ObfusQAte: A Proposed Framework to Evaluate LLM Robustness on Obfuscated Factual Question Answering

Anonymous ACL submission

Abstract

The rapid proliferation of **Large Language Models (LLMs)** has significantly contributed to the development of equitable AI systems capable of factual question-answering (QA). However, no known study tests the LLMs' robustness when presented with obfuscated versions of questions. To systematically evaluate these limitations, we propose a novel technique, `ObfusQAte` and leveraging the same, introduce **ObfusQA**, a comprehensive, first of its kind, framework, with *multi-tiered obfuscation levels* designed to examine LLM capabilities across three distinct dimensions: **(i) Named-Entity Indirection**, **(ii) Distractor Indirection**, and **(iii) Contextual Overload**. By capturing these fine-grained distinctions in language, ObfusQA provides a comprehensive benchmark for evaluating LLM robustness and adaptability. Our study observes that LLMs exhibit a tendency to fail or generate hallucinated responses, when confronted with these increasingly nuanced variations. To foster research in this direction, we make ObfusQAte publicly available.

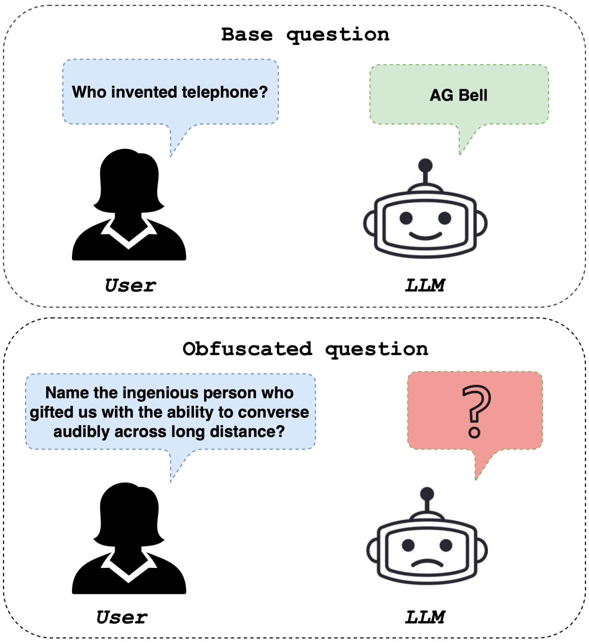

Base question
Who invented telephone?
AG Bell
User
LLM
Obfuscated question
Name the ingenious person who gifted us with the ability to converse audibly across long distance?
?
User
LLM

Figure 1: An example of Obfuscated QA between a human and an LLM. Obfuscation in the Question is expected to pose challenges for the LLM even for straightforward questions.

1 Introduction

In recent times, the Large Language Models (LLMs) like GPT (Achiam et al., 2023), LLaMA (Touvron et al., 2023), DeepSeek (Bi et al., 2024) have emerged as game-changers, showcasing unprecedented capabilities of generating coherent responses to a variety of prompts. These models have been applied to numerous tasks, such as report generation, virtual assistants, and summarization, to name a few (Manakul et al., 2023). Despite their efficacy, these models are plagued by their tendency to generate factually incorrect information with a tone of confidence often termed as *hallucination* (Azaria and Mitchell, 2023). The issue of hallucination critically hampers reliability and limits widespread adoption in real-world applications.

Based on the study by Chang et al. (2024), the current evaluation systems are categorized in many different ways, *factuality* being one of them. Factuality in the context of LLMs refers to the extent to which the information or answers provided by the model align with real-world truths and verifiable facts. Factuality in LLMs significantly impacts a variety of tasks and downstream applications, such as QA systems, information extraction, text summarization, dialogue systems, and automated fact-checking, where incorrect or inconsistent information could lead to substantial misunderstandings and misinterpretations. Therefore, evaluating factuality is critical to ensure trust in these models. This includes the ability of these models to maintain consistency with known facts, avoiding generating misleading or false information (known as "factual hallucination"). A range of methodologies have

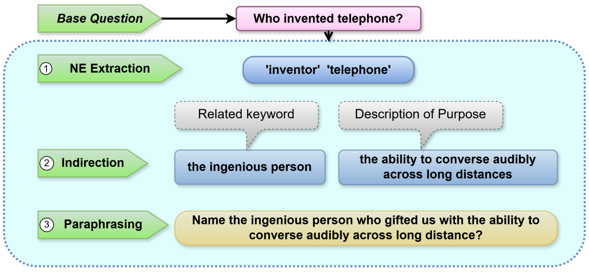


Figure 2: Named-Entity Indirection

been proposed to measure and improve the factuality of language models (Lin et al., 2021; Wang et al., 2023; Pezeshkpour, 2023; Honovich et al., 2022; Manakul et al., 2023).

There have been many attempts to improve or test QA capabilities of LLMs (Kamalloo et al., 2024; Zhuang et al., 2023), however, to our knowledge *there exists no study that tests the LLMs' ability to perceive semantically obfuscated or obscured variants of an otherwise straightforward question* (cf. A.4 for more related works). The analogy is: in an interview, the expert panel tests the knowledge depth of the candidate through intricately nuanced and yet equivalent variants of a potentially straightforward question, for which the candidate might have prepared or even *memorized* in advance. Fig. 1 depicts such a scenario, in the case of LLM evaluation.

To address this gap we propose a suite of techniques: `ObfusQAte` (Sec. 2) leading to obfuscated dataset **ObfusQA** (Sec. 2.1), a comprehensive, first of its kind, framework, with *multi-tiered obfuscation levels* designed to examine LLM capabilities across three distinct dimensions: **(i) Named-Entity Indirection**, **(ii) Distractor Indirection**, and **(iii) Contextual Overload**. Finally, we empirically benchmark the efficacy of state-of-the-art LLMs on our proposed setup (Sec. 3).

2 Proposed setup: ObfusQAte

To effectively evaluate the capabilities of LLMs across a diverse set of challenges, we establish a robust experimental framework built around the `ObfusQAte` technique. This framework leverages a comprehensive, *multi-tiered obfuscation* process to generate a diverse range of questions that are potentially more complex and challenging than a *base question* while maintaining the semantic essence and the expected answer. The generated are stored in the **ObfusQA** dataset (to be discussed in Section 2.1), which introduces obfuscations along three critical dimensions, as follows.

Named-Entity Indirection (NEI) or *Reasoning Through Indirect References*: The motivation behind introducing Named-Entity Indirection is to push LLMs towards deeper, more sophisticated reasoning by forcing them to infer relationships and entities from indirect or abstract cues. Rather than relying on explicit references, this approach challenges the model to connect disparate pieces of information through logical inferences. As shown in Fig.2, a simple question like *"Who invented the telephone?"* is transformed to *"Name the ingenious person who gifted us with the ability to converse audibly across long distance?"*. In this version, the model after named-entity (NE) extraction, must infer the *"inventor"* by connecting the concept of *"distant audible conversation"* with historical developments in communication technology. The model needs to deduct from the abstract idea of *distant audible conversation* to the *telephone* specifically. The question might then include additional references to related technologies—like the telegraph, wireless radios, and other communication innovations—which the model must logically connect to arrive at the correct answer. It is important to note that NEI involves reasoning not only about the entities within the question but also those expected in the answer. This process tests the model's ability to make nuanced connections and employs deeper inferencing (e.g., linking "telephone" to "ability to converse audibly across long distances") rather than relying on simple memorization of facts.

