When RAG Fails to Equalize: Geo-bias in Factual Question Answering over Public Companies

Abhinav Havaldar, Enrico Santus
Bloomberg
esantus@bloomberg.net

Abstract

Retrieval-augmented generation (RAG) is widely assumed to mitigate factual errors in large language models (LLMs), but it remains unclear whether retrieval uniformly compensates for missing knowledge. We study this question in a controlled factual QA setting over public companies, constructing a benchmark of ∼2,000 firms across global equity indices. We evaluate six LLMs on four atomic attributes under four conditions: no-context, perfect context, misleading context, and distraction context. We find strong geographic disparities in no-context accuracy, indicating uneven parametric knowledge. While perfect context improves performance, it does not eliminate these gaps: gains are correlated with baseline accuracy, suggesting retrieval effectiveness is coupled to internal representations. Under misleading context, models frequently copy incorrect information. Larger models improve overall performance but do not remove these structural effects. These results challenge the view of RAG as a universal corrective and highlight the interaction between model knowledge, context quality, and entity representation.

1 Introduction

Large language models (LLMs) are increasingly deployed as interfaces for factual question answering in professional workflows. In domains such as finance, corporate intelligence, due diligence, and internal knowledge search, users routinely ask seemingly simple questions—*Where is this company headquartered? When was it founded? Who are its key leaders? What industry does it operate in?* While these queries appear straightforward, they expose two core challenges: parametric knowledge is incomplete and uneven, and retrieval-augmented generation (RAG) does not always behave as a faithful evidence consumption mechanism.

RAG is widely used to improve factuality by supplying external evidence at inference time (Lewis et al., 2020; Guu et al., 2020; Borgeaud et al., 2021). Although it yields strong average gains, it remains unclear when retrieval genuinely compensates for missing knowledge and when it instead reinforces what models already know well.

This question is especially important in domains with uneven factual coverage. Public companies provide a natural example: firms in large, English-dominant markets have rich documentation and strong representation in training data, while companies in smaller or emerging markets often have sparse or fragmented coverage. As a result, identical queries can differ substantially in difficulty depending on the entity involved.

This raises a key issue: retrieval may not act as an independent corrective. If it were, providing perfect evidence should reduce performance gaps across markets. Instead, if retrieval effectiveness depends on existing knowledge—through entity salience or representation quality—it may reinforce existing disparities. In the worst case, imperfect context introduces a second failure mode, where models copy misleading information. Prior work supports this concern, showing a tension between internal priors and external evidence (Wu et al., 2024; Park & Lee, 2024).

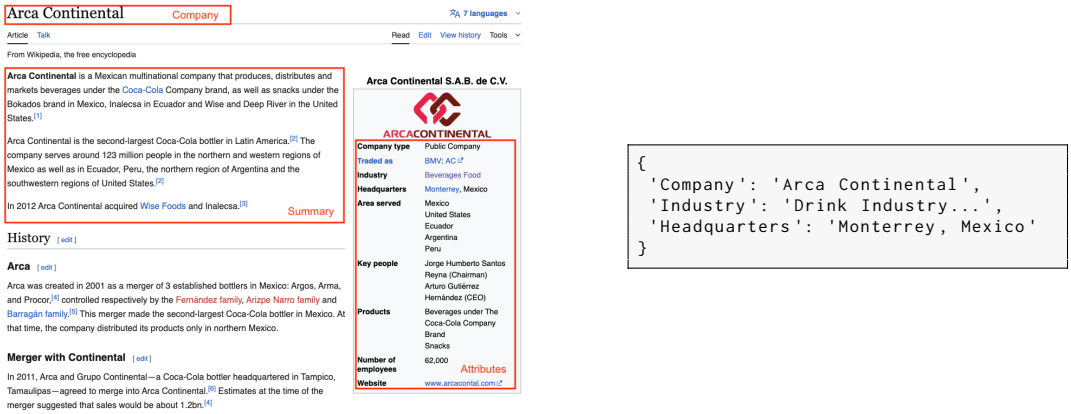


Figure 1: Paired question construction from the same fact.

At the same time, factual recall is known to be uneven across entities and domains (Petroni et al., 2019; Roberts et al., 2020; Kandpal et al., 2022). Recent work further highlights geographic disparities in LLM knowledge (Moayeri et al., 2024; Decoupes et al., 2024). However, existing benchmarks largely focus on aggregate accuracy, leaving open how retrieval interacts with these disparities in realistic settings.

We address this gap by introducing a structured evaluation framework for factual QA over public companies. Using Wikipedia, we construct a benchmark of ∼2,000 companies across 15 global equity indices and evaluate six LLMs on four corporate attributes (Industry, Founding Year, Headquarters, Key People). As shown in Figure 1, each fact is queried in paired inductive and deductive form. To isolate the interaction between parametric knowledge and evidence, we evaluate four conditions: *no-context*, *perfect context*, *misleading context*, and *distraction context*. The benchmark dataset is available at https://zenodo.org/records/19359640.

Our results reveal five key findings. First, no-context accuracy varies substantially across markets, reflecting uneven parametric knowledge. Second, inductive questions are consistently easier than deductive ones. Third, perfect context improves performance but does not eliminate disparities, with gains correlated to baseline knowledge. Fourth, misleading context induces systematic copying errors. Fifth, model scale improves overall performance but does not remove these structural effects.

These findings suggest that retrieval is not a universal corrective. Instead, factual reliability depends on the interaction between model knowledge, context quality, and entity representation. This has direct implications for both evaluation and deployment: beyond average accuracy, systems must be assessed for robustness across markets and sensitivity to imperfect evidence.

Our contributions are fourfold:

1. A benchmark for factual QA over public companies across 15 global equity indices.

2. A controlled framework separating parametric knowledge from contextual effects.

3. Evidence that retrieval gains are coupled with baseline knowledge rather than being uniform.

4. Identification of systematic failure under misleading context.

2 Related Work

A large body of work shows that pretrained language models encode substantial factual knowledge in their parameters. Early probing studies demonstrated that masked language models can recover many relational facts without retrieval (Petroni et al., 2019), while closed-book QA work framed model parameters as a form of compressed factual memory that improves with scale (Roberts et al., 2020). However, this knowledge is uneven: factual recall is brittle under paraphrase, sensitive to prompt variation, and degrades on rare or long-tail entities (Elazar et al., 2021; Kandpal et al., 2022). Popularity bias in factual recall and entity-centric QA predates the LLM era (Févry et al., 2020; Mallen et al., 2022; Lazaridou et al., 2022). These limitations are directly relevant to our setting, as public companies vary widely in salience and textual footprint, making them a natural testbed for structured long-tail knowledge across geography and markets.

Retrieval-augmented generation (RAG) was introduced to address the limits of parametric memory by conditioning models on external documents at inference time (Lewis et al., 2020). Subsequent work integrated retrieval into pretraining and large-scale modeling, while dense retrieval systems and reader architectures established strong baselines for knowledge-intensive QA (Guu et al., 2020; Karpukhin et al., 2020; Izacard & Grave, 2020; Borgeaud et al., 2021). Benchmark suites such as KILT unified evaluation across tasks (Petroni et al., 2020), complementing datasets like Natural Questions and FEVER (Kwiatkowski et al., 2019; Thorne et al., 2018). While this literature demonstrates that retrieval improves average performance, it typically treats retrieval as an exogenous benefit. Our work instead focuses on *differential gain*: whether retrieval helps uniformly across markets or remains conditioned by existing knowledge disparities.

Recent work questions whether models use retrieved context faithfully. Systems may hallucinate, ignore, or over-copy evidence (Gao et al., 2022; Min et al., 2023; Huang et al., 2023). Controlled evaluations show a *tug-of-war* between internal priors and external context: models may follow incorrect evidence when prior knowledge is weak (Wu et al., 2024), and even strong models can be misled by conflicting context (Wu et al., 2024). More broadly, imperfect retrieval and adversarial evidence degrade reliability in retrieval-augmented systems (Park & Lee, 2024). Our framework builds on this line of work but grounds it in a real-world domain and leverages geographic heterogeneity to study how evidence use depends on representational inequality.

Bias and uneven representation in NLP are well documented across social and linguistic dimensions (Blodgett et al., 2020; Hovy & Prabhumoye, 2021). More recent work shows that LLM factual recall varies systematically across countries and income groups (Moayeri et al., 2024), with measurable geographic distortions in model knowledge (Decoupes et al., 2024). We extend this perspective from country-level recall to entity-level factual QA. Public companies provide a structured domain where geography, market prominence, and information availability intersect.

Factual reliability is critical in finance, where LLMs are increasingly used for research and decision support. Existing benchmarks such as FinQA and TAT-QA focus on numerical reasoning over financial reports (Chen et al., 2021; Zhu et al., 2021), while surveys highlight ongoing challenges in factuality and robustness (Li et al., 2023). Our work addresses a complementary problem: atomic factual QA over companies, where the primary challenge lies in uneven coverage and sensitivity to context across global markets.

3 Research Questions and Hypotheses

We study how retrieval interacts with unevenly distributed knowledge by treating global market variation as a natural experiment. We analyze factual QA across indices, question formulations, and context regimes to disentangle parametric knowledge, evidence utilization, and robustness to imperfect context. Our goal is to determine whether retrieval acts as an independent corrective mechanism or remains conditioned by structural biases in model knowledge. The hypotheses are illustrated in Figure 2 and listed below.

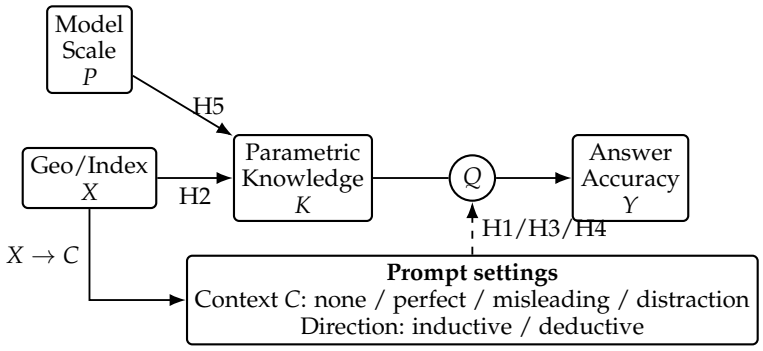


Figure 2: Hypothesis model. Geo/index position and model scale shape baseline parametric knowledge. Prompt settings determine the instantiated question through context regime and direction, thereby moderating how knowledge and evidence translate into answer accuracy.

