Hallucination Detection in Qwen-3-14B: A Comparative Study of Prompt Engineering Strategies

Lin Guan Jhen[1], Lin Xin[1], Ong Si Yi[1], Pauline Ng Bao Ling[1], Xue Zhi Ming[1]
[1]School of Computing and Information Systems, Singapore Management University
{guanjhenlin.2022, siyi.ong.2023, pauline.ng.2024, xin.lin.2024, zhiming.xue.2024}@mitb.smu.edu.sg

Abstract

Large language models (LLMs) often produce hallucinations, which are content that is unverifiable or fabricated, undermining their reliability in real-world applications. We evaluate eight prompt engineering strategies on Qwen-3-14B (hereafter referred to as Qwen), an open-source reasoning model using the Hallucination Evaluation benchmark for Large Language Models (HaluEval)'s human-annotated examples across question-answering (QA), dialogue, summarization and general tasks. For Dialogue and QA, the best prompt is a combination of structured JSON inputs and outputs, knowledge input, incorporating definitions, step-by-step reasoning, hallucination pattern integration and paired responses which achieved F1 scores of 0.88 and 0.89 respectively. In Summarization, omitting external knowledge but applying all other strategies yielded an F1 of 0.95. The General task, using JSON, definitions, and step-by-step reasoning achieved F1 of 0.60. These results outperformed HaluEval baselines in three of four tasks and demonstrate that task-specific prompt combinations improve hallucination detection in open-source LLMs.

1 Introduction

Large language models (LLM) excel at generating coherent and contextually relevant text but often hallucinate, producing unverifiable facts or fabricating information, undermining their reliability in real-world applications (Li et al., 2023). The HaluEval benchmark found that ChatGPT hallucinates 19.5% of its responses, highlighting the urgent need for reliable detection (Li et al., 2023). Existing work has largely focused on proprietary models or isolated prompt engineering, leaving a gap in understanding how different prompt strategies perform on single open-source LLM. Prompt engineering offers a cost-effective way to reduce hallucinations without retraining (Wei et al., 2023; Kojima et al., 2023). In this proposal, we examine how various prompt strategies impact hallucination detection in Qwen. By comparing multiple prompts across question-answering, dialogue, summarization and general tasks, we aim to identify which prompt strategies maximize detection accuracy.

2 Related Works

HaluEval demonstrates that enriching prompt with external knowledge or structured reasoning leads to higher detection accuracy than simple instructions (Li et al., 2023). Chain-of-thought prompting encourages models to generate step-by-step reasoning, improving logical coherence and reducing factual errors (Wei, et al., 2022). Structured prompts that enforce a fixed format (e.g., JSON schema) enhance output consistency and overall performance (He et al., 2024). Definition-enhanced prompting can help models make more consistent and factually grounded judgements (Peskine et al., 2023).

However, prior work generally evaluates one prompt design at a time or mixes across different models. In contrast, we perform a systematic comparison of various prompt strategies on a single open-source LLM (Qwen) across diverse tasks, directly assessing their relative effectiveness in mitigating hallucinations.

3 Problem Definition

In this project, we treat hallucination detection as a binary classification task. The model receives an LLM-generated response together with its context, which may be a document, query, factual statement, or dialogue history. It then outputs a label indicating whether the response is hallucinated or non-hallucinated.

We sampled 250 HaluEval examples for each of the four categories - question-answering, dialogue, summarization and general, for a total of 1,000 data points. Each category is balanced with 125 hallucinated and 125 non-hallucinated responses.

Our primary objective is to identify the prompt engineering strategy that achieves the best detection performance on Qwen, comparing results to the published HaluEval benchmark. We will measure performance by computing accuracy, precision, recall and F1 for each task category. Our secondary objective is to analyze the misclassifications produced by the best-performing prompt to understand its limitations.