Distractor Indirection (DI) or *Actively Steering Toward Wrong Answers*: Distractor Indirection introduces plausible but incorrect alternatives to steer the model toward false choices deliberately. When a question is framed using both **indirect references** and **distractions**, it becomes significantly more obscure by introducing multiple layers of information that include convincing yet incorrect options. This approach tests how well the model can distinguish between similarly plausible answers and how effectively it can sift through distracting, but related, information. For example, in Figure 3, the question could be framed as: *"Name the ingenious person who gifted us with the ability to converse audibly across long distances, a groundbreaking achievement that took place in 1876, amidst competitors like Thomas Edison, Nikola Tesla, and others pioneering advancements in electrical communication."* We aim to introduce incorrect but plausible options,

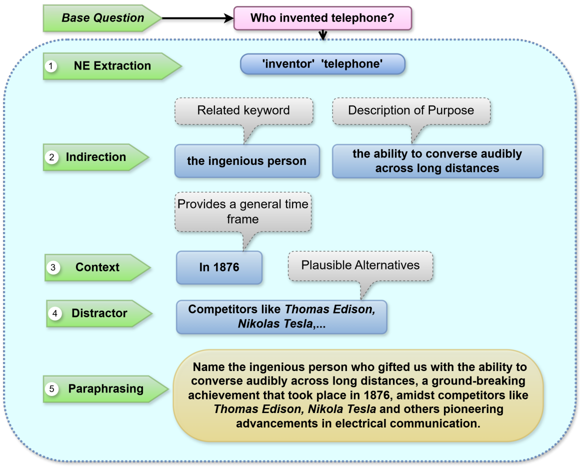

Base Question
Who invented telephone?
① NE Extraction
'inventor' 'telephone'
Related keyword
Description of Purpose
② Indirection
the ingenious person
the ability to converse audibly across long distances
Provides a general time frame
③ Context
In 1876
Plausible Alternatives
④ Distractor
Competitors like Thomas Edison, Nikolas Tesla,...
⑤ Paraphrasing
Name the ingenious person who gifted us with the ability to converse audibly across long distances, a ground-breaking achievement that took place in 1876, amidst competitors like Thomas Edison, Nikola Tesla and others pioneering advancements in electrical communication.

Figure 3: Distractor Indirection

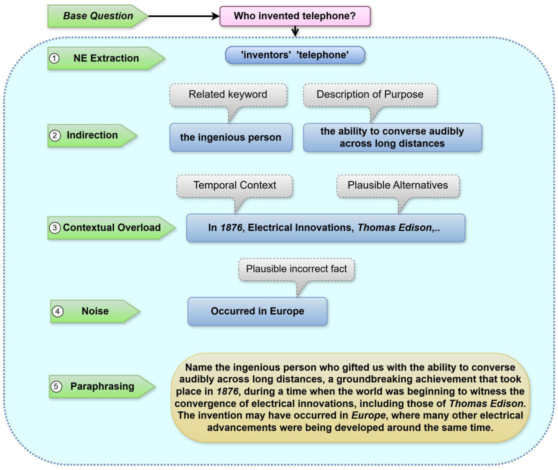

Base Question
Who invented telephone?
① NE Extraction
'inventors' 'telephone'
Related keyword
Description of Purpose
② Indirection
the ingenious person
the ability to converse audibly across long distances
Temporal Context
Plausible Alternatives
③ Contextual Overload
In 1876, Electrical Innovations, Thomas Edison,..
Plausible incorrect fact
④ Noise
Occurred in Europe
⑤ Paraphrasing
Name the ingenious person who gifted us with the ability to converse audibly across long distances, a groundbreaking achievement that took place in 1876, during a time when the world was beginning to witness the convergence of electrical innovations, including those of Thomas Edison. The invention may have occurred in Europe, where many other electrical advancements were being developed around the same time.

Figure 4: Contextual Overload

forcing the respondent to compare the actual inventor (Alexander Graham Bell) with well-known alternatives – "Thomas Edison" (renowned for electrical innovations) and "Nikola Tesla" (associated with pioneering electrical work). The presence of multiple figures in communication technology compels deeper reasoning, making the correct answer less immediately apparent. The correct answer is deliberately de-emphasized, while misleading alternatives seem equally viable, creating a choice dilemma.

Contextual Overloading (CO) or *Drowning the Core Question in Noise*: In the contextual overloading frame, we amplify the cognitive load by strategically incorporating **red herring facts** [1] within a heavily overloaded contextual ambiance. In contrast to DI, CO does not steer toward wrong answers but rather buries the correct one under a heavily overloaded contextual environment. This method adds layers of potentially misleading yet related information and adds noises that demand careful reasoning. For example, Figure 4 illustrates how a simple base question can be transformed into a heavily contextual one: *"Name the ingenious person who gifted us with the ability to converse audibly across long distances, a groundbreaking achievement that took place in 1876, during a time when the world was beginning to witness the convergence of electrical innovations, including those of Thomas Edison. The invention may have occurred in Europe, where many other electrical advancements were developed around the same time."* Contextual overload achieves the following: *Inject irrelevant but true information* ("the convergence of electrical innovations", "electrical advancements in Europe"): forcing the respondent to waste cognitive effort sorting signals from noise. *Adds excessive but factual complexity* (mentioning Thomas Edison, framing the invention within a global technological shift); making it harder to extract the essential clue. While all three introduce complexity, they do so in distinct ways: **NEI** *by requiring the respondent to uncover the intended entity through abstract reasoning and indirect linguistic cues*, **DI** *by subtly guiding the respondent toward incorrect answers*, and **CO** *by obscuring the core question with extraneous details*.

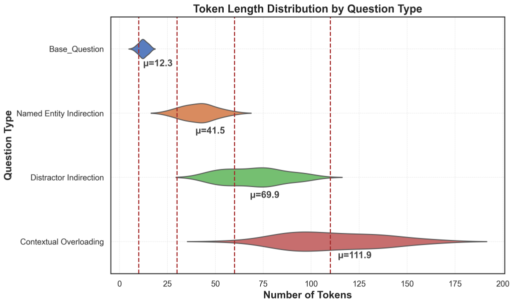


Figure 5: Token length distributions across ObfusQA types. Complexity increases with obfuscation; red dashed lines denote mean (μ) length.

2.1 Dataset Creation: ObfusQA

We source our base questions primarily from the TriviaQA dataset (Joshi et al., 2017), along with a small subset drawn from a government examination preparation website, GKToday [2]. Following this, we utilize Gemini 2.0 Flash (Deepmind, 2024; Team et al., 2024) LLM to generate obfuscations from the base questions using our designed

---

[1] as defined in the *Cambridge Dictionary*, red herring is a "fact, idea, or subject that takes people's attention away from the central point being considered"

[2] `https://www.gktoday.in/`

algorithm through well-designed prompts (cf. Appendix A.1). All generations were conducted at a temperature of 0.75. After obfuscating these base questions, we obtain a total of 1024 questions that include both the base questions and their three obfuscations, verified and corrected through a *human-in-the-loop* process by annotators (cf. Appendix A.3.1). Figure 5 shows average token length distributions for each variant. We also achieve an inter-annotator agreement score of 86.2% as measured by Cohen's κ. Human annotation safeguards dataset integrity by overseeing each transformation, focusing on **(i) Ground Truth Preservation** to ensure obfuscation avoids ambiguity or multiple valid answers, and validating that **(ii) Obfuscation Increases Cognitive Load, Not Ambiguity** by confirming transformations heighten reasoning difficulty without compromising semantic clarity or factual correctness.

Model	**Question Type**	**Zero-Shot (%)**	**Few-Shot (%)**	**CoT (%)**
GPT-4o	Base	67.97	73.05	**84.38**
	NEI	46.48	53.91	**55.86**
	DI	25.78	**33.59**	32.42
	CO	30.08	37.89	**38.67**
Claude 3.5 Sonnet	Base	**78.91**	74.61	75.00
	NEI	36.72	41.41	**54.30**
	DI	26.17	27.73	**38.28**
	CO	35.16	32.81	**39.45**

Table 1: EM (%) accuracy of GPT-4o and Claude 3.5 Sonnet across obfuscation types under zero-shot, few-shot, and CoT prompting. GPT-4o benefits most from CoT prompting; Claude 3.5 excels in zero-shot.