- **H1 Directional asymmetry:** How does question direction (entity → attribute vs. attribute → entity) affect difficulty and error patterns?
- **H2 Geo-economic disparity without context:** How does factual accuracy in the absence of retrieval context vary across indices and regions?
- **H3 Non-uniform benefit of retrieval:** Does perfect context improve accuracy uniformly, or are gains correlated with baseline performance?
- **H4 Contextual over-reliance:** To what extent does misleading or distraction context induce systematic errors or copying behavior?
- **H5 Scale and evidence use:** Does model scale improve the use of helpful context and robustness to misleading context, particularly in underrepresented markets?

4 Methodology

4.1 Benchmark Construction

Data sources and entity collection. We construct a benchmark of 2,135 unique publicly listed companies sourced from Wikipedia and mapped to 15 major global equity indices spanning North America, Europe, Asia, Latin America, Africa, and Oceania. For each index, we collect constituent companies and associate each entity with its corresponding index, country, and region, enabling geographically stratified analysis.

The resulting benchmark is intentionally heterogeneous in both market size and information coverage, ranging from large English-dominant markets (e.g., S&P 500, ASX 200) to mixed-resource settings (e.g., DAX, TSX, Hang Seng) and smaller or emerging markets (e.g., Ghana, Chile, Mexico, Brazil). A full index-level breakdown is provided in Table 3.

The dataset contains 2,165 total records due to multi-index membership, corresponding to 2,135 unique companies.

Attribute extraction and normalization. For each company, we extract four atomic factual attributes from Wikipedia infobox fields: *headquarters*, *founding year*, *industry*, and *key people*. These attributes are selected for their relevance in corporate analysis and broad availability across entities.

Table 1 summarizes both dataset metadata and the distribution of target attributes. The high cardinality of attributes such as *industry* and *key people*, combined with partial missingness, highlights the open-ended and heterogeneous nature of the benchmark.

To ensure consistency, we normalize all fields: headquarters are converted to a canonical city-country format; founding year is resolved to a single four-digit value; industry labels

Variable	Count	Unique	Freq	Top
Dataset Metadata				
Company	2165	2135	2	–
Index	2165	15	498	S&P 500 (US)
Industry	1703	643	147	Financial Serv.
Founding Year	1896	212	45	1994
Headquarters	1902	666	160	Tokyo, Japan
Key People	1619	1594	2	Peter Konieczny

Table 1: Summary statistics of the dataset. The top panel reports dataset-level metadata, while the bottom panel summarizes the target attributes used for question construction. Note that companies can be listed multiple times in different indices and therefore there are fewer (2135) unique companies than the total (2165).

are standardized using linked page titles and mapped to a controlled taxonomy; and list-valued attributes (e.g., key people) are whitespace-normalized, title-cased, and sorted into canonical order. This normalization is critical to avoid spurious mismatches (e.g., "Paris" vs. "Paris, France") that would otherwise introduce evaluation noise. Detailed preprocessing steps are described in Appendix A.3.

Question construction and distractors. As shown in Figure 1, each fact is converted into two multiple-choice questions that differ only in the direction of inference: **Inductive** (entity $\rightarrow$ attribute) and **Deductive** (attribute $\rightarrow$ entity). This paired design enables controlled analysis of directional asymmetry.

Each question is presented with one correct answer and three distractors. Distractors are selected to be semantically plausible but factually incorrect: company summaries are embedded using `bge-large-en-v1.5`, and for each target we retrieve nearby companies with different attribute values. Depending on the question direction, answer options are instantiated as company names or attribute values. This construction preserves semantic difficulty and reduces the likelihood of solving the task via shallow lexical cues.

Across all companies and attributes, this procedure yields approximately 15,000–17,000 multiple-choice questions, depending on attribute availability.

Context construction and context regimes. For each company, we use the lead paragraph of its Wikipedia page as contextual input. This provides a concise, standardized natural-language summary that contains relevant facts and remains fully reproducible.

To disentangle parametric knowledge from contextual evidence, each question is evaluated under four conditions:

1. **No-context**: model answers using parametric knowledge only.
2. **Perfect context**: the correct company summary is provided.
3. **Misleading context**: the masked Wikipedia summary of a distractor company is substituted, with `[COMPANY]` replaced by the target company name, creating locally coherent but factually incorrect evidence (see Appendix B.1.3).
4. **Distraction context**: the summary includes additional irrelevant information.

These regimes allow us to isolate how models balance internal knowledge with external evidence, and to evaluate their robustness to misleading or irrelevant context.

4.2 Experiment Design

Models and inference setup. We evaluate six language models spanning proprietary and open-weight families: GPT-5, GPT-5 mini, GPT-5 nano, Claude Sonnet 4, LLaMA-70B, and LLaMA-8B. All models are queried using a fixed multiple-choice prompt format with

consistent decoding settings. Responses are mapped to answer options and evaluated for correctness. Full model details and inference parameters are provided in Appendix B.4.

We evaluate four context conditions: **no-context**, **perfect context**, **misleading context**, and **distraction context**, which respectively test parametric recall, evidence utilization, and robustness to incorrect or irrelevant information.

Evaluation metrics. We evaluate model responses as binary correctness variables $x_i \in \{0, 1\}$ and compute accuracy $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} x_i$ with standard error $\sqrt{\hat{p}(1-\hat{p})/n}$. Confidence intervals are reported to reflect estimation uncertainty.

Stratified analysis. To study how context interacts with prior knowledge, we compute baseline no-context accuracy for each index m, denoted $\hat{p}_m$, and use it as a proxy for parametric knowledge. We then group indices into bins based on $\hat{p}_m$ and compute context-specific accuracy within each bin:

$$\hat{p}_{\text{context},b} = \frac{\sum_{i=1}^{n} x_{\text{context},i} \cdot \mathbb{I}(\hat{p}_{m_i} \in (b_{\min}, b_{\max}])}{\sum_{i=1}^{n} \mathbb{I}(\hat{p}_{m_i} \in (b_{\min}, b_{\max}])}.$$

Context sensitivity. We measure how predictions change relative to the no-context baseline using:

- **Correction rate:** incorrect $\rightarrow$ correct with context.
- **Misleading rate:** correct $\rightarrow$ incorrect under false context.
- **Distraction rate:** correct $\rightarrow$ incorrect under irrelevant context.

Statistical modeling. We model correctness for each example i and model m as

$$Y_i^{(m)} \sim \text{Bernoulli}(p),$$

and compare probabilities across conditions to test five hypotheses:

H1) Directional asymmetry: p differs between inductive and deductive questions.

H2) Geographic variation: p varies across indices g.

H3) Dependence on prior knowledge: accuracy with perfect context varies with baseline accuracy.

H4) Contextual errors: misleading and distraction effects depend on baseline knowledge.

H5) Model effects: p varies across models m.

5 Results

Directional asymmetry. We analyze whether question direction affects difficulty. Table 2 shows that inductive questions are consistently easier than deductive ones across models. This asymmetry reflects two factors: inductive questions resemble direct factual lookup, while deductive questions require identifying the correct entity among plausible alternatives, making them more sensitive to distractors and entity coverage. Deductive reasoning is therefore a stricter test of discriminative knowledge. The gap is not uniform across models, with smaller models showing the largest asymmetry and larger models narrowing it.

We next analyze no-context performance to characterize baseline parametric knowledge across markets.

No-context performance varies sharply across markets. Figure 3 reports no-context factual accuracy by index and reveals pronounced cross-market variation. For every model, performance differs substantially across indices, with higher accuracy in large, English-dominant markets and lower accuracy in smaller or less-represented ones. This indicates

Model	Deductive Accuracy	Inductive Accuracy	Odds Ratio (95% CI)	p-value
LLaMA-8B	0.57	0.67	0.66 [0.62, 0.70]	0.00
LLaMA-70B	0.70	0.76	0.75 [0.71, 0.81]	0.00
Claude Sonnet 4	0.78	0.80	0.76 [0.71, 0.82]	0.00
GPT-5 nano	0.82	0.82	0.92 [0.86, 0.99]	0.02
GPT-5 mini	0.89	0.89	1.00 [0.93, 1.08]	1.00
GPT-5†	0.73	0.78	1.01 [0.92, 1.10]	0.87

Table 2: Directional effect of question framing. Deductive questions are harder than inductive ones for most models. †GPT-5's lower deductive accuracy relative to GPT-5 mini is an exception to this trend; we did not identify a parsing or evaluation artifact underlying it and report it as observed.

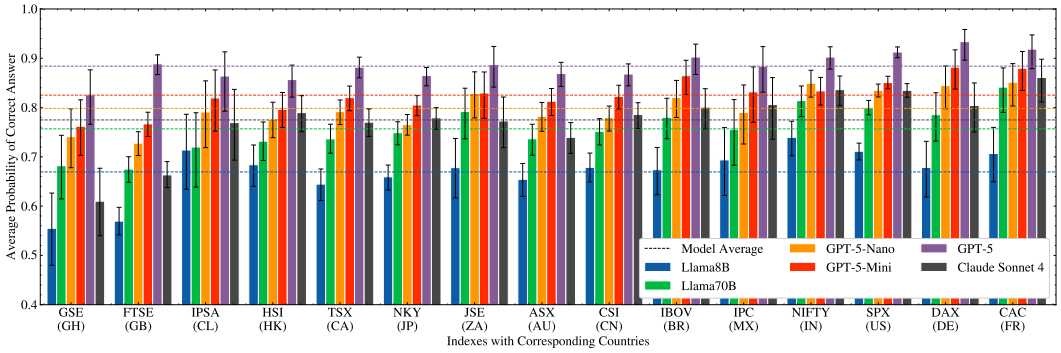
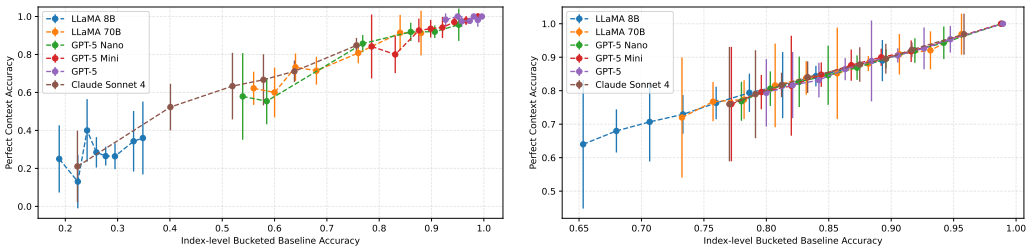


Figure 3: No-context factual accuracy across global equity indices. Grouped bars show the expected probability of a correct answer, with uncertainty intervals, for each model within each index. Horizontal reference lines indicate each model's global mean.

that parametric knowledge is systematically shaped by representational exposure, rather than being uniformly distributed. The consistency of this pattern across both proprietary and open-weight models suggests that it reflects a general property of LLMs rather than model-specific behavior. While larger models achieve higher overall accuracy, they preserve the same relative disparities, indicating that scale improves performance but does not eliminate uneven knowledge distribution. Attribute-level no-context results are provided in Appendix Figure 7.

