valuable information sources for LLMs.

5.2 Limitations and Future Work

Practical Considerations. Our approach, while effective, presents several methodological limitations that suggest directions for future research. The current pipeline's generation-first approach, where questions are created before applying filtering criteria, could be enhanced by incorporating

context-answer alignment and distractor plausibility directly into the generation process. Such an integrated approach could improve question quality while reducing computational overhead.

The scope of our evaluation faces an inherent challenge that manifested both as a limitation of our study and, ironically, in the process of writing this paper: identifying datasets that are definitively outside of LLMs' training data. While PhD manuscripts serve as a reasonable proxy for novel information, the rapid integration of online content into training sets makes finding suitable test collections increasingly difficult. This challenge underscores the timeliness of our work, as it becomes crucial to identify and preserve valuable information sources before they are absorbed into general training data.

Perhaps most significantly, our method deliberately focuses on information selection rather than distillation. This choice, while aligned with our goal of evaluating potential value, leaves open questions about how selected information can be optimally incorporated into LLM training. Recent work has begun to address this gap, with approaches such as SKILL (Moiseev et al., 2022) exploring efficient knowledge integration methods. However, Liu et al. (2024) highlight potential conflicts between newly added information and existing knowledge compressed in LLMs' internal representations, suggesting that information integration remains a complex challenge requiring further investigation.

Broader Considerations. While our approach requires digital text for MCQ generation, it offers a significantly more efficient alternative to full model training or fine-tuning. This enables strategic sampling approaches where representative portions of larger collections can be evaluated to inform broader digitization decisions. For instance, assessing a few chapters can inform decisions about entire book collections, making the method particularly valuable for resource-constrained scenarios. This sampling-based approach could revolutionize how institutions prioritize their digitization efforts, allowing for data-driven decisions about resource allocation in preservation projects.

Our effort to categorize high value information sheds light on fundamental challenges in distinguishing between information that is unknown to an LLM versus information it fails to retrieve. This raises important questions about the relationship between knowledge possession and practical competence in AI systems. When an LLM consistently fails to demonstrate knowledge in a specific domain, the distinction between these cases may become less relevant from a practical perspective. This observation has broader implications for how we conceptualize and evaluate knowledge in AI systems, suggesting that performance-based metrics might be more meaningful than theoretical attempts to map internal knowledge representations. Furthermore, this highlights the need for a more nuanced understanding of how LLMs process and utilize information, particularly when developing strategies for knowledge integration and model enhancement.

References

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Lin Shi, Chiyu Ma, Wenhua Liang, Weicheng Ma, and Soroush Vosoughi. 2024a. Judging the judges: A systematic investigation of position bias in pairwise comparative assessments by llms. *Preprint*, arXiv:2406.07791.