Model	**Question Type**	**Zero-Shot (%)**	**Few-Shot (%)**	**CoT (%)**
DeepSeek R1	Base	71.25	76.80	**82.15**
	NEI	42.30	48.67	**58.92**
	DI	28.45	31.20	**40.78**
	CO	33.90	39.15	**42.33**
o3-mini	Base	69.80	**79.60**	72.45
	NEI	40.85	**52.75**	45.20
	DI	27.60	30.85	**36.90**
	CO	32.40	**40.25**	36.70

Table 2: EM (%) accuracy of DeepSeek R1 and o3-mini across obfuscation types (100 samples). Both models benefit from CoT prompting; DeepSeek R1 showing stronger CoT gains on obfuscated inputs.

3 Evaluation Setup

We benchmarked seven SoTA LLMs on ObfusQA: GPT-4o (Hurst et al., 2024), GPT-4o mini (OpenAI, 2024a), LLaMA 3.3 70b (Dubey et al., 2024), Gemini 2.0 Flash, Claude 3.5 Sonnet (Anthropic, 2024); we include two strong *reasoning* models: DeepSeek R1 (Bi et al., 2024), and GPT o3-mini (OpenAI, 2024b). Model settings and responses are detailed in (cf. Appendix A.6, A.8). We evaluated performance using **Exact Match (EM) accuracy**, which measures the percentage of normalized outputs exactly matching a normalized ground truth answer (see Appendix.A.5).

3.1 Results and Analysis

We evaluate the LLMs across *zero-shot*, *few-shot*, and *chain-of-thought (CoT)* prompting strategies. While models perform well on base questions, their accuracy drops notably on obfuscated prompts, especially for DI and CO variants. We also evaluated Gemini 2.0 Flash (`gemini-2.0-flash-exp`) to assess its ability to answer its own obfuscated questions. Despite generating the queries, the model failed to answer most of the transformed queries correctly, highlighting LLMs' limited "self-awareness" (see Table 3 in Appendix A.6). Table 1, 2 shows a comparative study for four models with **bold** entries indicating row-wise best performance.. Due to budget constraints, 100 samples were used to evaluate reasoning-oriented models, DeepSeek R1 and GPT o3-mini; which showed reduced performance on obfuscated inputs, highlighting a potential vulnerability (cf. Appendix A.8). We also present sample queries where models demonstrate impressive capabilities in answering these obfuscated queries (cf. Appendix A.9). A detailed benchmarking analysis of the other listed models is provided in (cf. Appendix A.6).

Intrinsic Analyses. To further probe model behavior, we conducted three targeted internal analyses using LLaMA 3.1 8B (Dubey et al., 2024) and Mistral 7B v0.1 (Jiang et al., 2023), constrained by GPU availability. As these models perform poorly on ObfusQA, we omit their full results, but examine: **(i) Intrinsic Confidence**, which reveals a decline in self-assessed certainty across obfuscation types; **(ii) Memorization**, via membership inference, confirming the models cannot retrieve obfuscated answers from pre-training; and **(iii) Layer-wise Norm Drop Analysis**, showing early representational compression on obfuscated inputs. We also observe that shorter DI queries led to worse performance than longer CO ones (cf. Tables 1, 3), indicating that surface complexity alone does not explain model failure.

4 Conclusion

Our `ObfusQAte` shows novel light on LLMs robustness exhibited by LLMs' impairment caused by obfuscated versions of the base questions, hence opening a new avenue for interesting future work.

Limitations and Future Work

While our current study focuses on a single QA dataset composed of factual questions in English, it provides a foundational step toward broader generalization. In future work, we aim to extend our dataset to include *multilingual* data, particularly focusing on *low-resource* languages to enhance inclusivity. To better capture the diversity of real-world QA tasks, we plan to incorporate additional categories such as mathematical reasoning, comprehension-based tasks, and translation challenges. Our future work will additionally implement obfuscation-based techniques in white-box settings to evaluate these systems in greater detail, promoting the development of equitable and highly robust AI systems.

Ethics Statement

All human annotators were compensated fairly and commensurate with their contributions to ensure that their time and efforts were respected and valued. The recruitment process followed ethical standards, and all participants provided informed consent regarding the use of their annotations. The paid models employed in our study were accessed strictly via valid subscriptions, in accordance with the terms of service provided by the respective providers.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Anthropic. 2024. Claude 3.5 Sonnet.

Amos Azaria and Tom Mitchell. 2023. The internal state of an llm knows when it's lying. *arXiv preprint arXiv:2304.13734*.

Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, et al. 2024. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2024. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45.

Google Deepmind. 2024. Introducing Gemini 2.0: our new AI model for the agentic era.

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

Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. 2022. True: Re-evaluating factual consistency evaluation. *arXiv preprint arXiv:2204.04991*.

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

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Ehsan Kamalloo, Shivani Upadhyay, and Jimmy Lin. 2024. Towards robust qa evaluation via open llms. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2811–2816.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2021. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*.

Potsawee Manakul, Adian Liusie, and Mark JF Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *arXiv preprint arXiv:2303.08896*.

Seyedreza Mohseni, Seyedali Mohammadi, Deepa Tilwani, Yash Saxena, Gerald Ketu Ndawula, Sriram Vema, Edward Raff, and Manas Gaur. 2025. Can llms obfuscate code? a systematic analysis of large language models into assembly code obfuscation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 24893–24901.

OpenAI. 2024a. GPT-4o mini: advancing cost-efficient intelligence.

OpenAI. 2024b. OpenAI o3-mini.

Constantinos Patsakis, Fran Casino, and Nikolaos Lykousas. 2024. Assessing llms in malicious code deobfuscation of real-world malware campaigns. *Expert Systems with Applications*, 256:124912.

Pouya Pezeshkpour. 2023. Measuring and modifying factual knowledge in large language models. In *2023 International Conference on Machine Learning and Applications (ICMLA)*, pages 831–838. IEEE.

Adrian Swindle, Derrick McNealy, Giri Krishnan, and Ramyaa Ramyaa. 2024. Evaluation of large language models on code obfuscation (student abstract). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 23664–23666.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Cunxiang Wang, Sirui Cheng, Zhikun Xu, Bowen Ding, Yidong Wang, and Yue Zhang. 2023. Evaluating open question answering evaluation. *CoRR*.

Jingyang Zhang, Jingwei Sun, Eric Yeats, Yang Ouyang, Martin Kuo, Jianyi Zhang, Hao Frank Yang, and Hai Li. 2025. Min-k%++: Improved baseline for pre-training data detection from large language models. In *The Thirteenth International Conference on Learning Representations*.

Yuchen Zhuang, Yue Yu, Kuan Wang, Haotian Sun, and Chao Zhang. 2023. Toolqa: A dataset for llm question answering with external tools. In *Advances in Neural Information Processing Systems*, volume 36, pages 50117–50143. Curran Associates, Inc.

A Appendix

A.1 ObfusQAte Prompts

Named-Entity Indirection

1. Identify Named-Entities: First, find all the named entities in the question. Named entities can include people, organizations, locations, dates, etc.

2. Apply indirection to the named entities

- Using synonyms for the named entities.
- Replacing named entities with pronouns.
- Using definitions or descriptions instead of the actual named entity.
- Introducing family relations, if applicable, or other indirect forms.

3. Reframe the Question: After transforming the named entities, rephrase the question so that it becomes more difficult to answer, while still maintaining its original meaning.

4. Review the Final Question: Ensure that the final question is significantly more challenging and requires more reasoning or external knowledge to answer correctly.