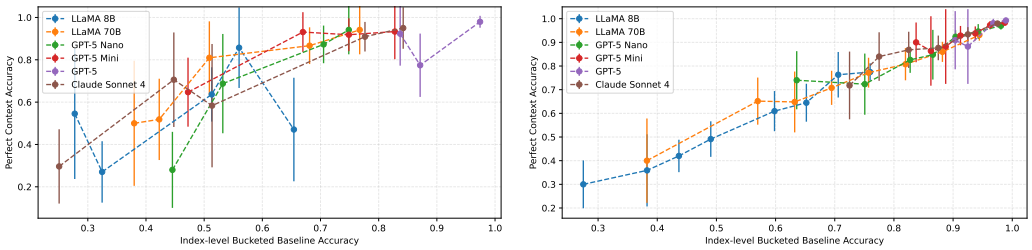
Perfect context improves performance, but not uniformly. Figure 4 shows that supplying correct contextual evidence improves accuracy across all models, confirming that retrieval can enhance factual QA. However, these gains are not uniform: markets with stronger no-context performance tend to remain stronger even with context. If retrieval acted as an independent corrective, accuracy would converge across markets; instead, we observe a coupling between baseline parametric knowledge and evidence utilization. This pattern is consistent across attributes and models, with a modest dampening effect for larger models. The corresponding stratified numerical results are reported in Appendix Tables 8–11.

Misleading context induces systematic copying failures. Figure 5 shows that when context is misleading, accuracy drops substantially, as models often adopt incorrect evidence rather than resist it. This is not merely the absence of benefit but an active degradation: models can perform worse with incorrect context than under parametric recall alone. This pattern aligns with prior-context conflict findings in RAG systems, here observed in a globally heterogeneous domain. Models that follow context too readily may appear grounded while remaining brittle, which is particularly problematic in high-stakes settings where visible evidence can create false trust. While stronger parametric knowledge can sometimes mitigate this effect, the pattern is inconsistent and depends on the task. In general, larger



(a) Founding Year (Appendix Table 8)

(b) Industry (Appendix Table 9)

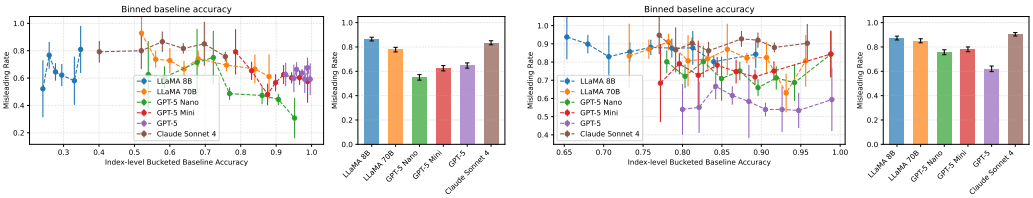


(c) Key People (Appendix Table 10)

(d) Headquarters (Appendix Table 11)

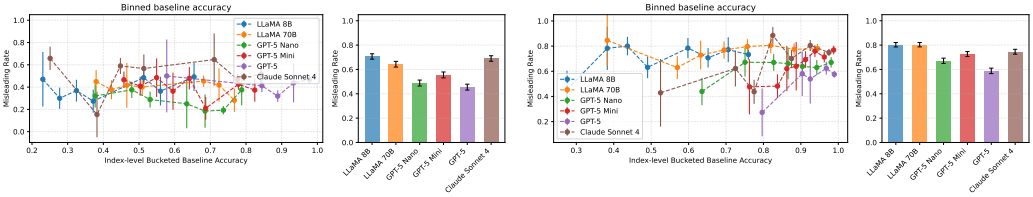
Figure 4: Effect of perfect context across indices. Positive shifts indicate that models benefit from provided evidence, with different magnitude by markets and model families.

models show greater robustness, though they remain susceptible to misleading information. The corresponding stratified numerical results are reported in Appendix Tables 12–15.



(a) Founding Year (Appendix Table 12)

(b) Industry (Appendix Table 13)

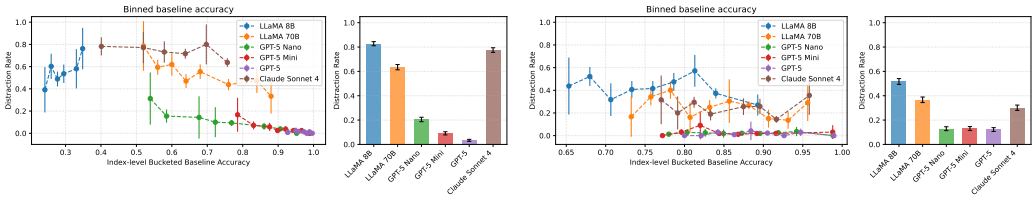


(c) Key People (Appendix Table 14)

(d) Headquarters (Appendix Table 15)

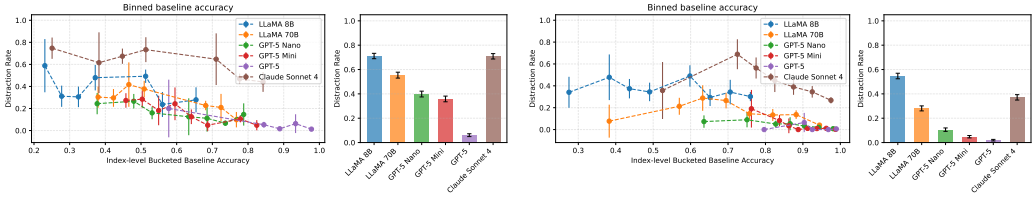
Figure 5: Effect of misleading context across indices. Decreasing misleading rates indicate the ability of models to use prior information to avoid relying on false information.

Distraction context degrades performance more mildly but systematically. Figure 6 shows a similar pattern under irrelevant rather than false context. The effect is more consistent and generally weaker than in the misleading setting, suggesting that models are more vulnerable to locally plausible false evidence than to merely irrelevant text. Larger models and indices with stronger baseline knowledge are more robust overall, though distractor difficulty may be attenuated in non-English-dominant markets due to the English-centric embedding model used for distractor selection. The corresponding stratified numerical results are reported in Appendix Tables 16–19.



(a) Founding Year (Appendix Table 16)

(b) Industry (Appendix Table 17)



(c) Key People (Appendix Table 18)

(d) Headquarters (Appendix Table 19)

Figure 6: Effect of distracting context across indices. Decreasing distraction rates indicate the ability of models to use prior information to avoid relying on irrelevant information.

Model scale improves performance but does not remove inequality. Across all conditions, larger models outperform smaller ones in absolute terms: they achieve higher baseline accuracy, benefit more from helpful context, and are more robust to misleading or distracting evidence. This suggests that scale strengthens both factual recall and evidence utilization.

However, scale does not alter the qualitative structure of the results. The same markets remain comparatively easy or difficult, and patterns of retrieval coupling and contextual brittleness persist. Thus, scale acts as an additive improvement—shifting performance upward without eliminating underlying geo-economic disparities in knowledge and evidence use.

6 Discussion

Taken together, the results support all five hypotheses: factual QA performance reflects an interaction between parametric knowledge and contextual evidence. **H1**: inductive questions are easier than deductive ones. **H2**: no-context accuracy varies sharply across indices, reflecting uneven knowledge. **H3**: correct context improves accuracy but does not close these gaps, with gains tied to baseline knowledge. **H4**: misleading context degrades performance, showing over-reliance on incorrect evidence. **H5**: larger models improve performance and robustness without removing structural disparities.

These results challenge retrieval as a universal corrective. Instead, it is a conditional amplifier: helping where models are already strong, while offering less benefit and more fragility where knowledge is sparse. Because aggregate accuracy hides systematic weaknesses, evaluation should be stratified by geography, entity coverage, and context condition; robustness to misleading evidence should be a core requirement, motivating validation layers, structured data integration, and provenance-aware design in high-stakes domains.

7 Conclusion

We introduced a benchmark for factual QA over public companies that uses geo-economic heterogeneity to study knowledge and retrieval in large language models. Accuracy is uneven across markets; correct context narrows but does not eliminate these gaps because gains remain tied to baseline accuracy; and misleading context induces systematic errors, a failure mode that persists even as larger models improve.

Overall, retrieval is not a universal corrective, underscoring the need for stratified, robustness-oriented evaluation in high-stakes settings. The benchmark dataset is available at https://zenodo.org/records/19359640.

Limitations. The benchmark relies on Wikipedia-derived fields and summaries—reproducible, but unlike production retrieval—and targets atomic facts rather than long-form or multi-hop reasoning. The misleading-context condition is synthetic; real-world errors may be subtler. Coverage quality and distractor difficulty may vary across markets, a confound multilingual or locally sourced corpora could reduce. Finally, the benchmark is English-centric and limited to US/European models; extending to non-Western models, open-ended factuality evaluation, and an "I don't know" option for calibration are natural next steps.