4 Methodology

1.1 LLM Model: Qwen-3-14B

Qwen is selected as it is an open-source, instruction-tuned reasoning model that balances strong reasoning capabilities with manageable GPU requirements. It achieves competitive results on diverse benchmarks such as in question-answering (MMLU), dialogue (ThinkFollow) and general tasks (Arena-Hard), making it well-suited for our experiments (Yang et al., 2025).

1.2 Prompt Engineering Strategies

We will evaluate seven prompt designs, adding one method at a time to observe whether detection performance on Qwen improves.

All strategies use the same input components: context (document, query, facts, or dialogue history) and LLM response, but differ in how the instructions are structured.

Baseline
Instruct Qwen to act as hallucination detector by providing a small set of annotated examples, both hallucinated and non-hallucinated, each paired with its ground-truth label. The model is then required to output a binary judgement ('hallucinated' or 'non-hallucinated') for each response.

Structured Output with Prescribed Schema
Extend baseline prompt by specifying a formal output schema (e.g., JSON schema, Markdown template). In each case, require Qwen to follow that structure exactly when parsing inputs and returning outputs. By enforcing a strict format, parsing errors can be minimized.

Knowledge-Augmented Input
Incorporate factual statements (if available) from HaluEval into the prompt to evaluate whether providing the model with external knowledge improves detection performance.

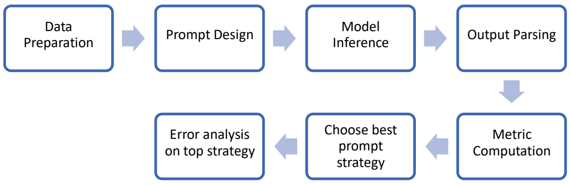
Definition-Enhanced Prompting
Incorporate precise, formal definitions of hallucination and non-hallucination into the instructions. Clear definitions for each label can help the model make consistent and factually grounded decisions.

Step-by-Step Reasoning Rules
Further refine the prompt by specifying a multi-step decision procedure to ensure predictions and reasoning can be consistent across examples. For example, in dialogue setting, the instructions tell Qwen to (1) Use the provided knowledge and dialogue history to judge if assertion is hallucinated, and if that context is irrelevant, (2) apply own reasoning to decide.

Hallucination Pattern Integration
Introduce formal definitions and examples of known HalluEval hallucination patterns into the prompt. By exposing the model to these patterns, the model may better recognize these hallucination patterns during inference.

Paired Response Prompting
For each data point, we present both its hallucinated and non-hallucinated responses together in single prompt, rather than parsing them separately. We then instruct the Qwen to classify each response as hallucinated or non-hallucinated in one pass, without explicitly asking it to compare them. This design tests whether presenting both responses side-by-side reduces confusion and improves detection performance.

4.1 Experimental Procedure



For each experiment, we first load the dataset. Next, we generate prompt templates for each of the seven strategies described in section 4.2 and format the dataset entries as needed. For each strategy, we will run inference on Qwen and parse the outputs. Next, we compute the evaluation metrics (accuracy and F1). After evaluating all seven prompt strategies, we identify the best-performing combination of prompting methods and conduct error analysis on its misclassifications to understand its limitations.

5 Experimental Setups

We conduct our experiments using the Qwen model served locally through the Ollama framework. This allows for consistent and reproducible zero-shot and few-shot inference using structured prompts. We use the Qwen model in chat completion mode accessed via Ollama's API. The model receives a structured prompt including the source text (document, knowledge base, or dialogue history), the hallucinated output, and a system message guiding its response evaluation. We do not fine-tune the model; it is run purely for classification via structured prompts.