Distractor Indirection

1. Identify Named Entities: First, find all the named entities in the question. Named entities can include people, organizations, locations, dates, etc.

2.Apply indirection to the named entities, this can involve:

- Using synonyms for the named entities.
- Replacing named entities with pronouns.
- Using definitions or descriptions instead of the actual named entity.
- Introducing family relations, if applicable or apply other indirect forms.

3. Add Very Little Context: Add a tiny bit of context (one short phrase or clause) to the question without giving away the answer.

4. Add Distractor: Introduce distractors to make the question more challenging:

- Similar Concepts: Add semantically or topically related concepts that might confuse the answer (e.g. “Was Hamlet published the same year as Romeo and Juliet?”).
- Plausible Alternatives: Offer incorrect but reasonable alternatives that could distract the user (e.g. “Was Hamlet written by Marlowe, Shakespeare or Jonson?”).

5. Rephrase the Question:
Change the structure of the question so that it forces the user to think harder or search for answers in various places.

Contextual Overload

1. Identify Named Entities: First, find all the named entities in the question. Named entities can include people, organizations, locations, dates, etc.

2. Apply indirection to the named entities, this can involve:

- Using synonyms for the named entities.
- Replacing named entities with pronouns.
- Using definitions or descriptions instead of the actual named entity.
- Introducing family relations, if applicable, or apply other indirect forms.

3. Contextual Overload:

- Irrelevant Details: Add information that is not directly relevant to the answer but might confuse or mislead the respondent.
- Red Herring Facts: Include incorrect facts that might seem plausible.
- Temporal/Spatial Context: Frame the question within a specific period or geographic location, requiring a more specific answer.

4. Rephrase the Question: Change the structure of the question so that it forces the user to think harder or search for answers in various places.

A.1.1 Prompting strategies

Chain-of-Thought

You are an expert at answering complex and obfuscated objective-type questions.

Think step by step to deconstruct the question, identify the core information needed, and derive the correct answer.

Finally, state the answer clearly and concisely.

Follow these steps:

1. Read the entire question carefully, even if it includes extra indirection or distractors.

2. Isolate the core query by stripping away any added layers of indirection, irrelevant details,or red herrings.

3. Reverse any transformations to recover the original meaning of the question.

4. Apply logical reasoning and your domain knowledge to determine the correct answer.

5. Finally, output only the concise final answer without showing any internal reasoning or extra text.

Few-shot Prompt

You are an assistant that answers only with the objective answer. Do not include any additional information. When responding, carefully review the examples that include both the base question and the modifications, and use these to infer the intended meaning of the asked question and deliver answer:

NAMED-ENTITY-INDIRECTION = """

Examples:

Example 1: base-Question: <INSERT YOUR EXAMPLE 1 HERE> Answer: <GROUND TRUTH ANSWER 1>

Example 2: base-Question: <INSERT YOUR EXAMPLE 2 HERE> Answer: <GROUND TRUTH ANSWER 2> """

DISTRACTION INDIRECTION = """

Examples:

Example 1: base-Question: <INSERT YOUR EXAMPLE 1 HERE> Answer: <GROUND TRUTH ANSWER 1>

Example 2: base-Question: <INSERT YOUR EXAMPLE 2 HERE> Answer: <GROUND TRUTH ANSWER 2> """

CONTEXTUAL-OVERLOAD = """

Examples:

Example 1: base-Question: <INSERT YOUR EXAMPLE 1 HERE> Answer: <GROUND TRUTH ANSWER 1>

Example 2: base-Question: <INSERT YOUR EXAMPLE 2 HERE> Answer: <GROUND TRUTH ANSWER 2> """

A.2 Dataset Distribution

The ObfusQA dataset comprises 256 unique base factual questions, each transformed into three progressively challenging obfuscated variants: *Named-Entity Indirection (NEI)*, *Distractor Indirection (DI)*, and *Contextual Overload (CO)*, resulting in a total of 1024 samples. Each variant preserves the semantic intent of the original while introducing distinct cognitive challenges. The dataset is evenly distributed across the four types (25% each), enabling controlled experiments across difficulty levels. An analysis of average token lengths confirms increasing verbosity and complexity: base questions average 11.6 tokens, NEI variants 41.9, DI 62.3, and CO variants 116.1 tokens (cf. Figure 5).

A.3 Annotation

A.3.1 Annotator's Details

We engaged a team of seven undergraduate students from an Indian university who are part of an AI research lab. They are well-trained and have relevant course experience to manually annotate the generated questions. Their role involved interpreting and analyzing the questions, reasoning through them, and making subtle edits to tackle hallucinations, which generally occur when the automated system gradually deviate from the original intent or meaning of the base question. This process ensured robustness throughout our evaluation. After thorough human annotation, we get **ObfusQA** dataset.

A.3.2 Annotation Example

We illustrate our annotation process, starting with the **base question**:

> *"What is the capital of Australia?"*

From this base question, our automated system (Google's Gemini 2.0 Flash) generates three obfuscated variants:

(i) Named-Entity Indirection (NEI): *"Which urban center, situated within the Commonwealth realm that witnessed the dawning of the new millennium with a quadrennial celebration of athletic prowess, serves as the locus of governance for a continent-spanning island nation, characterized by its distinctive fauna and a political system shaped by the Westminster tradition?"*

(ii) Distraction Indirection (DI): *"Amidst ongoing debates about regional development, and considering the political and administrative heart of the land Down Under, is the principal federal city—which we'll call X—more populous than the metropolis that annually celebrates equestrian prowess, or does it rival the city that serves as the harbour and is also known for an architectural marvel in terms of size? By what name, then, is this city designated on official maps?"*

(iii) Contextual Overload (CO): *"Amidst the echoes of the Great Emu War and the ongoing debate over the Pavlova's true origins, can you identify the city, nestled within the Australian Capital Territory, that serves as the seat of the Governor-General, currently held by the King's representative, and where the Old Parliament House, a relic of the era before self-government was fully realized and a structure often mistaken for the primary legislative building due to its prominent position near Lake Burley Griffin, is located, remembering that the nation's highest court is actually located elsewhere? Furthermore, disregard the spurious claims that Sydney or Melbourne hold this distinction, as they are merely the most populous and historically significant metropolises, respectively."*

Ground Truth: *"Canberra"*

The annotation process involves the following steps:

1. **Synthetic Generation:** Each obfuscation question is produced by prompting the LLM with instructions to transform the base question into NEI, DI, and CO formats.

2. **Human Review:** The annotators read the generated questions carefully and checked for: *(i) Factual Consistency:* Does the obfuscated question still refer to the correct entity (i.e., capital of Australia) and avoid contradictory statements?

   (ii) Semantic Faithfulness: Does the question still ask for the same information as the base question?

   (iii) Hallucinations: Are there any introduced inaccuracies (e.g., attributing the capital city to the wrong country)?

3. **Edits and Corrections:** The annotators make subtle wording adjustments to remove or fix any detected hallucinations while preserving the intended obfuscation style to prevent any semantic "drift" bias.

4. **Final Verification:** Each revised question is confirmed to be semantically aligned with the original base query, ensuring that all four variants (Base, NEI, DI, CO) ask for the same underlying fact.

Through this process, we obtain the final, human-verified version of the obfuscated questions, each pointing to the same *ground truth* answer, **Canberra**.

A.3.3 Inter-Annotator Reliability

To measure inter-annotator reliability, we compute *Cohen's Kappa* (κ) between the two annotators for entire the dataset across all the obfuscated variants. Cohen's Kappa accounts for agreement occurring by chance, making it more robust than raw accuracy in evaluating categorical labeling tasks.

We compute κ as:

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

where p_o is the observed agreement and p_e is the expected agreement by chance.