Acknowledgments

We would like to thank Sebastian Gehrmann, Ivailo Dimov, David Rosenberg, and Arun Verma for the insightful discussions, which helped improve this paper.

References

Su Lin Blodgett, Solon Barocas, Hal Daum'e, and Hanna M. Wallach. Language (technology) is power: A critical survey of "bias" in nlp. In *Annual Meeting of the Association for Computational Linguistics*, pp. 5454–5476. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.485.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. Improving language models by retrieving from trillions of tokens. In *International Conference on Machine Learning*, 2021.

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matthew I. Beane, Ting-Hao 'Kenneth' Huang, Bryan R. Routledge, et al. Finqa: A dataset of numerical reasoning over financial data. In *Conference on Empirical Methods in Natural Language Processing*, pp. 3697–3711. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.emnlp-main.300.

Rémy Decoupes, Roberto Interdonato, Mathieu Roche, Maguelonne Teisseire, and Sarah Valentin. Evaluation of geographical distortions in language models. *Machine-mediated learning*, pp. 86–100, 2024. doi: 10.1007/s10994-025-06916-9.

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, E. Hovy, Hinrich Schütze, and Yoav Goldberg. Measuring and improving consistency in pretrained language models. *Transactions of the Association for Computational Linguistics*, 9:1012–1031, 2021. doi: 10.1162/tacl_a_00410.

Thibault Févry, Livio Baldini Soares, Nicholas FitzGerald, Eunsol Choi, and T. Kwiatkowski. Entities as experts: Sparse memory access with entity supervision. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Conference on Empirical Methods in Natural Language Processing*, pp. 4937–4951. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.400.

Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Zhao, N. Lao, Hongrae Lee, Da-Cheng Juan, et al. Rarr: Researching and revising what language models say, using language models. In *Annual Meeting of the Association for Computational Linguistics*, pp. 16477–16508. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2023.acl-long.910.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. Realm: Retrieval-augmented language model pre-training. In *ICML*, 2020.

Dirk Hovy and Shrimai Prabhumoye. Five sources of bias in natural language processing. *Language and Linguistics Compass*, 15(8), 2021. doi: 10.1111/lnc3.12432.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.*, 43(2):1–55, 2023. doi: 10.1145/3703155.

Gautier Izacard and Edouard Grave. Leveraging passage retrieval with generative models for open domain question answering. In *Conference of the European Chapter of the Association for Computational Linguistics*, pp. 874–880. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2021.eacl-main.74.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. Large language models struggle to learn long-tail knowledge. In *International Conference on Machine Learning*, 2022.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *Conference on Empirical Methods in Natural Language Processing*, pp. 6769–6781. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.550.

T. Kwiatkowski, J. Palomaki, Olivia Redfield, Michael Collins, Ankur P. Parikh, Chris Alberti, D. Epstein, I. Polosukhin, Jacob Devlin, Kenton Lee, et al. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466, 2019. doi: 10.1162/tacl_a_00276.

Angeliki Lazaridou, Elena Gribovskaya, Wojciech Stokowiec, and Nikolai Grigorev. Internet-augmented language models through few-shot prompting for open-domain question answering, 2022. URL https://arxiv.org/abs/2203.05115.

Patrick Lewis, Ethan Perez, Aleksandara Piktus, F. Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Kuttler, M. Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Neural Information Processing Systems*, 2020.

Yinheng Li, Shaofei Wang, Han Ding, and Hang Chen. A survey of large language models in finance (finllms). *4th ACM International Conference on AI in Finance*, pp. 374–382, 2023. doi: 10.1145/3604237.3626869.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Annual Meeting of the Association for Computational Linguistics*, pp. 9802–9822. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2023.acl-long.546.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12076–12100. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.741.

Mazda Moayeri, Elham Tabassi, and Soheil Feizi. Worldbench: Quantifying geographic disparities in llm factual recall. In *Conference on Fairness, Accountability and Transparency*, pp. 1211–1228. ACM, 2024. doi: 10.1145/3630106.3658967.

Seong-Il Park and Jay-Yoon Lee. Toward robust ralms: Revealing the impact of imperfect retrieval on retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 12:1686–1702, 2024. doi: 10.1162/tacl_a_00724.

F. Petroni, Tim Rocktäschel, Patrick Lewis, A. Bakhtin, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. Language models as knowledge bases? In *Conference on Empirical Methods in Natural Language Processing*, pp. 2463–2473. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1250.

F. Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vassilis Plachouras, Tim Rocktaschel, et al. KILT: A benchmark for knowledge-intensive language tasks. In *North American Chapter of the Association for Computational Linguistics*, pp. 2523–2544. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2021.naacl-main.200. URL https://aclanthology.org/2021.naacl-main.200/.

Adam Roberts, Colin Raffel, and Noam Shazeer. How much knowledge can you pack into the parameters of a language model? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 5418–5426. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.437.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. Fever: a large-scale dataset for fact extraction and verification. In *North American Chapter of the Association for Computational Linguistics*, pp. 809–819. Association for Computational Linguistics, 2018. doi: 10.18653/v1/N18-1074.

Kevin Wu, Eric Wu, and James Zou. Clasheval: Quantifying the tug-of-war between an llm's internal prior and external evidence. In *Neural Information Processing Systems*, pp. 33402–33422. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2024. doi: 10.52202/079017-1053.

Fengbin Zhu, Wenqiang Lei, Youcheng Huang, Chao Wang, Shuo Zhang, Jiancheng Lv, Fuli Feng, and Tat-seng Chua. Tat-qa: A question answering benchmark on a hybrid of tabular and textual content in finance. In *Annual Meeting of the Association for Computational Linguistics*, pp. 3277–3287. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.acl-long.254.

A Data

A.1 Dataset Retrieval and Composition

Data retrieval proceeds in two stages. First, we collect Wikipedia pages corresponding to the target equity indices and parse their constituent company lists. Second, for each company page, we extract the lead summary paragraph and relevant infobox fields. Duplicate entities are merged after normalization.

Table 3 summarizes the dataset composition and corresponding Wikipedia sources.

A.2 Company pages

For each company page, we extract the lead paragraph and the fields used to instantiate the four target attributes as shown in Figure 1. Each of the constituent companies was accessed on 2026-01-03. We omit all but the four fields (Headquarters, Founding Year, Key People, and Industry) along with the company's name and leading paragraph.

A.3 Preprocessing and Normalization

Many of the fields are not normalized (i.e., headquarters may be "Paris" or "Paris, France"), which can cause confusion. To avoid these issues, we more aggressively normalize our raw data. In a few cases, to parse the strings, we use `LLaMA 3.1 70B` with *temperature* $= 0.02$. We select infobox fields that are both common enough and semantically appropriate for atomic factual QA. Fields with extreme sparsity or highly unstable formatting are excluded.

A.3.1 Normalizing Founding Year

We resolve the founding field to the first valid four-digit year after removing markup and auxiliary text.

Index	Country	Reg	Wiki Page	#
GSE	Ghana	AF	GSE_Composite_Index	33
N225	Japan	AS	Nikkei_225	458
FTSE	United Kingdom	EU	FTSE_250_Index	250
IBOV	Brazil	LA	List_of_companies_listed_on_B3	68
ASX200	Australia	OC	S&P/ASX_200	146
HSI	Hong Kong	AS	Hang_Seng_Index	81
JSE	South Africa	AF	FTSE/JSE_Top_40_Index	37
TSX	Canada	NA	S&P/TSX_Composite_Index	171
IPSA	Chile	LA	Indice_de_Precios_Selectivo_de_Acciones	23
CSI300	China	AS	CSI_300_Index	191
IPC	Mexico	LA	Indice_de_Precios_y_Cotizaciones	30
SPX	United States	NA	List_of_S&P_500_companies	498
DAX	Germany	EU	DAX	41
NIFTY50	India	AS	NIFTY_50	98
FRA	France	EU	CAC_40	40
Total	–	–	–	2165

Table 3: Dataset composition and sources (Reg: region).

	Missingness	Uniqueness
Company Name	0.00	96.98
Index Name	0.00	0.81
Summary	0.54	94.74
Website	8.95	90.17
Headquarters	12.19	68.55
Founding Year	13.70	72.75
Company Type	14.94	5.33
Traded As	15.91	93.97
Industry	16.40	43.87
Key People	16.94	94.09
Revenue	25.62	93.33
Number of Employees	26.91	91.37
Net Income	29.61	92.80
Total Assets	45.36	93.29
Products	46.55	83.75
Operating Income	48.81	92.62
Total Equity	50.81	92.43
Area Served	60.84	25.48
Subsidiaries	68.45	93.33
Founder	71.41	91.51

Table 4: Raw Wikipedia-derived fields, with fields used in this study highlighted in gray.

A.3.2 Normalizing Key People

These list-valued fields are first split by their delimiters, then stripped of excess whitespace, title-cased, and finally sorted into a canonical order.

A.3.3 Normalizing Industry

Industry is by far the most open-ended field in this study and required multiple normalization steps. Industry items are normalized using linked page titles where available. All values are then title-cased and sorted. The prompt used for determining the GICS sector is as follows:

> Look at the unstructured industries mentioned in the following text and extract a list of all the industries for this company. Resolve each to one of the Global Industry Classification Standard (GICS) sectors:
> - Communication Services
> - Consumer Discretionary
> - Consumer Staples
> - Energy
> - Financials
> - Health Care
> - Industrials
> - Information Technology
> - Materials
> - Real Estate
> - Utilities
>
> Respond with a JSON object such as {"industries": ["Health Care", "Financials"]}

The results are then title-cased, and duplicates are removed. We also verify if the results are in the list of GICS sectors. Finally, we sort the industries lexically to give a canonical list.

A.3.4 Normalizing Headquarters

Each location string is split on commas and reduced to city and country by passing it to the aforementioned model with the following prompt.

> Extract the city and country (Full Name) as a list of JSON objects. Leave any unknown field blank.
>
> Response should follow { "locations": [{"city": "Sydney", "country": "Australia"}]}

Country aliases are mapped to canonical names by a human reviewer who resolves names like "People's Republic of China" to "China". Furthermore, we resolve all variations caused by differences in diacritics and accents.