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

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

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

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

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

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

A Appendix

A.1 Supplementary Methods

A.1.1 LLM Prompts

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

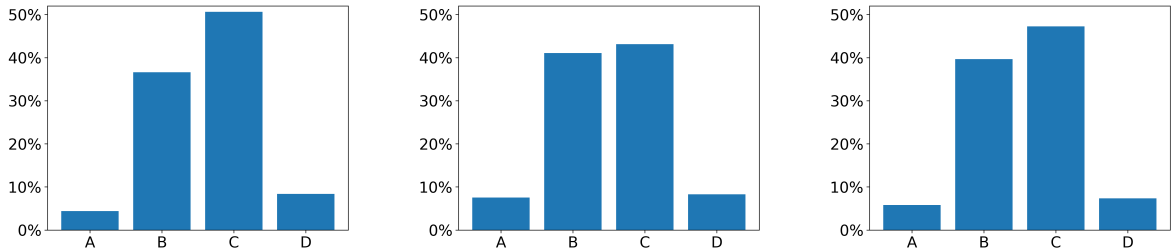
Table 2: LLM Task Prompt Templates

A.2 Supplementary Results

A.2.1 Positional Bias

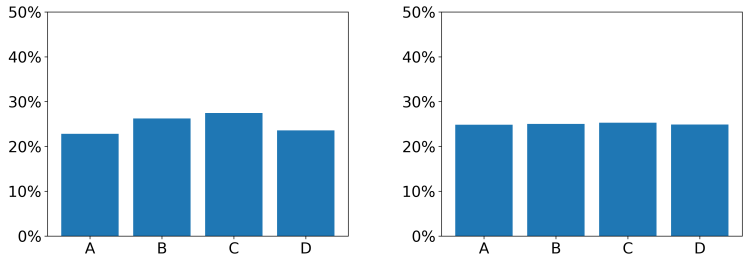
For all three generated MCQ datasets, GPT-4o shows a strong tendency to place the correct answer in the answer options B and C over A and D in around 80% of the time. This tendency may come from training biases where datasets exhibited a similar distribution in MCQ formats.

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

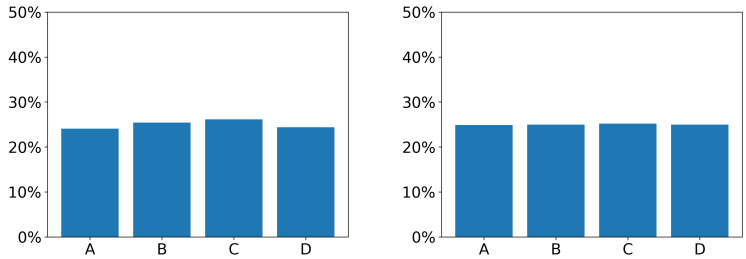


(a) Baseline (b) Wikipedia (c) Venice

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

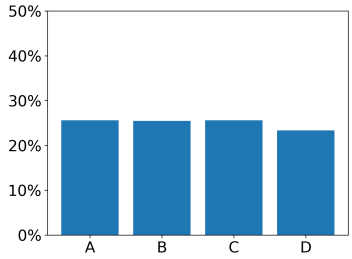


(a) EPFL - No Context (b) EPFL - With Context

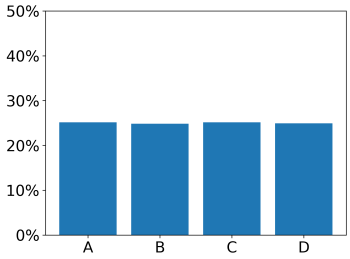


(c) Wikipedia - No Context (d) Wikipedia - With Context

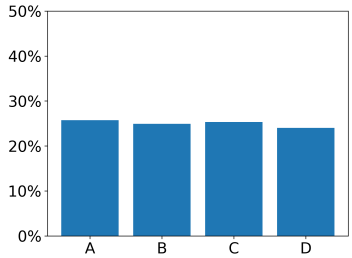
Figure 5: Distribution of Correct Answer Letter Prediction for EPFL and Wikipedia MCQ Datasets Evaluated on GPT-4o After Positional Bias Mitigation.