In our case, annotations were compared at the sentence level for semantic equivalence and factual alignment. The resulting Kappa score was $\kappa = 0.862$, indicating *strong agreement* according to standard interpretation thresholds. Disagreements were resolved via adjudication to ensure high-quality supervision for evaluation.

A.4 Related Work

Although obfuscation has been discussed in recent LLM studies, very few have systematically assessed its impact on LLM performance.(Mohseni et al., 2025) investigates LLMs' potential to generate obfuscated assembly code, developing the MetamorphASM benchmark to evaluate this capability across various models. (Swindle et al., 2024) explore how effectively LLMs can detect and analyze obfuscated C++ code, with results highlighting that LLMs struggle particularly with layered and insertion-based obfuscations, revealing a significant gap in their code reasoning abilities. (Patsakis et al., 2024) investigates LLMs' ability to de-obfuscate malicious PowerShell scripts in real-world malware like Emotet, showing promising results even without specialized training.

A.5 Evaluation Metric

We evaluate our system using an **Exact Match (EM) Accuracy** metric, which quantifies the percentage of samples for which the normalized model answer exactly matches one of the normalized ground truth answers. For example, given the ground truth "Pope", a generated answer like "Pope, and his relative" would be considered incorrect due to the exact match requirement; we tabulate these examples in the Appendix. Mathematically, let N be the total number of samples. For each sample i, let $\mathcal{Y}^{(i)}$ be the set of normalized ground truth answers and $\hat{y}^{(i)}$ the normalized answer produced by the model. We define an indicator function $I^{(i)}$ such that $I^{(i)} = 1$ if $\hat{y}^{(i)} \in \mathcal{Y}^{(i)}$ and $I^{(i)} = 0$ otherwise. The overall EM accuracy is then computed as:

$$\text{EM Accuracy} = \frac{1}{N} \sum_{i=1}^{N} I^{(i)} \times 100\%. \qquad (1)$$

For each of the obfuscation variant, the same methodological formulation is applied. The normalization process standardizes case, removes punctuation, and ensures whitespace uniformity, making comparisons resilient to peripheral textual variations.

A.6 Benchmarking Study on other LLMs

To evaluate the robustness of different large language models (LLMs) against query obfuscation, we conducted a comparative benchmarking study on LLaMA 3.3 70B (`llama-3.3-70b-versatile`), GPT-4o mini, and Gemini Flash 2.0, (cf. Appendix: Table 3). We analyzed their performance across three prompting strategies: Zero-Shot, Few-Shot, and Chain-of-Thought (CoT) prompting across all Query types. The specific prompt templates used in this study can be found in (cf. Appendix A.1.1).

A.7 Intrinsic Analysis

A.7.1 Understanding Intrinsic Confidence

In order to internally assess what these systems truly understand, as well as to verify the validity of their self-assessed claims and predict which questions they are likely to answer correctly, we evaluate the $P(\text{IK})$ scores for our obfuscations. Here, $P(\text{IK})$ denotes the probability that a model assigns to the phrase "I know", i.e. the proposition that it will answer a given question correctly when samples are generated at unit temperature (Kadavath et al., 2022). We performed this analysis by probing the model using ObfusQA, performing a token-level assessment of the $P(\text{IK})$ scores. During inference, the model performs poorly on ObfusQA due to its limited knowledge and smaller parameter size. Our results show a consistent decline in $P(\text{IK})$ across obfuscations, particularly for DI and CO types, indicating reduced internal confidence and comprehension are adversely affected by the perturbations (see Table 4).

A.7.2 Memorization

To determine whether our obfuscation queries are incorporated during the model's pre-training phase, we apply a Membership Inference Attack (MIA) on the LlaMA 3.1 8b and Mistral 7b v0.1 models using the Min-K%++ method (Zhang et al., 2025) to detect pre-training data in these LLMs. In this framework, the parameter K specifies the percentage of token sequences with the lowest scores that are used to compute the final score. (Figure 6) and (Figure 7) plot the AUROC performance across the parameter K for our question categories for both these models. High AUROC values confirm Min-K%++'s reliability in detecting pre-training data.

Model	Question Type	Zero-Shot (%)	Few-Shot (%)	CoT (%)
LlaMA 3.3 70b	Base	75.69	**77.34**	74.61
	Named-Entity Indirection	**43.14**	40.23	41.41
	Distractor Indirection	29.80	**30.08**	**30.08**
	Contextual Overload	32.55	32.81	**35.55**
GPT 4o mini	Base	57.81	57.42	**61.72**
	Named-Entity Indirection	31.64	32.42	**36.72**
	Distractor Indirection	23.05	24.22	**26.17**
	Contextual Overload	23.44	26.95	**30.08**
Gemini Flash 2.0	Base	72.27	76.95	**78.91**
	Named-Entity Indirection	44.92	48.44	**50.78**
	Distractor Indirection	32.03	**36.72**	33.59
	Contextual Overload	36.72	**37.50**	35.55

Table 3: Evaluation of GPT 4o and LlaMA 3.3 70b, GPT 4o mini and Gemini Flash 2.0 on different obfuscation types, under zero-shot, few-shot, and chain-of-thought (CoT) prompt conditioning. Each value represents the EM accuracy (%), where higher values indicate better performance. **Bold-faced** entries highlight the best accuracy within each row