A.3.5 Normalizing Summaries

The leading and trailing whitespace is removed from each leading paragraph on the Wikipedia pages. We then query the aforementioned LLM to identify all direct mentions of the company and mask them with `[COMPANY]` using the following prompt:

> Consider the following summary for {{company}} and mask all mentions of {{company}} with [COMPANY]. Do not change any other part of the text.
> Summary: {{summary}}
> Response should follow { "masked_summary": "..." }

Each of the responses is then compared to the original to ensure that all non-mention text remains unchanged. This additional field is referred to as the `masked_summary`.

B Methodology

B.1 Prompt Templates

Below is the prompt structure used for multiple-choice evaluation.

B.1.1 No-context Prompt

> You are answering a factual multiple-choice question about a public company.
> Question: {{question}}
> A) {{choice_A}}
> B) {{choice_B}}
> C) {{choice_C}}
> D) {{choice_D}}
> Select the most appropriate choice and respond with only one of A, B, C, or D as a JSON object like { "response": "A" }.

B.1.2 Perfect/Misleading/Distraction Context Prompt

> You are answering a factual multiple-choice question about a public company. Use the provided context if helpful.
> Context: {{context}}
> Question: {{question}}
> A) {{choice_A}}
> B) {{choice_B}}
> C) {{choice_C}}
> D) {{choice_D}}
> Select the most appropriate choice and respond with only one of A, B, C, or D as a JSON object like { "response": "A" }.

B.1.3 Context Construction

Correct-context examples use the company's original Wikipedia lead paragraph. Misleading-context examples are generated by replacing the target attribute value in the lead paragraph with an incorrect but plausible value drawn from another entity, while preserving the surrounding text as much as possible. This creates minimally perturbed evidence that is locally coherent but factually wrong with respect to the queried attribute.

B.2 Constructing Contexts

To study the effects of providing contexts with varying veracity and relevance, we construct perfect, misleading, and distraction contexts. These contexts are constructed per field, per company, and per index, and correspond to the same question asked with different additional information. This section details how these contexts are constructed.

B.2.1 Perfect Context

The perfect context was constructed using the unchanged leading paragraph from each company's Wikipedia page. For certain attributes, the summary did not contain the fact being tested. For example, the founding year may not be mentioned in the summary for Apple. To control for these absences, we filter our questions and responses for the perfect context scenario. These filters are specific to the fields being queried.

- Headquarters: We use the regex `\b{{city}}\b` to verify if the city of the correct headquarters is mentioned in the summary.
- Founding Year: We use the regex `\b{{founding year}}\b` to verify if the founding year of the company is mentioned in the summary.
- Key People: We use the regex `\b{{last name}}\b` to verify if the last name of the correct key person is mentioned in the summary.
- Industry: Due to the central nature of the industry of a company, we make the assumption that each summary contains at least some information regarding the industry. Hence, no filtering is applied to questions regarding Industry.

- Across the benchmark, these filters exclude 61.0% of Founding Year questions, 29.6% of Headquarters questions, and 86.0% of Key People questions from the perfect-context condition, reflecting how rarely Wikipedia lead paragraphs restate infobox facts verbatim (Table 5). Exclusion is highest for Key People across every index, consistent with lead paragraphs rarely naming executives by surname.

Index	Founding Year	Headquarters	Key People
GSE	38.7	73.3	100.0
N225	56.6	31.0	93.2
FTSE	73.0	29.7	84.9
IBOV	54.5	31.8	87.9
ASX200	65.2	62.6	94.0
HSI	60.5	31.2	79.7
JSE	55.9	32.4	97.1
TSX	64.3	25.4	85.5
IPSA	76.2	56.5	95.0
CSI300	62.0	29.5	96.3
IPC	78.6	39.3	76.0
SPX	58.7	18.7	82.7
DAX	41.5	17.1	80.0
NIFTY50	63.2	21.6	70.3
FRA	37.5	25.6	72.5

Table 5: Perfect-context exclusion rate (%) by attribute and index.

B.2.2 Misleading Context

To construct a purposefully misleading context, we follow these steps:

1. Select any of the incorrect options for this question.
2. Find its `masked_summary`.
3. Replace all mentions of `[COMPANY]` with the name of the company we are studying.

Consider the example in which we ask the model about Apple's industry and have provided the industry for Johnson & Johnson (Health Care) as a choice. A misleading context here would replace all mentions of Johnson & Johnson in its summary with Apple, as shown below.

> **Apple** is an American multinational pharmaceutical, biotechnology, and medical technologies corporation headquartered in New Brunswick, New Jersey, and publicly traded on the New York Stock Exchange. Its common stock is a component of the Dow Jones Industrial Average, and the company is ranked No. 48 on the 2025 Fortune 500 list of the largest United States corporations. In 2025, the company was ranked 42nd in the Forbes Global 2000. **Apple** has a global workforce of approximately 138,000...

B.2.3 Distraction Context

The distraction context is constructed in the same manner as the misleading context with one small change. Instead of replacing the name of the target company, the incorrect company's name is left in the context. Using the same example as above, we leave Johnson & Johnson in the summary untouched.

B.3 Robustness and Coverage Diagnostics

Summary length. Mean Wikipedia lead-paragraph length varies by index (60–158 words) but does not significantly predict perfect-context accuracy (Pearson $r = 0.432$, $p = 0.11$, $n = 15$ indices), suggesting that source-text length is not the primary driver of cross-market performance gaps.

Distractor similarity. Since distractors are selected using an English-centric embedding model (`bge-large-en-v1.5`), we test whether this makes distractors easier to rule out for non-English-market companies. Mean TF-IDF cosine similarity between target and distractor summaries is higher for English-market indices (GSE, FTSE, JSE, TSX, ASX200, SPX; mean $= 0.061$) than non-English-market indices (mean $= 0.046$), though the difference is not significant (t-test, $p = 0.24$, $n = 15$). This does not support a strong confound: distractors are, if anything, no easier for non-English markets.

Company-size (revenue quintile) robustness. To test whether geographic disparities persist after controlling for company size, we regress answer accuracy on within-index Wikipedia revenue quintiles (quintile 1 = smallest), together with index fixed effects. Across all four context conditions, size is a significant predictor of accuracy (quintile coefficient $p < 0.01$ in every condition; Table 6), but index fixed effects remain jointly significant after controlling for it ($p < 0.001$ in every condition), and only 1 of 15 indices (SPX) shows strictly monotonic accuracy across quintiles. Geo-economic disparities are therefore not fully explained by company size.

Condition	Quintile coef.	Quintile p	Index FE p
No-Context	0.036	< 0.001	< 0.001
Perfect Context	0.041	< 0.001	< 0.001
Distraction Context	0.040	< 0.001	< 0.001
Misleading Context	0.030	0.002	< 0.001

Table 6: Company-size (revenue quintile) regression, controlling for index fixed effects.

B.4 Models and Inference Metadata

B.4.1 Model Details

Here we specify the hyperparameter settings used when querying the different models, along with the dates on which the models were called and their specific versions. These details are listed in Table 7.

Each of these models was accessed using the Python OpenAI Chat Completions API with internal Bloomberg-hosted endpoints. Each question was asked individually in its own query (i.e., the queries were not batched).

B.4.2 Parsing LLM Responses

All of our prompts ask the models to respond using JSON objects. We find the first instance of the { character and the last instance of the } character. The string enclosed by these two characters is extracted, parsed as a JSON object, and used for the appropriate analysis. Any deviations that make it impossible to parse the JSON response are excluded from the analysis.

C Results

In all result tables, *Center* denotes the midpoint of the corresponding baseline-accuracy bin used for stratified analysis. Rows correspond to quantiles of baseline no-context accuracy, and tables report either *Accuracy*, *Misleading Rate*, or *Distraction Rate*, depending on the evaluation setting, each with its associated uncertainty.

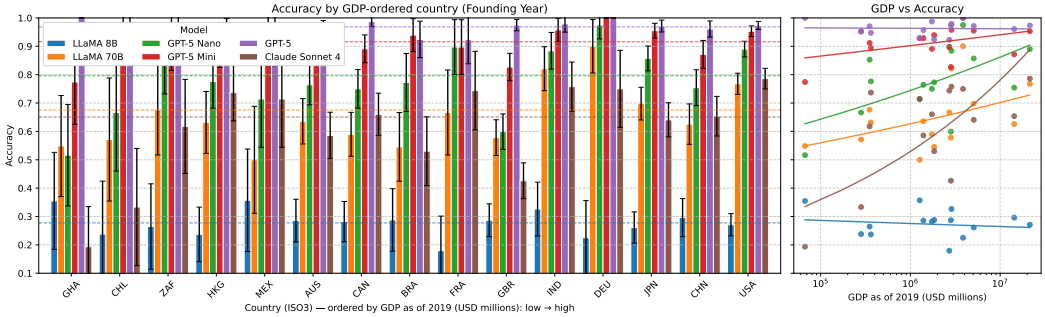
C.1 No-Context

We begin by evaluating model performance without external evidence, isolating parametric knowledge. For stratified analyses, companies are grouped into bins according to baseline no-context accuracy.

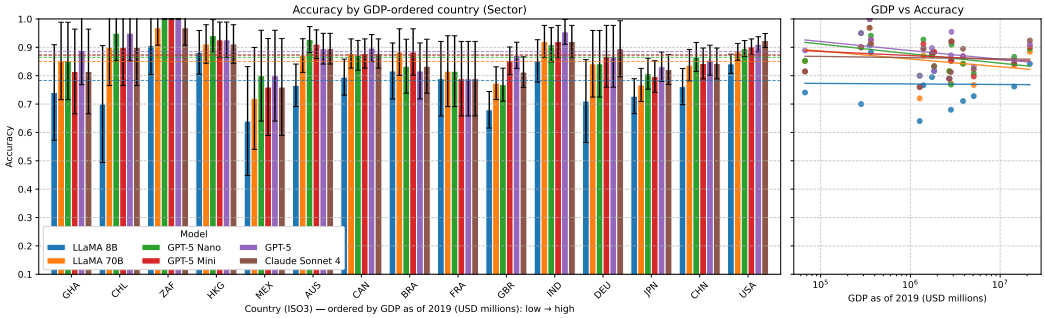
Model	Parameters	Version	Temperature	Max Output Tokens
LLaMA	8B	3.1	1.0	4,096
LLaMA	70B	3.1	1.0	8,192
GPT-5 nano	$-^{*}$	5.0	$-^{+}$	128,000
GPT-5 mini	$-^{*}$	5.0	$-^{+}$	128,000
GPT-5	$-^{*}$	5.0	$-^{+}$	128,000
Claude Sonnet	$-^{*}$	4.0	1.0	64,000

Table 7: Language Model Access Details. *: Closed source models do not disclose details regarding parameters. $^{+}$: OpenAI does not permit users to specify temperature.