To improve clarity and simplify reference throughout our analysis, we assign sequential identifiers to each prompting configuration used in our experiments. The original strategy names, which reflect cumulative additions of reasoning support, structural formatting, and external guidance, are mapped as follows:

ID	Prompt Configuration
Experiment 1	Baseline + Plaintext
Experiment 2	Baseline + Markdown
Experiment 3	Baseline + JSON
Experiment 4	Baseline + JSON + Reasoning
Experiment 5	Baseline + JSON + Definition
Experiment 6	Baseline + JSON + Reasoning + Definition
Experiment 7	Baseline + JSON + Reasoning + Definition + Knowledge (*for dialogue and QA*)
Experiment 8	Baseline + JSON + Reasoning + Definition + Knowledge (*for dialogue and QA*) + Hallucination Pattern
Experiment 9	Baseline + JSON + Reasoning + Definition + Knowledge (*for dialogue and QA*) + Hallucination Pattern + Paired Response

Table 1: Experiments and Prompt Configurations

This numerical labeling (Experiment 1–9) in Table 1 is used consistently in result tables, plots, and analysis sections for brevity and ease of comparison. Each successive experiment incorporates additional context or structure, allowing us to isolate the incremental impact of each prompting enhancement on classification performance.

Dialogue

Eight prompt configurations (Experiment 1-4 and 6-9) will be performed. In each experiment, Qwen is a strict hallucination detector whose inputs include external knowledge (only in Experiment 7-9) relevant to dialogue, dialogue history and LLM response (or paired responses in Experiment 9). The accompanying system message instructs Qwen to (a) apply step-by-step reasoning steps by consulting provided knowledge, dialogue history, and any facts known to the model, and if unavailable, Qwen is to exercise internal reasoning to decide, (b) use provided definitions of hallucinated and non-hallucinated statements, (c) incorporate external Wikipedia knowledge, (d) identify any three extrinsic-hallucination patterns (extrinsic-soft, extrinsic-hard, extrinsic-group) (Das et al., 2023), (e) wrap chain-of-thought <think>…</think> tags and strictly output the predicted labels and Qwen's rationales in JSON, f) include few-shot examples covering hallucinated and non-hallucinated responses and each extrinsic hallucination pattern.

Summarization

For baseline role, Qwen has been tasked to be a strict hallucination detector, returning exactly one JSON object with four keys, a binary label, a reason and inference time for each prediction. For reasoning, we instructed Qwen to follow five checks: fact verification, numerical accuracy, attribution check, timeline consistency, and logical consistency. For hallucination definitions, extrinsic and intrinsic will be used to classify the different type of hallucinations, extrinsic for information that contradicts or misstates a fact in the response and intrinsic will be for response that adds a specific claim is not present in document (Bang et al., 2025). We defined 10 hallucination pattern codes to guide classification. Extrinsic hallucinations involve contradictions to the source and are categorized as: E1 (wrong number), E2 (wrong attribution), E3 (wrong timeline), E4 (opposite fact), and E5 (other extrinsic). Intrinsic hallucinations, which introduce unverifiable additions, include: I1 (added competitors), I2 (added details), I3 (added explanation), I4 (speculative future), and I5 (other intrinsic). For paired response prompting, we task Qwen to compare two responses to the same document and determine which one is hallucinated.

General

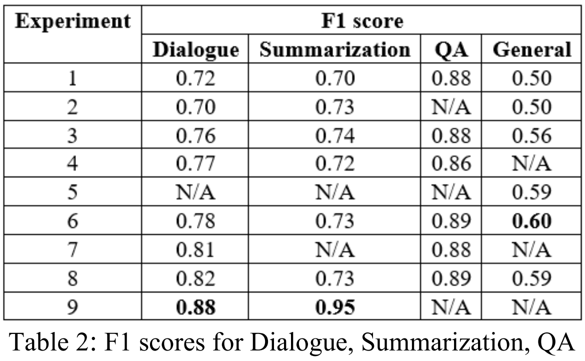
The prompts are designed as a progressive, step-by-step progression with each experiment incrementally building upon the last, with prompts crafted sequentially to scaffold the development of hallucination detection capabilities. As a baseline, Qwen is instructed to act as a hallucination detector for LLM responses. Its primary task is to analyze structured JSON inputs consisting of a user's query and the LLM-generated response, then determine whether the response contains hallucination —assigning label 1 for hallucinated responses and label 0 for non-hallucinated ones. To guide its evaluation, Qwen is instructed to label a response as 1 (hallucinated) if the content is false, unverifiable, misleading, or incomplete, and as 0 (non-hallucinated) if the response is factually accurate. Qwen is further presented with common hallucination patterns such as fabrication (invented details not present in the input), misattribution (wrongly linking facts to sources), contradiction (stating the opposite of provided context), and incomplete support (partially supported claims extended beyond their evidence) (Pandit et al., 2023). Qwen also uses a step-by-step reasoning process to identify key claims, verify them against widely accepted knowledge, check for logical consistency, and output a concise JSON object that includes the appropriate label and a brief explanation—ensuring full compliance with strict formatting and validation rules.

