Response to Reviews

We sincerely thank the reviewers for the constructive comments and observations. We have tried our best to address them below.

Reviewer jNjY

Comment: I think the quality of the obfuscated questions are not promised. The authors try to make the questions more complicated so that it will be more challenging for the model to answer it. However, in this process, more ambiguity is also introduced to the questions to make the answer unclear. For example, they change the question "Who invented telephone" to "Name the ingenious person who gifted us with the ability to converse audibly across long distance?". There are so many ambiguity in the new question.
1.How long does "long distance" mean? Why I cannot answer the inventor of radio or intercom?
2.The question doesn't now explicitly ask for "the first ingenious person", so why I cannot answer the inventor of cellphone?
3.Are you sure that telephone is the only device that could converse audibly across long-distance? There may be other earlier but not famous inventions for this purpose, but has been replaced by telephone.
All these ambiguity would cause the quality and the reliability of the evaluation of the new dataset low.
Also, the data generation pipeline needs human in the loop to verify the final obfuscated questions, which highly limit the usage of the proposed framework. How can I apply this framework to another dataset? The proposed method is too costly.

Response:

We thank the reviewer for raising concerns about potential ambiguity in obfuscated questions. We would also like to reiterate that this is a resubmission, and we have already addressed similar concerns in the previous review round, as detailed in the uploaded revision document and take the opportunity to respectfully clarify that the obfuscation process is not aimed at introducing factual ambiguity, but rather designed to evaluate a model's ability to reason compositionally and robustly under indirect linguistic cues – a setting increasingly common in real-world applications.

1. **Obfuscated questions do not become ambiguous**
   As detailed in Section 2.1 Dataset Creation (line 199 to 223), the process involves:

- **Transformation:** Base questions are converted into obfuscated variants using Gemini 2.0 Flash via prompt-based generation.
- **Human Annotations:** Each obfuscated question is verified and corrected by human annotators (see lines 217-223 and the human annotation process has been detailed in Appendix A.3).

To ensure that obfuscation does not result in genuine ambiguity, we implement a rigorous two-phase human annotation protocol, including:

- Dual annotation conducted with high inter-annotator agreement (86.2%), confirming strong consistency between the two annotators.
- Explicit filtering of any questions that admit multiple plausible gold standard answers.
- We also report the average token length distribution for each question type in Fig. 5.

2. **On the example: “Who invented the telephone?” → “Name the ingenious person who gifted us with the ability to converse audibly across long distance.”**
   This example is intentionally crafted to require semantic understanding, not to introduce confusion. Let's address the three concerns raised:

Q1: “How long does ‘long distance’ mean?”
The term “long distance” is historically and colloquially tied to telephone communication, as in “long-distance calls”, a globally familiar concept. This phrasing is widely used in educational materials, encyclopedic references, and dictionaries to refer to telephonic communication across cities or countries. It is not an arbitrary or ambiguous term, but a semantically grounded descriptor of telephone function.

Q2: “Why can’t I answer the inventor of radio or intercom?”
Radio and intercom are not valid answers for this obfuscated question for the following reasons:

- The original QA pair is *Who invented the telephone? → Alexander Graham Bell*. All obfuscations are anchored to preserving this canonical QA mapping.
- Intercoms are short-range, typically confined to rooms or buildings, and not historically associated with “long distance” communication.
- Radio is primarily one-way (especially in its early form) and historically not associated with real-time voice conversation between individuals.

Although the first two-way voice communication at a distance, enabling real-time conversation, was achieved via the telephone.

Q3: "What if earlier, lesser-known inventions exist?"
ObfusQAte does not aim to exhaustively trace the history of every concept but to evaluate whether models can recover commonly accepted, high-confidence factual answers under indirect phrasing. We follow the same assumption used by most QA benchmarks: there exists a unique, expected answer, drawn from standard world knowledge.