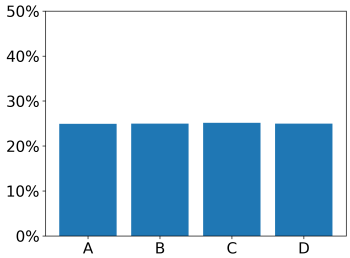
(a) EPFL - No Context



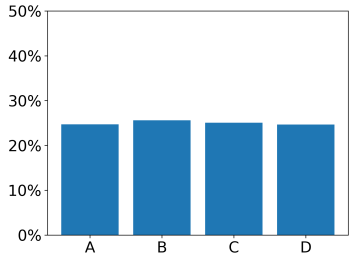
(b) EPFL - With Context



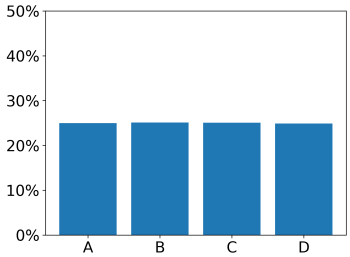
(c) Wikipedia - No Context



(d) Wikipedia - With Context



(e) Baseline - No Context

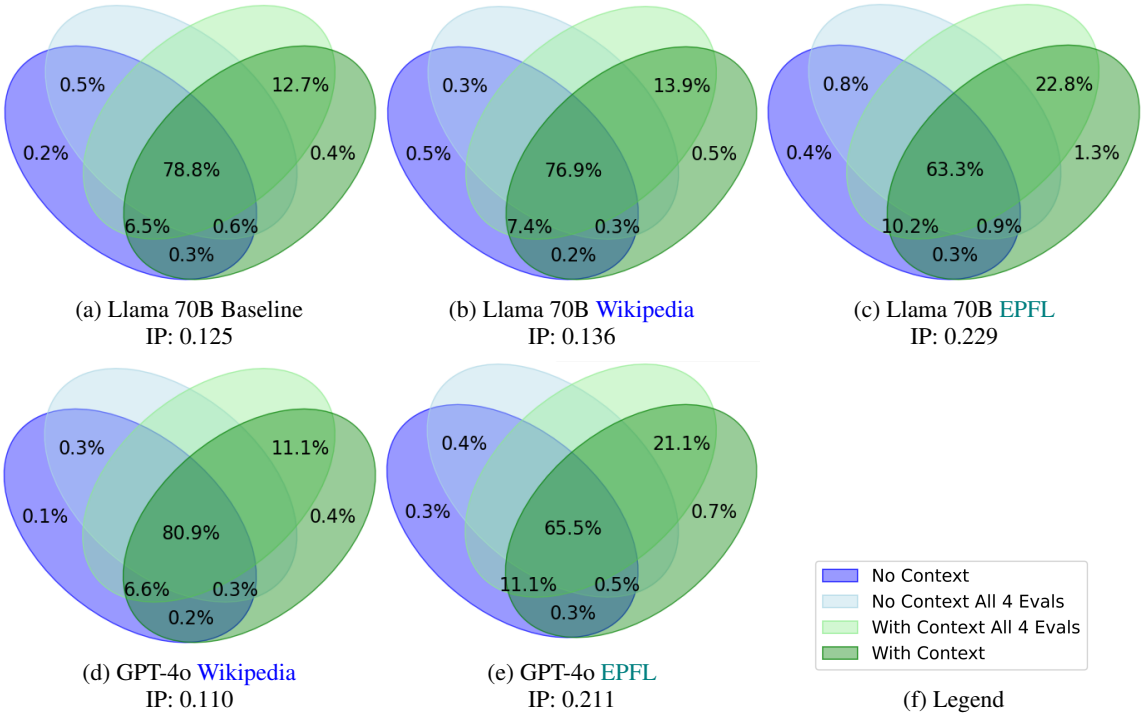


(f) Baseline - With Context

Figure 6: Distribution of Correct Answer Letter Prediction for EPFL, Wikipedia, and Baseline MCQ Datasets Evaluated on Llama 70B After Positional Bias Mitigation.

A.2.2 Information Potential

Figure 7 shows the models' performance with an emphasis on the positional bias mitigation strategy. As every MCQ is evaluated four times, this allows the analysis of a model's consistency. It is clearly visible that the models become less consistent without context in datasets with higher information potential while this trend is less pronounced with context. Figure 8 shows the performance of Llama 70B along with the information potential across the datasets and cutoff percentiles.