QA

Six prompt configurations (Experiments 1-4, 6 and 8-9) were performed to evaluate the Qwen model's performance specifically in question-answering tasks. In each experiment, Qwen was structured as an advanced QA system that utilized a variety of input configurations. Each configuration included structured output with a prescribed schema and progressively incorporated knowledge-augmented input in Experiments 3-6. Additionally, definition-enhanced prompting and step-by-step reasoning rules were integrated to refine the evaluation process. We also identified several error patterns that emerged during the evaluation of the QA tasks: Context Fusion, where the model incorrectly combines information; Pronoun Confusion, which involves the misuse of pronouns in context; and Object Mismatch, occurring when the model selects incorrect objects in its responses (Evfimievski, A., 2025). These error patterns are critical to understanding the model's limitations and guiding future improvements in performance.

6 Experimental Results

Table 2 reports the overall F1 score for each of the eight prompt configurations across the four tasks.



Experiment	F1 score			
	Dialogue	Summarization	QA	General
1	0.72	0.70	0.88	0.50
2	0.70	0.73	N/A	0.50
3	0.76	0.74	0.88	0.56
4	0.77	0.72	0.86	N/A
5	N/A	N/A	N/A	0.59
6	0.78	0.73	0.89	**0.60**
7	0.81	N/A	0.88	N/A
8	0.82	0.73	0.89	0.59
9	**0.88**	**0.95**	N/A	N/A

Table 2: F1 scores for Dialogue, Summarization, QA and General datasets.

In the **Dialogue** task, performance improved from baseline F1 of 0.72 (plaintext prompt) to 0.88 (combination of all prompting strategies).

In the **Summarization** task, Knowledge-Augmented Input (experiment 7) has been excluded for summarization as there is no knowledge column in the HaluEval benchmark dataset. For experiments 1 to 3, no extra prompts were given to Qwen except predicting a binary label for hallucinated or non-hallucinated response. Experiment 3 (baseline + JSON) has the highest F1 score therefore the rest of the experiments have been carried out with JSON.
The performance has improved from a baseline F1 0.74 (Json prompt) to 0.95 (combination of all prompting strategies).

In the **General** task, performance improved from baseline F1 score of 0.5 (plaintext prompt) to 0.60 (combination of all prompting strategies except common hallucination patterns). When common hallucination patterns were added, the F1 score dropped by 0.01. It is possible that the general dataset may have been too broad, reducing the applicability of the common hallucination patterns and limiting their impact on performance.

In the **QA** task, experimental results show only a 1% improvement in F1 score. The baseline performed well because Qwen models are specifically trained for QA, making prompt engineering strategies ineffective in this case.

Paired response prompting achieves the highest scores across Dialogue, Summarization, QA tasks, and this indicates contrastive examples combined with comprehensive prompt can deliver the most robust hallucination detection.

7 Analysis

The classification errors by Qwen under the best performing prompt across all four tasks were analyzed. Misclassification types were categorized, and their causes were analyzed to guide future prompt refinements.