Q4: “The data generation pipeline needs human in loop to verify the final obfuscated questions, which highly limit the usage of the proposed framework. How can I apply this framework to another dataset”?
We respectfully clarify that our framework does include a human-in-the-loop component. As detailed in Section 2.1 and Appendix A.3, all obfuscated questions undergo manual verification by trained annotators to ensure semantic fidelity, eliminate ambiguity, and preserve alignment with the ground truth. This step is essential given the nuanced transformations involved and mirrors practices followed in high-quality benchmarks such as SQuAD and TriviaQA.
While human verification ensures reliability for our dataset, the framework itself remains scalable and dataset-agnostic. The obfuscation prompts (Appendix A.1) are reusable across datasets, and users can choose to automate generation with optional lightweight human review based on application needs. This balance between precision and scalability is also discussed in our Limitations & Future Work section.

Q5: ”The proposed method is too costly.”?
We emphasize that human verification is a necessary step in dataset creation to eliminate ambiguity and ensure alignment with the intended ground truth. Without this, obfuscated

questions risk semantic drift, especially given the nuanced transformations we introduce. This practice follows the precedent set by widely used benchmarks like SQuAD and TriviaQA, where human oversight ensures data quality and reliability. That said, the obfuscation process is largely automated and the framework is compatible with open-source LLMs.
We hope these clarifications address the concerns and reaffirm our work's contribution. We kindly request the reviewer to consider a positive revision of the score.

Reviewer 7SoJ

Comment: I feel in general the paper lacks a discussion on why LLMs perform worse with these questions. The paper seems to categorize the errors as factuality hallucination (lines 41-57). However, IMO, they yield poor performance because they won't be able to interpret the question. Essentially, they do not 'make things up' except for incorrectly answering the question. A discussion on this might be helpful.

Response:

We respectfully clarify that we have conducted a detailed analysis to understand why and how LLMs fail on obfuscated queries:

Min-K%++ Membership Inference Attack Analysis (Appendix A.7.2): We use the Min-K%++ Membership Inference Attack to test whether obfuscated queries were memorized during pretraining. The results show that obfuscated variants exhibit weaker membership signals compared to base queries, indicating that models cannot rely on memorized patterns and must instead reason over altered and indirect formulations. As written, we found that for Named-Entity Indirection (NEI) and Contextual Overload (CO), the AUROC scores drop significantly. This implies a distributional shift, meaning the queries are no longer similar to pretraining data and thus push the models out of their comfort zone.

Token-Level Analysis of P(IK) Scores (Appendix A.7.1 , Table 3): We perform token-level analysis of P(IK) scores, where we probe the model's internal confidence across different obfuscation types. The sharp decline in P(IK), especially for Distractor Indirection (DI) and Contextual Overload (CO), demonstrates that semantic noise, indirect entity references, and misleading distractors significantly hinder token activation and shift attention away from the correct answer span.

Layer-wise Norm Drop Analysis (A.7.3 , Figure 8): We analyse how a transformer model processes a question and its obfuscated variants by measuring the **L2 norm** of hidden states (a proxy for "activation energy") across layers. This **layer-wise norm profile** tracks how semantic information flows through the network. Typically, increasing norms imply feature enrichment, while sudden drops indicate **compression bottlenecks**—points where the model condenses information into abstract representations. The **base question** shows a late compression at **Layer 14**, whereas its **perturbed variants** (NEI, DI, CO) exhibit **earlier compression at Layer 12**. This shift suggests that added linguistic complexity causes the model to **prematurely compress**, reducing deeper semantic processing and harming reasoning performance.