(a) Llama 70B Baseline
IP: 0.125

(b) Llama 70B Wikipedia
IP: 0.136

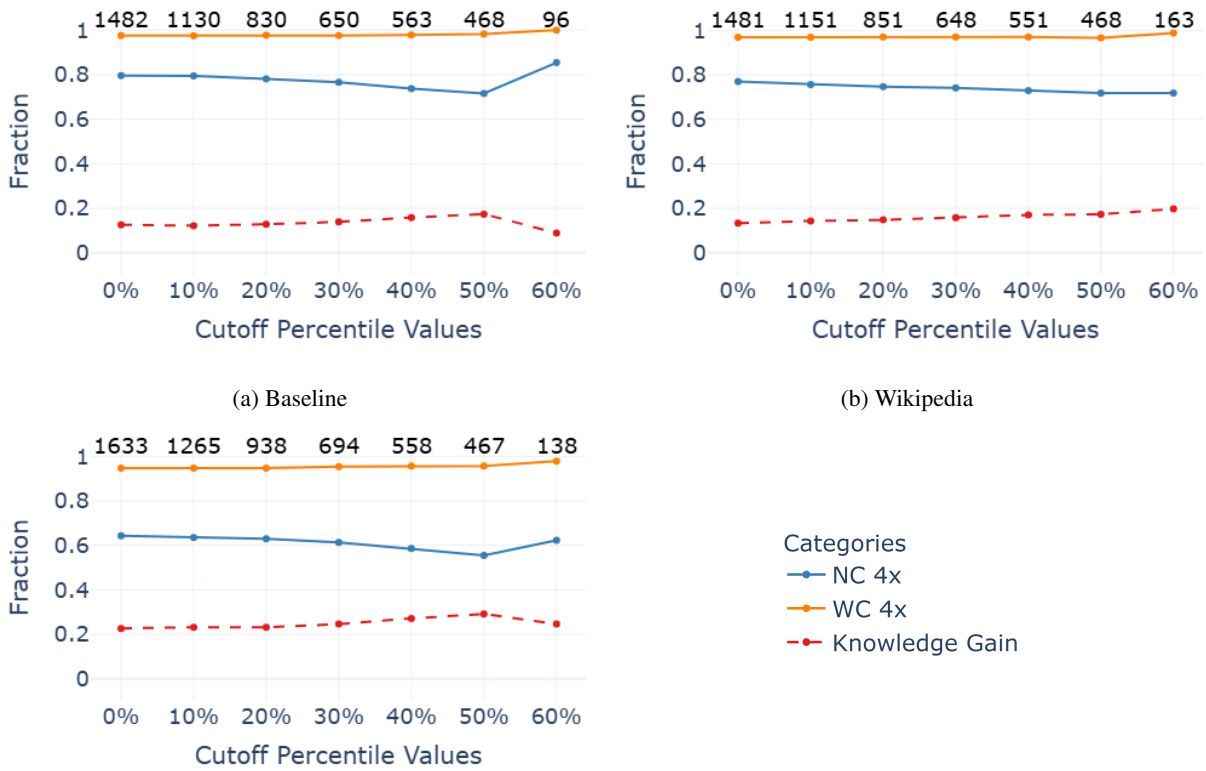
(c) Llama 70B EPFL
IP: 0.229

(d) GPT-4o Wikipedia
IP: 0.110

(e) GPT-4o EPFL
IP: 0.211

(f) Legend

Figure 7: Statistics of Model Performances Including whether Model was Correct on all Four Evaluations of Each MCQ
Venn diagrams of the number of correctly answered questions when answered with or without context.
IP: Information Potential computed with Equation 3



(a) Baseline

(b) Wikipedia

(c) EPFL

(d) Legend

Figure 8: Llama 3 70B Performances Across Different Cutoff Percentiles with Joint Cosine and Jaccard-ROUGE-l Metric Thresholding and Number of MCQs for the Corresponding Percentiles.
NC 4x: Evaluation with no context correct for all 4 bias mitigation evaluations of each question.
WC 4x: Evaluation with context correct for all 4 bias mitigation evaluations of each question.
Information Potential: The value computed with Equation 3

A.2.3 Qualitative Analysis of EPFL MCQ Dataset

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

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