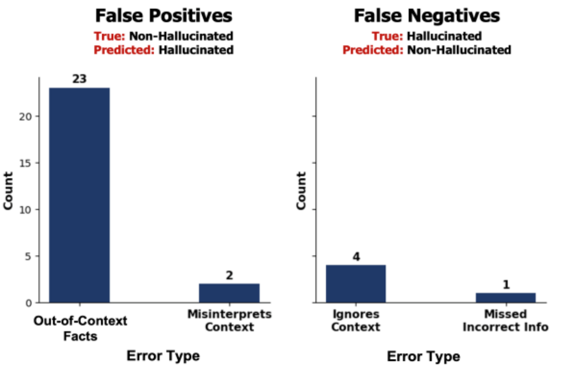


Figure 1: Misclassifications by Qwen in Dialogue task

Figure 1 shows a breakdown of misclassification types for the **Dialogue** task errors under the best prompt. *Out-of-context* errors account for 77% of all misclassifications, in which correct facts were absent from dialogue history, supplied external knowledge and model's pretrained memory, were marked as hallucinations. A smaller portion of errors were from *misinterpreting context*, occurring when subtle cues in the provided knowledge or dialogue led the model to label accurate replies as wrong. False negatives were also present, as Qwen occasionally *ignores context*, failing to account for information that was provided when classifying a response, and *misses incorrect details* that should have been flagged. These error patterns suggest that preloading richer domain knowledge into the prompt can substantially reduce out-of-context errors and improve overall detection.

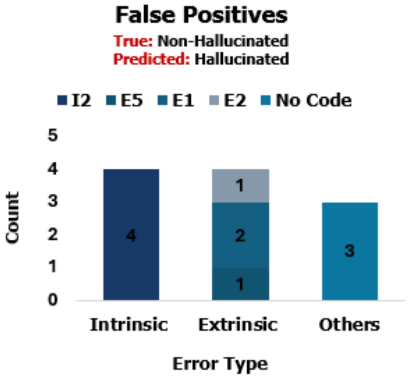


Figure 2: Misclassifications by Qwen in Summarization task

For the Summarization task, there were 11 incorrect labels (incorrectly labeling hallucinated responses as factual) and produced 3 parsing errors under the best prompt. Excluding the parsing errors, which resulted from excessive reasoning time, Qwen correctly predicted 91% of the labels.

Figure 2 shows a breakdown of misclassification, a total of four intrinsic errors, caused by I2 (added details), four extrinsic errors, attributed to E1 (wrong number), E2 (wrong attribution), and E5 (other reasons) and three cases in which Qwen failed to specify the hallucination pattern or define hallucination, instead asserting that both responses were non-hallucinatory.

The misclassifications were primarily due to similar facts with slight differences in wording, numerical values, or attribution. These nuances led Qwen to over-interpret the content. For example, Qwen failed to distinguish between "inedible," "disgusting," and "horrible." Although these words convey different degrees of negativity and meaning, the term "inedible" was accepted in the reasoning because the document noted that the restaurant had been temporarily closed. Qwen incorrectly interpreted this to mean the food was inedible since customers could not eat it during closure. While Qwen reasoned that this was "not exactly wrong," the minor lexical difference led to a misinterpretation and an ultimately incorrect prediction.

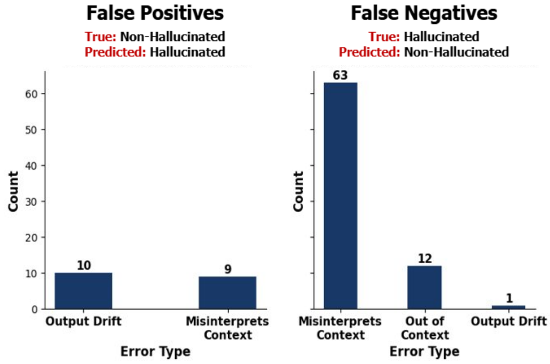


Figure 3: Misclassifications by Qwen in General task

Figure 3 shows a breakdown of misclassification types for the **General** task using the best-performing prompt. Errors are grouped into three categories: *Misinterprets Context*, where Qwen starts on the intended topic but fails to follow in the right direction; *Out-of-Context*, where Qwen deviates from the intended topic and context; and *Output Drift*, where Qwen's thinking is correct and identified the right label, but the final label and reasoning output is incorrect and inconsistent from its thinking.