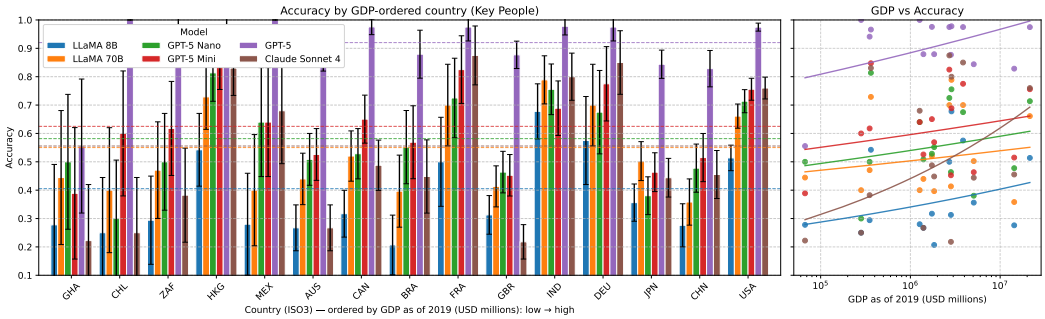
Figure 7 summarizes no-context performance across attributes and indices.



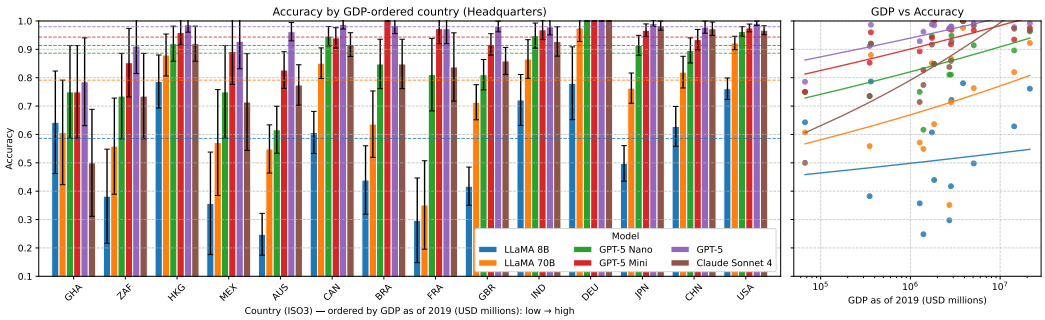
(a) Founding Year



(b) Industry



(c) Key People



(d) Headquarters

Figure 7: Accuracy by Index by GDP

C.2 Perfect Context

We next evaluate performance when correct contextual evidence is provided. This setting measures the extent to which retrieval improves factual QA and whether gains depend on prior knowledge.

Stratified results across attributes are reported in Tables 8–11.

model	Claude Sonnet 4		GPT-5		GPT-5 Mini		GPT-5 Nano		LLaMA-70B		LLaMA-8B	
	Accuracy	Center	Accuracy	Center	Accuracy	Center	Accuracy	Center	Accuracy	Center	Accuracy	Center
Quantile												
0.1	0.21 ± 0.10	0.22	0.98 ± 0.02	0.93	0.84 ± 0.09	0.79	0.58 ± 0.12	0.54	-	-	0.25 ± 0.09	0.19
0.2	-	-	-	-	-	-	0.55 ± 0.06	0.58	0.62 ± 0.04	0.56	-	-
0.3	-	-	-	-	0.80 ± 0.05	0.83	-	-	0.60 ± 0.07	0.60	0.13 ± 0.07	0.22
0.4	0.52 ± 0.06	0.40	1.00 ± 0.00	0.95	-	-	-	-	0.73 ± 0.04	0.64	0.40 ± 0.08	0.24
0.5	-	-	0.97 ± 0.01	0.96	0.93 ± 0.03	0.88	-	-	0.71 ± 0.04	0.68	0.28 ± 0.04	0.26
0.6	0.63 ± 0.09	0.52	-	-	0.94 ± 0.02	0.90	0.86 ± 0.02	0.77	-	-	0.26 ± 0.03	0.28
0.7	0.67 ± 0.07	0.58	0.98 ± 0.01	0.97	0.94 ± 0.03	0.92	-	-	0.81 ± 0.03	0.76	0.26 ± 0.03	0.29
0.8	0.71 ± 0.03	0.64	1.00 ± 0.00	0.98	0.97 ± 0.01	0.94	0.92 ± 0.02	0.86	-	-	-	-
0.9	-	-	0.98 ± 0.02	0.99	0.98 ± 0.01	0.97	0.92 ± 0.02	0.91	0.91 ± 0.05	0.84	0.34 ± 0.08	0.33
1.0	0.85 ± 0.02	0.76	1.00 ± 0.00	1.00	1.00 ± 0.00	0.99	0.96 ± 0.04	0.95	0.91 ± 0.06	0.88	0.36 ± 0.10	0.35

Table 8: Statistics for Accuracy given Perfect Context for questions about Founding Year.

model	Claude Sonnet 4		GPT-5		GPT-5 Mini		GPT-5 Nano		LLaMA-70B		LLaMA-8B	
	Accuracy	Center	Accuracy	Center	Accuracy	Center	Accuracy	Center	Accuracy	Center	Accuracy	Center
Quantile												
0.1	0.76 ± 0.09	0.77	0.79 ± 0.05	0.80	0.76 ± 0.09	0.77	0.77 ± 0.03	0.78	0.72 ± 0.09	0.73	0.64 ± 0.10	0.65
0.2	0.79 ± 0.07	0.79	0.82 ± 0.05	0.82	0.80 ± 0.03	0.80	0.81 ± 0.03	0.80	0.77 ± 0.03	0.76	0.68 ± 0.03	0.68
0.3	0.82 ± 0.02	0.81	0.83 ± 0.03	0.84	0.81 ± 0.08	0.82	0.83 ± 0.04	0.83	0.77 ± 0.03	0.78	0.71 ± 0.06	0.71
0.4	0.84 ± 0.02	0.83	0.86 ± 0.02	0.86	0.85 ± 0.02	0.84	0.85 ± 0.05	0.85	0.82 ± 0.06	0.81	0.73 ± 0.03	0.73
0.5	-	-	0.89 ± 0.06	0.88	0.88 ± 0.02	0.87	0.87 ± 0.02	0.87	0.84 ± 0.03	0.83	0.76 ± 0.02	0.76
0.6	0.88 ± 0.03	0.87	0.91 ± 0.01	0.91	0.90 ± 0.01	0.89	0.90 ± 0.01	0.90	0.85 ± 0.07	0.86	0.79 ± 0.03	0.79
0.7	0.90 ± 0.02	0.90	0.93 ± 0.03	0.93	0.92 ± 0.02	0.92	0.92 ± 0.02	0.92	0.88 ± 0.01	0.88	0.82 ± 0.05	0.81
0.8	0.92 ± 0.01	0.92	0.95 ± 0.02	0.95	-	-	0.94 ± 0.02	0.94	0.91 ± 0.03	0.91	0.84 ± 0.02	0.84
0.9	-	-	-	-	-	-	-	-	0.92 ± 0.03	0.93	-	-
1.0	0.97 ± 0.03	0.96	1.00 ± 0.00	0.99	1.00 ± 0.00	0.99	1.00 ± 0.00	0.99	0.97 ± 0.03	0.96	0.89 ± 0.03	0.89

Table 9: Statistics for Accuracy given Perfect Context for questions about Industry.

model	Claude Sonnet 4		GPT-5		GPT-5 Mini		GPT-5 Nano		LLaMA-70B		LLaMA-8B	
	Accuracy	Center	Accuracy	Center	Accuracy	Center	Accuracy	Center	Accuracy	Center	Accuracy	Center
Quantile												
0.1	0.30 ± 0.09	0.25	0.92 ± 0.08	0.84	0.65 ± 0.08	0.47	-	-	0.50 ± 0.15	0.38	-	-
0.2	-	-	-	-	-	-	0.28 ± 0.09	0.45	0.52 ± 0.10	0.42	0.55 ± 0.16	0.28
0.3	-	-	0.77 ± 0.08	0.87	-	-	-	-	-	-	0.27 ± 0.07	0.32
0.4	0.71 ± 0.11	0.45	-	-	-	-	0.69 ± 0.12	0.53	0.81 ± 0.09	0.51	-	-
0.5	0.58 ± 0.15	0.51	-	-	-	-	-	-	-	-	-	-
0.6	-	-	-	-	0.93 ± 0.05	0.67	-	-	-	-	-	-
0.7	-	-	-	-	-	-	-	-	-	-	0.64 ± 0.07	0.51
0.8	-	-	-	-	0.92 ± 0.04	0.75	0.87 ± 0.05	0.71	0.87 ± 0.04	0.68	0.86 ± 0.10	0.56
0.9	0.91 ± 0.04	0.78	0.98 ± 0.01	0.97	-	-	0.94 ± 0.06	0.75	-	-	-	-
1.0	0.95 ± 0.05	0.84	-	-	0.93 ± 0.07	0.83	-	-	0.94 ± 0.06	0.77	0.47 ± 0.12	0.65

Table 10: Statistics for Accuracy given Perfect Context for questions about Key People.