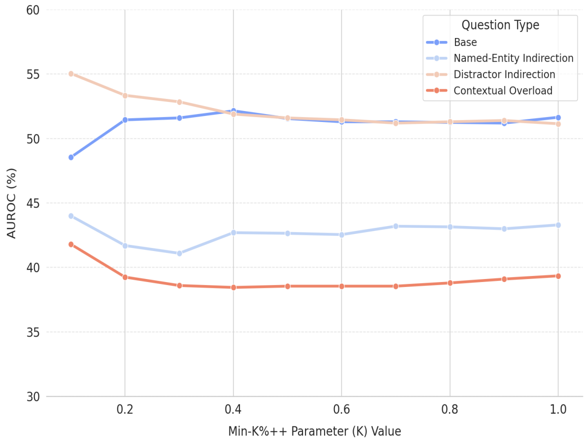


Figure 6: AUROC Performance Across Min-K%++ Parameter K on LlaMA 3.1 8b

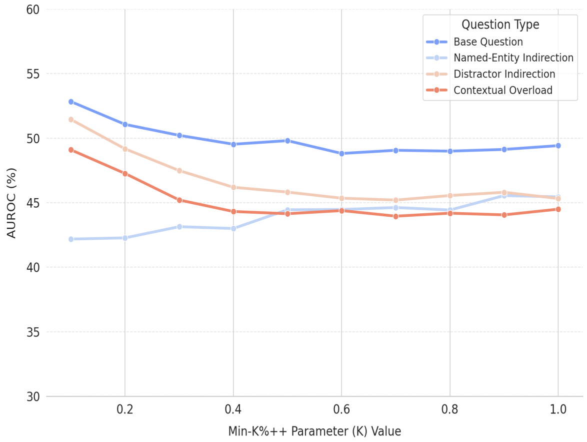


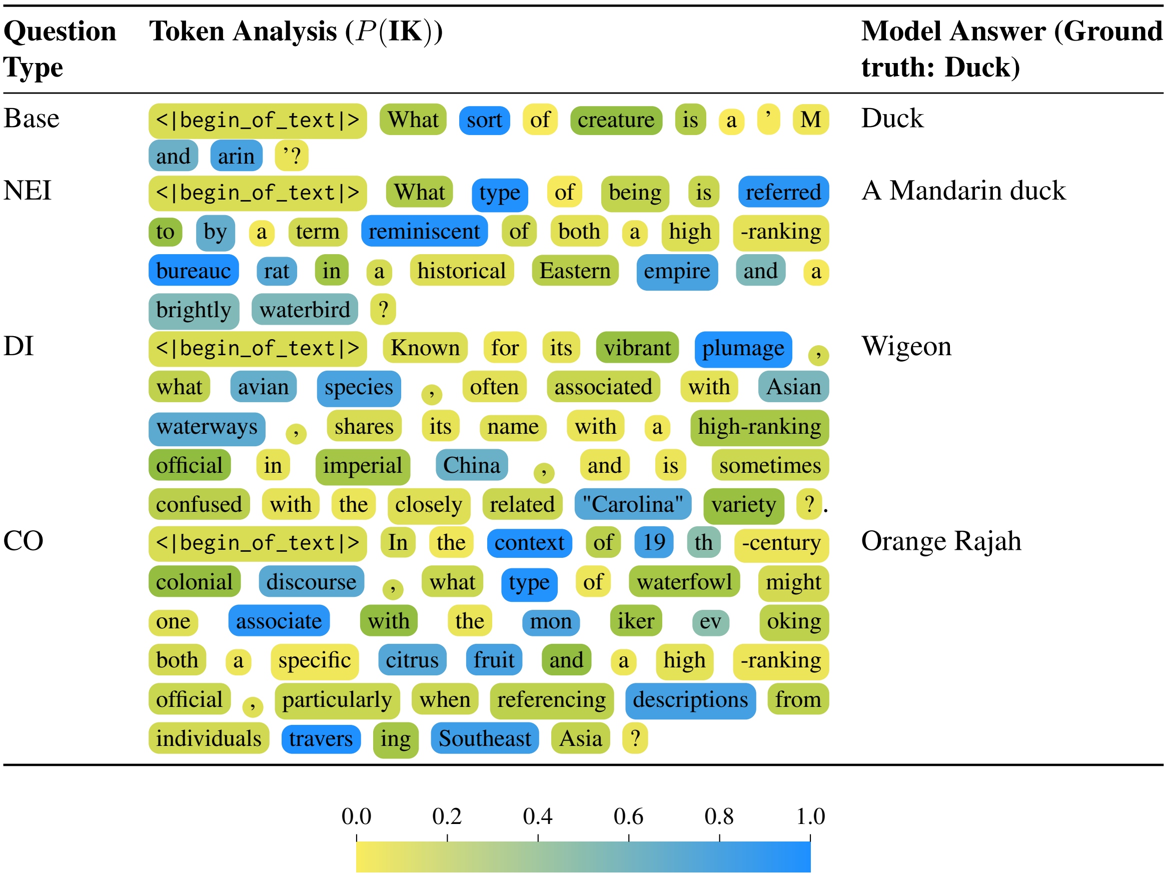
Figure 7: AUROC Performance Across Min-K%++ Parameter K on Mistral 7b v0.1

LlaMA 3.1 8b. Base questions yield the highest AUROC (47–55%), showing a brief initial dip before increasing. Distractor Indirection starts relatively high, then gradually decreases and converges near the Base Question curve. Contextual Overload remains the lowest (38–44%), while Named Entity Indirection begins around 43–45% and fluctuates slightly before ending near the others. However, DI obfuscations are detected at levels comparable to base questions. While QA performance is lower, detection relies on statistical alignment with training data, requiring further investigation.

Mistral 7b v0.1. Base questions range from 53% AUROC at $K = 0.0$ down to about 50% at $K = 1.0$. Distractor Indirection similarly decreases (52% to 45%), while Named Entity Indirection rises from 42% to about 45%. Contextual Overload starts at 48%, gradually declines, and stabilizes near $k = 0.6$. Here, despite some fluctuations, the overall results indicate that simpler query forms more clearly reveal membership signals, while obfuscation reduces pre-training data detection.

A.7.3 Layer-wise Norm Drop Analysis

Figure 8 illustrates the layer-wise hidden-state norm profiles, $\bar{n}_\ell$ for our base question and its obfuscated variants about the largest planet in the solar system. At each transformer layer ℓ, every token t in the input sequence is represented by a



<table>
<tr><th>Question Type</th><th>Token Analysis (P(IK))</th><th>Model Answer (Ground truth: Duck)</th></tr>
<tr><td>Base</td><td><|begin_of_text|> What sort of creature is a ’ M and arin ’?</td><td>Duck</td></tr>
<tr><td>NEI</td><td><|begin_of_text|> What type of being is referred to by a term reminiscent of both a high -ranking bureauc rat in a historical Eastern empire and a brightly waterbird ?</td><td>A Mandarin duck</td></tr>
<tr><td>DI</td><td><|begin_of_text|> Known for its vibrant plumage , what avian species , often associated with Asian waterways , shares its name with a high-ranking official in imperial China , and is sometimes confused with the closely related "Carolina" variety ? .</td><td>Wigeon</td></tr>
<tr><td>CO</td><td><|begin_of_text|> In the context of 19 th -century colonial discourse , what type of waterfowl might one associate with the mon iker ev oking both a specific citrus fruit and a high -ranking official , particularly when referencing descriptions from individuals travers ing Southeast Asia ?</td><td>Orange Rajah</td></tr>
</table>

Table 4: Representation of token-wise $P(\text{IK})$ scores for various obfuscated query variants used to probe the LLaMA 3.1 8b model. The Base query, with its straightforward phrasing, concentrates activations on key tokens, leading to the correct answer, “Duck”. In the NEI variant, despite added descriptive elements, the essential cues remain sufficiently prominent to yield an answer close to the ground truth ("A Mandarin duck"). However, the DI and CO queries introduce further obfuscation through extra contextual and indirect references, which shift the model’s focus and alter the token activation pattern, ultimately resulting in incorrect responses ("Wigeon" and "Orange Rajah").

high-dimensional hidden state vector $\mathbf{h}_{\ell,t}$. The ℓ_2 norm of this vector, $\|\mathbf{h}_{\ell,t}\|_2$, serves as a proxy for the token’s “activation energy” or semantic richness at that layer. To track the flow of information through the network, we compute the average norm across all T tokens at each layer:

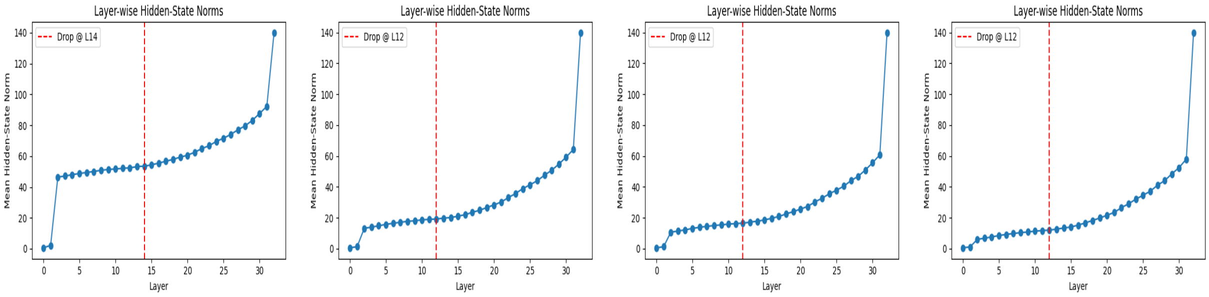
$$\bar{n}_\ell = \frac{1}{T}\sum_{t=1}^{T}\left\|\mathbf{h}_{\ell,t}\right\|_2$$

This layer-wise profile $\bar{n}_\ell$ reveals how the model processes and transforms information. Rising norms across layers typically indicate feature amplification, where token-level representations gain semantic detail; while sudden drops in $\bar{n}_\ell$ signal compression bottlenecks. These bottlenecks correspond to stages where the model collapses distributed features into more abstract, high-level representations.