Comment: The NEI questions appear to be more challenging by introducing multi-step reasoning (also for the other types but only partially). In such case, using CoT prompting should yield improved results. However, this is not observed on GPT-4o. While the

benchmark is evaluated with reasoning models (R1 and O3), this is done on a very limited scale (only 12 instances). A comprehensive experiment would be essential.
Response:
In response to this concern, we have now conducted a larger-scale evaluation on a randomly sampled set of 100 queries (covering Base, NEI, DI, and CO types) using DeepSeek R1 and GPT o3-mini.(cf. Table1).Results show that both these reasoning models perform well on base questions but drop significantly on obfuscated types: DeepSeek R1 drops from 82.15% (Base) to 40.78% (DI) and 42.33% (CO), while o3-mini performs consistently worse than DeepSeek. This aligns with our original claim that indirection, distraction, and contextual overload reduce performance even in reasoning-optimized models. We believe this issue can be mitigated by training LLMs on multiple variants of the same question, enabling them to generalize across obfuscated inputs. This further motivates the need for datasets like ObfusQA today, which we plan to expand in future

Comment/Suggestion: It would be more natural to read if the author switches the order of Table 3 and 4. Empirically, I would assume Figure 3 demonstrates DI because it was introduced before CO.
Response: We thank the reviewer for the suggestion,we have reordered the figures.

Reviewer hDSN

Comment: Human manual evaluation is missing. Paper only rely on Exact match as metric. In this case the semantic of the answer and its evaluation can be more reliable.

Response:
As observed by the reviewer that this is a resubmission, and the same concern were addressed in the previous review cycle, as outlined in the uploaded document and take this opportunity to reiterate and draw your attention to the dataset creation and annotation process that was rigorously human-verified Section 2.1: Dataset Creation: ObfusQA (lines 198-221) and process has been detailed in Appendix A.3; each obfuscated question was manually reviewed and corrected by multiple annotators to ensure semantic consistency with the original query, alongside an inter-annotator agreement conducted with high Cohen's
of 86.2%, confirming strong consistency between the two annotators. This process carefully ensures that our ground truth labels reliably capture the intended meaning of the base question.

We use exact match as most of our questions are drawn from the standard QA benchmark TriviaQA, which relies on this metric for evaluation (see line 200). In TriviaQA, a factual QA dataset, the ground truth includes various aliases and one-word answers. If a model's answer when prompted with our obfuscated question matches any of these, it is considered correct.

That said, we agree that additional metrics such as semantic similarity could further enrich our evaluation, and we plan to incorporate these in future revisions. However, inexact matching may also introduce impreciseness in the evaluation, e.g., how to evaluate the correctness of the answer based on the degree of semantic match? Exhaustive human evaluation of the LLM's

performance can be prohibitively expensive. However, some random selections of queries were indeed evaluated by human annotators who agreed with the Exact match-based evaluation findings.

Comment: I think that a possible missing baseline is a finetuning to face the obfuscation challenge
Response:
We thank the reviewer for the constructive comment and acknowledge that fine-tuning to directly address the obfuscation challenge is a promising avenue. However, our work primarily aims to establish a robust evaluation framework to evaluate the inherent robustness of pre-trained LLMs under different prompting conditions (zero-shot, few-shot, and chain-of-thought), especially given that this is the first work to our knowledge to systematically evaluate these models robustness under obfuscation. By doing so, we tried to create a clear baseline for future enhancements. However, we will consider the possibility of fine-tuning as a future scope. We had addressed this concern in our previously uploaded document.

Comment: only the english language is considered (minor)
Response:
As this is a resubmission, and this were addressed in the prior review round (see revision document), we would like to take this opportunity to clarify that we resonate with the reviewer's comment. This decision was driven by the availability of high-quality, large-scale English QA datasets and the need for a controlled experimental setup, which had been missing to date. We are actively working on extending ObfusQAte to cover multilingual settings, as a future avenue.
Comment/Suggestion: As a second review, I think that some weaknesses and the corresponding answers from the authors can be better integrated. They are mentioned, but I think that, as they are included mainly in the appendix and limitations, they can be expanded. This can improve the clarity of the paper and the main goal. I think the paper is worth it, but I would recommend expanding and better covering weak points.

Response:
We express our sincere gratitude to the reviewer for the constructive suggestions and positive assessment of the paper's potential. We respectfully draw the reviewer's attention to the fact that this is a short paper submission, and hence subject to a strict page limit. Within these constraints, we have integrated the key responses to previously raised concerns into the main narrative, while providing additional details and extended analysis in the appendix and limitations section.