model	Claude Sonnet 4		GPT-5		GPT-5 Mini		GPT-5 Nano		LLaMA-70B		LLaMA-8B	
	Accuracy	Center	Accuracy	Center	Accuracy	Center	Accuracy	Center	Accuracy	Center	Accuracy	Center
Quantile												
0.1	0.62 ± 0.18	0.52	0.88 ± 0.12	0.80	0.75 ± 0.16	0.76	0.74 ± 0.06	0.64	0.40 ± 0.09	0.38	0.30 ± 0.05	0.28
0.3	-	-	-	-	-	-	-	-	-	-	0.36 ± 0.08	0.38
0.4	-	-	-	-	0.90 ± 0.04	0.84	0.72 ± 0.07	0.75	0.65 ± 0.05	0.57	0.42 ± 0.03	0.44
0.5	0.72 ± 0.07	0.72	-	-	0.86 ± 0.07	0.86	-	-	0.65 ± 0.07	0.63	0.49 ± 0.04	0.49
0.6	0.84 ± 0.05	0.78	0.91 ± 0.06	0.90	0.88 ± 0.08	0.89	0.83 ± 0.03	0.83	0.71 ± 0.04	0.69	-	-
0.7	0.87 ± 0.04	0.82	0.88 ± 0.08	0.92	0.93 ± 0.02	0.91	0.85 ± 0.05	0.87	0.77 ± 0.03	0.76	0.61 ± 0.04	0.60
0.8	0.88 ± 0.03	0.88	-	-	0.94 ± 0.02	0.94	0.92 ± 0.01	0.90	0.81 ± 0.03	0.82	0.64 ± 0.04	0.65
0.9	0.93 ± 0.02	0.92	0.98 ± 0.01	0.97	0.97 ± 0.01	0.96	0.95 ± 0.02	0.94	0.86 ± 0.02	0.88	0.76 ± 0.05	0.71
1.0	0.98 ± 0.01	0.98	0.99 ± 0.00	0.99	0.98 ± 0.01	0.99	0.97 ± 0.01	0.98	0.93 ± 0.01	0.94	0.77 ± 0.02	0.76

Table 11: Statistics for Accuracy given Perfect Context for questions about Headquarters.

C.3 Misleading Context

We evaluate robustness under misleading context, where provided evidence is locally coherent but factually incorrect.

model	Claude Sonnet 4		GPT-5		GPT-5 Mini		GPT-5 Nano		LLaMA-70B		LLaMA-8B	
Quantile	Center	Rate	Center	Rate	Center	Rate	Center	Rate	Center	Rate	Center	Rate
0.1	-	-	0.93	0.63 ± 0.04	0.79	0.79 ± 0.08	0.54	0.62 ± 0.12	0.52	0.93 ± 0.07	-	-
0.2	-	-	-	-	-	-	0.58	0.61 ± 0.04	0.56	0.74 ± 0.03	-	-
0.3	-	-	-	-	0.83	0.65 ± 0.03	-	-	0.60	0.73 ± 0.05	-	-
0.4	0.40	0.79 ± 0.04	0.95	0.60 ± 0.05	-	-	0.68	0.71 ± 0.13	0.64	0.67 ± 0.03	0.24	0.52 ± 0.11
0.5	-	-	0.96	0.66 ± 0.03	0.88	0.48 ± 0.04	0.72	0.75 ± 0.10	0.68	0.74 ± 0.03	0.26	0.77 ± 0.05
0.6	0.52	0.80 ± 0.07	-	-	0.90	0.57 ± 0.03	0.77	0.49 ± 0.02	-	-	0.28	0.65 ± 0.03
0.7	0.58	0.87 ± 0.04	0.97	0.64 ± 0.02	0.92	0.63 ± 0.04	-	-	0.76	0.69 ± 0.02	0.29	0.62 ± 0.04
0.8	0.64	0.82 ± 0.02	0.98	0.59 ± 0.05	0.94	0.60 ± 0.02	0.86	0.47 ± 0.03	-	-	-	-
0.9	0.70	0.85 ± 0.08	0.99	0.68 ± 0.04	0.97	0.60 ± 0.03	0.91	0.44 ± 0.02	0.84	0.67 ± 0.05	0.33	0.58 ± 0.09
1.0	0.76	0.76 ± 0.02	1.00	0.59 ± 0.06	0.99	0.57 ± 0.08	0.95	0.31 ± 0.07	0.88	0.61 ± 0.08	0.35	0.81 ± 0.09

Table 12: Statistics for Misleading Rate given Misleading Context for questions about Founding Year.

model	Claude Sonnet 4		GPT-5		GPT-5 Mini		GPT-5 Nano		LLaMA-70B		LLaMA-8B	
Quantile	Center	Rate	Center	Rate	Center	Rate	Center	Rate	Center	Rate	Center	Rate
0.1	0.77	0.95 ± 0.05	0.80	0.54 ± 0.07	0.77	0.68 ± 0.11	0.78	0.80 ± 0.03	0.73	0.83 ± 0.09	0.65	0.94 ± 0.06
0.2	0.79	0.87 ± 0.06	0.82	0.55 ± 0.07	0.80	0.79 ± 0.03	0.80	0.72 ± 0.03	0.76	0.87 ± 0.03	0.68	0.90 ± 0.03
0.3	0.81	0.90 ± 0.02	0.84	0.67 ± 0.04	0.82	0.73 ± 0.10	0.83	0.80 ± 0.04	0.78	0.91 ± 0.02	0.71	0.83 ± 0.06
0.4	0.83	0.86 ± 0.02	0.86	0.62 ± 0.03	0.84	0.78 ± 0.02	0.85	0.71 ± 0.06	0.81	0.81 ± 0.07	0.73	0.86 ± 0.03
0.5	-	-	0.88	0.58 ± 0.10	0.87	0.75 ± 0.03	0.87	0.75 ± 0.03	0.83	0.82 ± 0.03	0.76	0.88 ± 0.02
0.6	0.87	0.93 ± 0.02	0.91	0.54 ± 0.02	0.89	0.72 ± 0.02	0.90	0.66 ± 0.02	0.86	0.87 ± 0.07	0.79	0.88 ± 0.03
0.7	0.90	0.92 ± 0.02	0.93	0.54 ± 0.06	0.92	0.75 ± 0.03	0.92	0.71 ± 0.03	0.88	0.82 ± 0.01	0.81	0.88 ± 0.05
0.8	0.92	0.88 ± 0.01	0.95	0.53 ± 0.05	-	-	0.94	0.69 ± 0.05	0.91	0.82 ± 0.04	0.84	0.80 ± 0.02
0.9	-	-	-	-	-	-	-	-	0.93	0.63 ± 0.05	-	-
1.0	0.96	0.90 ± 0.05	0.99	0.59 ± 0.09	0.99	0.84 ± 0.07	0.99	0.84 ± 0.07	0.96	0.81 ± 0.07	0.89	0.84 ± 0.04

Table 13: Statistics for Misleading Rate given Misleading Context for questions about Industry.

Tables report the *Misleading Rate*, defined as the proportion of previously correct no-context predictions that become incorrect under misleading context. The *Center* column denotes the midpoint of the corresponding baseline-accuracy bin.

Results are shown in Tables 12–15.

C.4 Distraction Context

We evaluate robustness under distraction context, where additional information is irrelevant rather than incorrect.

Tables report the *Distraction Rate*, defined as the proportion of previously correct no-context predictions that become incorrect under irrelevant context. The *Center* column denotes the midpoint of the baseline-accuracy bin.

Results are shown in Tables 16–19.

model	Claude Sonnet 4		GPT-5		GPT-5 Mini		GPT-5 Nano		LLaMA-70B		LLaMA-8B	
Quantile	Center	Rate	Center	Rate	Center	Rate	Center	Rate	Center	Rate	Center	Rate
0.1	0.25	0.66 ± 0.05	0.58	0.50 ± 0.17	-	-	-	-	0.38	0.45 ± 0.05	0.23	0.47 ± 0.12
0.2	-	-	-	-	0.46	0.47 ± 0.04	0.38	0.32 ± 0.05	0.42	0.38 ± 0.04	0.28	0.30 ± 0.05
0.3	0.38	0.15 ± 0.10	-	-	0.50	0.41 ± 0.04	-	-	0.47	0.42 ± 0.10	0.32	0.37 ± 0.05
0.4	0.45	0.59 ± 0.04	-	-	0.55	0.48 ± 0.09	0.48	0.38 ± 0.04	0.51	0.40 ± 0.04	0.37	0.27 ± 0.05
0.5	0.51	0.57 ± 0.06	-	-	0.60	0.36 ± 0.09	0.53	0.29 ± 0.04	-	-	-	-
0.6	-	-	-	-	0.64	0.48 ± 0.05	-	-	-	-	-	-
0.7	-	-	0.84	0.41 ± 0.03	0.69	0.21 ± 0.05	0.63	0.25 ± 0.11	-	-	0.51	0.48 ± 0.03
0.8	0.71	0.65 ± 0.12	0.89	0.32 ± 0.03	-	-	0.69	0.19 ± 0.08	0.68	0.45 ± 0.03	0.56	0.36 ± 0.07
0.9	0.78	0.47 ± 0.02	0.93	0.44 ± 0.09	0.78	0.43 ± 0.03	0.74	0.19 ± 0.02	0.72	0.42 ± 0.08	-	-
1.0	0.84	0.52 ± 0.05	0.98	0.47 ± 0.02	0.82	0.37 ± 0.05	0.79	0.38 ± 0.07	0.77	0.28 ± 0.05	0.65	0.49 ± 0.06

Table 14: Statistics for Misleading Rate given Misleading Context for questions about Key People.