Observation: We observe that, the base question exhibits a relatively late drop in hidden-state norms (at Layer 14), whereas all perturbed variants (NEI, DI, CO) show an earlier drop (at Layer 12). This consistent shift toward earlier compression suggests that injected linguistic complexity or semantic distraction leads the model to prematurely reduce representational richness. Such early bottlenecks may truncate deeper semantic processing and negatively impact the model’s ability to reason through nuanced input.

A.8 Qualitative Examples of Model Behavior

We present representative examples to illustrate model behavior under different obfuscation types. Correct (green) and incorrect (red) answers highlight successful reasoning and failure cases, respectively.



(a) Base Question **(b)** NEI **(c)** DI **(d)** CO

Figure 8: Comparison of model responses to four question variants about the largest planet in the solar system. **(a)** *Base Question*: "What is the largest planet in our solar system?" **(b)** *NEI (Named Entity Indirection)*: "What celestial body, renowned as the most expansive in the star system we call home, would its diameter compare to when measured against the circumference of the third rock from the sun, assuming its equatorial dimension?" **(c)** *DI (Distractor Indirection)*: "Given its gaseous composition and significant influence on neighboring celestial bodies, which orbiter, often referred to as the "king" of our local star's family, surpasses all others in diameter, and is it larger than, say, Neptune or Uranus?" **(d)** *CO (Contextual Overload)*: "Amidst the celestial bodies influenced by the radiant star at the heart of our local cosmic neighborhood, and considering the understanding of astronomical entities held by ancient Greco-Roman civilizations, which gas giant, whose namesake ruled over the gods, surpasses all others in volumetric magnitude, dwarfing even the terrestrial sphere that cradles the city known for the Colosseum, and also has a storm that is larger than three times the area of Earth? Also, is it true that the planet is mostly made of solid rock?" **Ground truth:** Jupiter.

Example 1: Ground Truth — **Anita Loos**

Base Question: Who wrote the novel *Gentlemen Prefer Blondes*?

Claude 3.5 Sonnet Answer: Anita Loos

Deepseek R1 Answer: Anita Loos

GPT o3-mini Answer: Anita Loos

GPT 4o Answer: Anita Loos

NEI: Who authored the widely read novel concerning the experiences of a blonde woman, a work that subsequently served as the inspiration for a renowned musical production, and whose author is also related, through her granddaughter, to the creator of a highly popular series of books chronicling the life of a young wizard?

Claude 3.5 Sonnet Answer: Lauren Weisberger

Deepseek R1 Answer: Dodie Smith

GPT o3-mini Answer: Gregory Maguire

GPT 4o Answer: Anita Loos

Example 1: **Anita Loos** (continued)

DI: Within the vibrant cultural landscape of the Jazz Age, which author, whose creative spark is said to have been ignited by anecdotes circulating within the opulent settings of Manhattan's grand hotels, wrote the satirical narrative focusing on the exploits of two captivating women known for their pursuit of wealth and advantageous marriages, a work frequently misattributed to the cousin of Anita Loos or a contemporary writer such as Elinor Glyn?
Claude 3.5 Sonnet Answer: (Anita Loos)

Deepseek R1 Answer: Anita Loos

GPT o3-mini Answer: Edith Wharton

GPT 4o Answer: Edith Wharton

CO: During the intellectually and artistically fertile period of the Roaring Twenties, characterized by the iconic flapper fashion and the infectious rhythms of jazz music, a certain comedic novel emerged, encapsulating the spirit of this dynamic era.

Example 1: **Anita Loos** (continued)

The granddaughter of the author celebrated for a compilation of poems frequently recited at Vassar College penned a narrative centered on the adventures of a blonde woman. This author, whose sibling was a distinguished academic specializing in ancient Greek literature, crafted a work that later achieved acclaim as a celebrated musical. Could you identify the individual responsible for writing this narrative, a story depicting the escapades of a protagonist with fair hair, often perceived as naive, as she navigates the intricate social dynamics of a world populated by immense fortunes, aristocratic titles, and transatlantic voyages, while acknowledging that her contemporary, Anita Loos, was also a significant literary figure of that time?

Claude 3.5 Sonnet Answer: Edith Wharton

Deepseek R1 Answer: Lorelei Lee

GPT o3-mini Answer: Anita Loos

GPT 4o Answer: Evelyn Waugh

Example 2: Ground Truth — **Walter**

Base Question: What is Bruce Willis' real first name?

Claude 3.5 Sonnet Answer: Walter

Deepseek R1 Answer: Walter

GPT o3-mini Answer: Walter

GPT 4o Answer: Walter

NEI: The actor who played John McClane had one name and the other part of the name sounds similar to `Walter`. What is his first name?

Claude 3.5 Sonnet Answer: Walter

Deepseek R1 Answer: Walter

GPT o3-mini Answer: Bruce

GPT 4o Answer: Walter

Example 2: **Walter** (continued)

DI: An actor renowned for portraying resilient, often wisecracking, characters in action films chose a stage name that resonated with a strong, assertive image. Before achieving global fame, this performer adopted a first name that diverges from the more common appellation he was given at birth. Knowing that his career took off in the 1980s and that he has German ancestry, what was his birth first name?

Claude 3.5 Sonnet Answer: Bruce

Deepseek R1 Answer: Walter

GPT o3-mini Answer: Bruce

GPT 4o Answer: Bruce

CO: While he adopted a stage name synonymous with Hollywood action stardom, the actor who brought to life iconic characters known for their grit and resilience entered the world bearing a different designation. This individual, whose surname suggests a determined disposition, hails from a family with roots extending to post-war Germany, a nation undergoing a period of significant cultural redefinition in the mid-20th century. Though he embraced a persona associated with American action heroes, his given name actually shares etymological roots with a term denoting a military commander. Was this birth name more commonly associated with artistic figures, biblical figures, or a character in World War I? Was it Arnold, Walter, or Bruno?

Claude 3.5 Sonnet Answer: Sylvester

Deepseek R1 Answer: Arnold

GPT o3-mini Answer: Arnold

GPT 4o Answer: Bruce

Example 3: Ground Truth — **United States / United States of America**

Base Question: In which country was Emilio Estevez born?

Claude 3.5 Sonnet Answer: United States

Deepseek R1 Answer: United States

GPT o3-mini Answer: United States

GPT 4o Answer: United States

NEI: In what nation, where the silver screen is celebrated and its sibling shares a patronym, did the offspring of the "Apocalypse Now" narrator first draw breath?

Claude 3.5 Sonnet Answer: Phillipines

Deepseek R1 Answer: United States

GPT o3-mini Answer: United States

GPT 4o Answer: United States

DI: Given the backdrop of his father's artistic journey, in what nation did Martin Sheen's eldest offspring first draw breath, noting that he shares a moniker with a saint?

Claude 3.5 Sonnet Answer: Spain

Deepseek R1 Answer: United States

GPT o3-mini Answer: United States

GPT 4o Answer: United States

CO: In the nation renowned for its Hollywood heart and where a distinguished family, including the son of Martin Sheen (recognized by a title echoing a medieval military rank), first breathed life, despite his father's deep connections to another land celebrated for its shamrocks and folklore?

Claude 3.5 Sonnet Answer: United States of America

Deepseek R1 Answer: United States

GPT o3-mini Answer: United States

GPT 4o Answer: United States

A.9 Example Questions

In this section, we present a collection of sample questions along with their corresponding answers produced by GPT-4o and Claude 3.5 Sonnet, the top-performing models on ObfusQA. We have predominantly selected questions that our scoring method has identified as correct. By presenting these curated examples, we try to showcase the adept answering abilities and current proficiency of these LLMs. (cf. Appendix: table: 5, 6, 7).

A.10 Hosting & Maintenance

Once the dataset is made public, we plan to host it on Hugging Face.

A.11 Intended Usage

The **ObfusQAte** framework and dataset are designed primarily for research and development purposes, with the goal of evaluating and improving large language models (LLMs) resilience to obfuscated or indirectly phrased queries. By systematically challenging these LLMs with varied levels of semantic, distractive, and contextual complexity, researchers can pinpoint vulnerabilities in current models and devise strategies (e.g., improved prompt conditioning, fine-tuning/alignment, or adversarial training) to enhance their factual consistency. Importantly, ObfusQAte also exposes instances where models rely on mere memorization of pre-trained data rather than genuine reasoning, thereby highlighting the need for approaches that foster true understanding. We discourage using this dataset to deliberately deceive or mislead end-users, as the intent is to foster robust, transparent AI systems that better serve real-world needs.

Table 5: **Example 1:** Sample Queries and Model Answers from **GPT-4o** and **Claude 3.5 Sonnet**

Query Type	**GPT-4o**	**Claude 3.5 Sonnet**
Base Question	*Query:* What is the largest ocean on Earth? ***Ground Truth:*** Pacific Ocean	*Query:* What is the chemical symbol for silver? ***Ground Truth:*** Ag
Named-Entity Indirection	*Query:* Which body of water, known as the one bordering the Asian and American continents, and also referred to as the one Ferdinand Magellan crossed, holds the title for greatest surface area among all the world's interconnected hydrosphere components? *Answer:* The Pacific Ocean	*Query:* What alphanumeric designator is assigned to the element whose monetary applications historically rivaled those of aurum...? *Answer:* Ag
Distraction Indirection	*Query:* Considering its vastness and the ring of fire that surrounds it, which watery expanse, often navigated by vessels crossing from the land of the rising sun to the Americas, reigns supreme in terms of surface area when compared to the Atlantic, Indian, Arctic, and Southern oceans? *Answer:* The Pacific Ocean	*Query:* Considering its vastness and the ring of fire that surrounds it, which watery expanse, often navigated by vessels crossing from the land of the rising sun to the Americas, reigns supreme in terms of surface area when compared to the Atlantic, Indian, Arctic, and Southern oceans? *Answer:* Ag
Contextual Overload	*Query:* During the administration of the 45th U.S. President, amidst debates about climate change and maritime boundaries, what body of water, often associated with tales of krakens and explorations by Magellan, holds the greatest surface area, exceeding that of the Atlantic and Indian combined, despite some cartographers disputing its northernmost reaches due to Arctic ice formations, mistakenly suggesting the Arctic Ocean's dominance in size? *Answer:* The Pacific Ocean	*Query:* During the administration of the 45th U.S. President, amidst debates about climate change and maritime boundaries, what body of water, often associated with tales of krakens and explorations by Magellan, holds the greatest surface area, exceeding that of the Atlantic and Indian combined, despite some cartographers disputing its northernmost reaches due to Arctic ice formations, mistakenly suggesting the Arctic Ocean's dominance in size? *Answer:* Ag

Table 6: **Example 2:** Sample Queries and Model Answers from **GPT-4o** and **Claude 3.5 Sonnet**

Query Type	**GPT-4o**	**Claude 3.5 Sonnet**
Base Question	*Query:* What is the smallest country in the world? ***Ground Truth:*** Vatican City.	*Query:* Who painted the ceiling of the Sistine Chapel? ***Ground Truth:*** Michelangelo
Named-Entity Indirection	*Query:* What is the designation given to the sovereign state with the smallest area, an entity whose leader governs from a location within the boundaries of the city known as the center of the Catholic Church, a religious institution led by the successor to Saint Peter? *Answer:* Vatican City	*Query:* Who was the artistic genius commissioned to adorn the Vatican's most sacred chamber, a space renowned for its depiction of biblical narratives across its uppermost surface? *Answer:* Michelangelo
Distraction Indirection	*Query:* Considering its unique relationship with the Holy See, which sovereign nation, possessing the least expanse of terrestrial territory when measured by conventional cartographical means, is globally recognized for its diminutive size, rivaling others such as Nauru or Tuvalu? *Answer:* Vatican City	*Query:* Considering his extensive Florentine background and influence from Donatello, who was commissioned by the same pontiff who also initiated the rebuilding of St. Peter's Basilica to decorate the palatial sacred space, and not Raphael, complete the frescoes? *Answer:* Michelangelo
Contextual Overload	*Query:* Considering the spectrum of political geography, from sprawling continental federations to ephemeral assertions of territorial autonomy, and acknowledging the unique status of polities whose authority transcends mere physical boundaries: Which internationally acknowledged sovereign entity, whose terrestrial footprint is eclipsed even by the gardens of the aforementioned pontifical enclave, represents the nadir of territorial extension among independent nation-states absent ongoing disputes over legitimacy, and falls demonstrably below the minimum spatial threshold required for county-level designation within that transatlantic republic founded upon ideals of representative self-governance, thus distinguishing it from insular micro-polities adrift within the ocean named for peacefulness, whose census rolls scarcely register four-figure population counts while also explicitly ignoring a known sea platform from WW2? *Answer:* Vatican City	*Query:* Amidst the artistic fervor of the Renaissance, and considering the era's patronage system where familial influence often dictated commissions, who was the individual, renowned for sculpting David and whose artistic journey was significantly shaped by his Florentine origins, responsible for the fresco adorning the apex of the papal sanctuary inaugurated by Sixtus IV, a space where cardinals convene to elect the spiritual leader of a global faith, even though some falsely attribute sections depicting the Genesis narrative to Raphael due to their shared period of prolific creation and similar mastery of perspective? *Answer:* Michelangelo Buonarroti

Table 7: **Example 3:** Sample Queries and Model Answers from **GPT-4o** and **Claude 3.5 Sonnet**

Query Type	**GPT-4o**	**Claude 3.5 Sonnet**
Base Question	*Query:* What is the name of the Earth's natural satellite? ***Ground Truth:*** The Moon	*Query:* What is the name of the process by which plants make their own food? ***Ground Truth:*** Photosynthesis
Named-Entity Indirection	*Query:* What is the appellation given to the celestial body, orbiting our pale blue dot, that ancient cultures revered as a deity of the night, and which Neil Armstrong visited? *Answer:* The Moon	*Query:* What is the appellation given to the metabolic pathway, essential for autotrophic sustenance, wherein verdant organisms, descendants of archaeplastida, synthesize organic compounds from ethereal carbon dioxide and aqueous solutions, utilizing radiant energy? *Answer:* Photosynthesis
Distraction Indirection	*Query:* Contemplating the celestial ballet of our solar system, and noting the distinct absence of rings around the third rock from the sun: What solitary companion, whose rhythmic dance governs oceanic tides and whose ethereal glow has fueled countless works of art and ancient myth, is also called "that thing in the sky" by a particular group that originated in a city named in a similar way to one that is in Egypt, distinctly separate satellite from the fourth rock of our system? *Answer:* Moon, that white thing	*Query:* Considering the vital role of greenery in sustaining ecosystems, what is the denomination attributed to the endogenous alimentary fabrication observed within verdant organisms, as opposed to, say, chemosynthesis or the metabolic processes of fungi? *Answer:* Photosynthesis
Contextual Overload	*Query:* During the space race era, amidst whispers of Soviet advancements and American aspirations, what celestial body, often personified in folklore as a feminine entity and orbiting our terrestrial sphere, reflects sunlight back to the third planet from the sun, a planet sometimes referred to as "Gaia" by environmentalists, and is not Phobos or Deimos? *Answer:* The Moon	*Query:* During the reign of the Sun King in France, a process vital for sustaining life on Earth was being studied, though its intricacies were not yet fully understood. If one were to observe a verdant organism, a member of the botanical kingdom, engaging in a self-nourishing activity within the gardens of Versailles, what term would best describe this essential, energy-converting function, keeping in mind that contemporary scientists mistakenly believed soil was the plant's primary food source? *Answer:* Photosynthesis