model	Claude Sonnet 4		GPT-5		GPT-5 Mini		GPT-5 Nano		LLaMA-70B		LLaMA-8B	
	Center	Rate	Center	Rate	Center	Rate	Center	Rate	Center	Rate	Center	Rate
Quantile												
0.1	0.53	0.43 ± 0.14	0.80	0.27 ± 0.10	0.76	0.48 ± 0.11	0.64	0.44 ± 0.06	0.38	0.85 ± 0.10	0.28	0.45 ± 0.08
0.3	-	-	-	-	-	-	-	-	-	-	0.38	0.78 ± 0.09
0.4	-	-	-	-	0.84	0.48 ± 0.05	0.75	0.67 ± 0.06	0.57	0.63 ± 0.05	0.44	0.80 ± 0.04
0.5	0.72	0.62 ± 0.07	-	-	0.86	0.62 ± 0.09	-	-	0.63	0.73 ± 0.06	0.49	0.63 ± 0.04
0.6	0.78	0.44 ± 0.05	0.90	0.58 ± 0.09	0.89	0.64 ± 0.10	0.83	0.67 ± 0.03	0.69	0.77 ± 0.03	-	-
0.7	0.83	0.89 ± 0.03	0.93	0.54 ± 0.10	0.91	0.69 ± 0.03	0.87	0.66 ± 0.06	0.76	0.80 ± 0.03	0.60	0.78 ± 0.04
0.8	0.88	0.70 ± 0.03	-	-	0.94	0.76 ± 0.02	0.90	0.64 ± 0.02	0.82	0.81 ± 0.03	0.65	0.71 ± 0.04
0.9	0.93	0.81 ± 0.02	0.97	0.62 ± 0.03	0.96	0.71 ± 0.02	0.94	0.63 ± 0.03	0.88	0.77 ± 0.02	0.71	0.77 ± 0.05
1.0	0.97	0.75 ± 0.01	0.99	0.58 ± 0.01	0.99	0.77 ± 0.02	0.98	0.67 ± 0.02	0.94	0.78 ± 0.02	0.76	0.73 ± 0.02

Table 15: Statistics for Misleading Rate given Misleading Context for questions about Headquarters.

model	Claude Sonnet 4		GPT-5		GPT-5 Mini		GPT-5 Nano		LLaMA-70B		LLaMA-8B	
	Center	Rate	Center	Rate	Center	Rate	Center	Rate	Center	Rate	Center	Rate
Quantile												
0.1	-	-	0.93	0.01 ± 0.01	0.79	0.17 ± 0.08	0.54	0.31 ± 0.12	0.52	0.79 ± 0.11	-	-
0.2	-	-	-	-	-	-	0.58	0.15 ± 0.03	0.56	0.59 ± 0.03	-	-
0.3	-	-	-	-	0.83	0.07 ± 0.02	-	-	0.60	0.62 ± 0.05	-	-
0.4	0.40	0.78 ± 0.04	0.95	0.02 ± 0.02	-	-	0.68	0.14 ± 0.10	0.64	0.47 ± 0.03	0.24	0.39 ± 0.10
0.5	-	-	0.96	0.02 ± 0.01	0.88	0.06 ± 0.02	0.72	0.10 ± 0.07	0.68	0.55 ± 0.03	0.26	0.60 ± 0.06
0.6	0.52	0.77 ± 0.07	-	-	0.90	0.02 ± 0.01	0.77	0.09 ± 0.01	-	-	0.28	0.49 ± 0.03
0.7	0.58	0.73 ± 0.05	0.97	0.01 ± 0.00	0.92	0.04 ± 0.01	-	-	0.76	0.44 ± 0.03	0.29	0.54 ± 0.04
0.8	0.64	0.72 ± 0.02	0.98	0.00 ± 0.00	0.94	0.03 ± 0.01	0.86	0.06 ± 0.02	-	-	-	-
0.9	0.70	0.80 ± 0.09	0.99	0.01 ± 0.01	0.97	0.02 ± 0.01	0.91	0.04 ± 0.01	0.84	0.47 ± 0.06	0.33	0.58 ± 0.09
1.0	0.76	0.64 ± 0.02	1.00	0.00 ± 0.00	0.99	0.00 ± 0.00	0.95	0.00 ± 0.00	0.88	0.33 ± 0.08	0.35	0.76 ± 0.10

Table 16: Statistics for Distraction Rate given Irrelevant Context for questions about Founding Year.

model	Claude Sonnet 4		GPT-5		GPT-5 Mini		GPT-5 Nano		LLaMA-70B		LLaMA-8B	
	Center	Rate	Center	Rate	Center	Rate	Center	Rate	Center	Rate	Center	Rate
Quantile												
0.1	0.77	0.32 ± 0.11	0.80	0.00 ± 0.00	0.77	0.00 ± 0.00	0.78	0.01 ± 0.01	0.73	0.17 ± 0.09	0.65	0.44 ± 0.13
0.2	0.79	0.20 ± 0.07	0.82	0.00 ± 0.00	0.80	0.03 ± 0.01	0.80	0.03 ± 0.01	0.76	0.34 ± 0.04	0.68	0.52 ± 0.04
0.3	0.81	0.29 ± 0.02	0.84	0.03 ± 0.01	0.82	0.09 ± 0.06	0.83	0.02 ± 0.02	0.78	0.40 ± 0.04	0.71	0.32 ± 0.07
0.4	0.83	0.19 ± 0.03	0.86	0.01 ± 0.00	0.84	0.02 ± 0.01	0.85	0.02 ± 0.02	0.81	0.16 ± 0.07	0.73	0.41 ± 0.04
0.5	-	-	0.88	0.04 ± 0.04	0.87	0.01 ± 0.01	0.87	0.02 ± 0.01	0.83	0.25 ± 0.03	0.76	0.42 ± 0.03
0.6	0.87	0.26 ± 0.04	0.91	0.02 ± 0.01	0.89	0.02 ± 0.01	0.90	0.02 ± 0.01	0.86	0.30 ± 0.10	0.79	0.47 ± 0.04
0.7	0.90	0.26 ± 0.03	0.93	0.00 ± 0.00	0.92	0.02 ± 0.01	0.92	0.03 ± 0.01	0.88	0.27 ± 0.02	0.81	0.57 ± 0.07
0.8	0.92	0.14 ± 0.01	0.95	0.03 ± 0.02	-	-	0.94	0.04 ± 0.02	0.91	0.15 ± 0.04	0.84	0.37 ± 0.02
0.9	-	-	-	-	-	-	-	-	0.93	0.14 ± 0.04	-	-
1.0	0.96	0.35 ± 0.09	0.99	0.00 ± 0.00	0.99	0.03 ± 0.03	0.99	0.00 ± 0.00	0.96	0.29 ± 0.08	0.89	0.27 ± 0.05

Table 17: Statistics for Distraction Rate given Irrelevant Context for questions about Industry.

model	Claude Sonnet 4		GPT-5		GPT-5 Mini		GPT-5 Nano		LLaMA-70B		LLaMA-8B	
	Center	Rate	Center	Rate	Center	Rate	Center	Rate	Center	Rate	Center	Rate
Quantile												
0.1	0.25	0.75 ± 0.05	0.58	0.20 ± 0.13	-	-	-	-	0.38	0.30 ± 0.05	0.23	0.59 ± 0.12
0.2	-	-	-	-	0.46	0.27 ± 0.03	0.38	0.24 ± 0.05	0.42	0.30 ± 0.04	0.28	0.31 ± 0.05
0.3	0.38	0.62 ± 0.14	-	-	0.50	0.28 ± 0.04	-	-	0.47	0.42 ± 0.10	0.32	0.31 ± 0.05
0.4	0.45	0.67 ± 0.04	-	-	0.55	0.18 ± 0.07	0.48	0.27 ± 0.03	0.51	0.38 ± 0.04	0.37	0.48 ± 0.06
0.5	0.51	0.73 ± 0.06	-	-	0.60	0.24 ± 0.08	0.53	0.16 ± 0.03	-	-	-	-
0.6	-	-	-	-	0.64	0.12 ± 0.03	-	-	-	-	-	-
0.7	-	-	0.84	0.05 ± 0.01	0.69	0.05 ± 0.03	0.63	0.12 ± 0.09	-	-	0.51	0.49 ± 0.03
0.8	0.71	0.65 ± 0.12	0.89	0.02 ± 0.01	-	-	0.69	0.11 ± 0.06	0.68	0.23 ± 0.02	0.56	0.24 ± 0.06
0.9	0.78	0.47 ± 0.02	0.93	0.06 ± 0.04	0.78	0.11 ± 0.02	0.74	0.07 ± 0.01	0.72	0.21 ± 0.06	-	-
1.0	0.84	0.44 ± 0.05	0.98	0.01 ± 0.00	0.82	0.05 ± 0.02	0.79	0.15 ± 0.05	0.77	0.10 ± 0.04	0.65	0.28 ± 0.06

Table 18: Statistics for Distraction Rate given Irrelevant Context for questions about Key People.

model	Claude Sonnet 4		GPT-5		GPT-5 Mini		GPT-5 Nano		LLaMA-70B		LLaMA-8B	
	Center	Rate	Center	Rate	Center	Rate	Center	Rate	Center	Rate	Center	Rate
Quantile												
0.1	0.52	0.36 ± 0.13	0.80	0.00 ± 0.00	0.76	0.19 ± 0.09	0.64	0.07 ± 0.03	0.38	0.08 ± 0.08	0.28	0.34 ± 0.07
0.3	-	-	-	-	-	-	-	-	-	-	0.38	0.48 ± 0.11
0.4	-	-	-	-	0.84	0.08 ± 0.03	0.75	0.09 ± 0.04	0.57	0.21 ± 0.04	0.44	0.37 ± 0.05
0.5	0.72	0.69 ± 0.07	-	-	0.86	0.03 ± 0.03	-	-	0.63	0.29 ± 0.06	0.49	0.34 ± 0.04
0.6	0.78	0.56 ± 0.05	0.90	0.06 ± 0.04	0.89	0.00 ± 0.00	0.83	0.05 ± 0.02	0.69	0.27 ± 0.04	-	-
0.7	0.82	0.45 ± 0.05	0.92	0.00 ± 0.00	0.91	0.01 ± 0.01	0.87	0.05 ± 0.03	0.76	0.14 ± 0.03	0.60	0.49 ± 0.05
0.8	0.88	0.39 ± 0.04	-	-	0.94	0.01 ± 0.01	0.90	0.03 ± 0.01	0.82	0.13 ± 0.03	0.65	0.29 ± 0.04
0.9	0.92	0.35 ± 0.03	0.97	0.00 ± 0.00	0.96	0.01 ± 0.01	0.94	0.02 ± 0.01	0.88	0.14 ± 0.02	0.71	0.34 ± 0.06
1.0	0.98	0.27 ± 0.01	0.99	0.01 ± 0.00	0.99	0.00 ± 0.00	0.98	0.01 ± 0.00	0.94	0.04 ± 0.01	0.76	0.30 ± 0.02

Table 19: Statistics for Distraction Rate given Irrelevant Context for questions about Headquarters.