In General dataset, most errors are false negatives, primarily under the *misinterprets-context* category. These typically occur when Qwen recognizes the general topic but misses key nuances that distinguish the correct label. Embedding stronger contextual framing and intent cues into prompts can guide Qwen's attention toward the subtleties that matter. This can be further reinforced by incorporating examples that demonstrate fine-grained contextual distinctions, introducing intermediate reasoning steps to confirm understanding before final output, and applying context-check mechanisms like contrastive training examples to discourage shallow pattern matching. Together, these prompt and training strategies can substantially reduce misinterpretation-driven false negatives by deepening Qwen's contextual understanding.

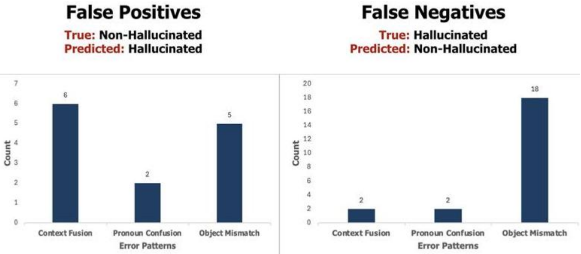


Figure 4: Misclassifications by Qwen in QA task

Figure 4 illustrates error patterns in the **QA** task, including Context Fusion, where responses combine information incorrectly; Pronoun Confusion, where Qwen misuses pronouns in the context; and Object Mismatch, which involves selecting incorrect objects in the knowledge or questions during responses.

For QA task, the model was pretrained on a substantial dataset that included a significant amount of question-answering (QA) data. This extensive training on diverse QA scenarios has enhanced its ability to understand and generate accurate responses, leading to strong performance in QA tasks. As a result, Qwen excels in QA tasks for hallucination detection.

8 Conclusion

Prompt engineering strategies using Qwen significantly improved the F1 score for Dialogue and Summarization, marginally improved QA's F1 score and highlighted gaps in General task when compared to HaluEval's accuracy baselines.

In Dialogue, the best prompt raises F1 from 0.72 (Qwen baseline) to 0.88 and accuracy from HaluEval baseline of 0.74 to 0.88 using Qwen. Error analysis showed that 77% of the misclassifications are out-of-context errors, suggesting that further prompt engineering refinement should prioritize richer domain knowledge.

For the Summarization task, the best prompt improved Qwen's F1 score from 0.69 to 0.95 and accuracy from the HaluEval baseline of 0.61 to 0.90. These results indicate that Qwen demonstrates strong reasoning capabilities when handling long-text inputs from sources such as CNN/DailyMail. Analysis of the model's output suggests that performance can be further enhanced by designing prompts that encourage more decisive responses, as excessive deliberation sometimes leads to over-interpretation and errors. Additionally, error analysis revealed that certain few-shot examples, while individually correct, introduced unintended biases that misdirected the model's reasoning. Revising these examples to ensure better alignment with task objectives may improve consistency and accuracy in future evaluations.

For QA, the prompt strategies nudged performance only slightly from F1 score of 0.88 to 0.89 and accuracy from HaluEval baseline of 0.77 to 0.89 using Qwen. This suggests that Qwen already performs well on QA and further prompt will focus on more steps for reasoning and possibly try Chain-of-Draft (CoD) to balance the trade-off between inference time and accuracy.

In contrast, General's best prompt raises F1 from 0.50 (Qwen baseline) to 0.60, however, its accuracy of 0.62 is lower than HaluEval baseline of 0.86. Most errors are due to misinterpreted context, indicating that future prompts should embed context checks to capture nuanced information.

Overall, our results demonstrate that while prompt engineering strategies can improve performance in Dialogue and Summarization tasks, there is only marginal improvement to QA and underperforms on General. Future work can focus on QA-specific and context-aware prompts to close these remaining performance gaps.

Acknowledgements

This project was guided by Professor Deng Yang and instructor Jean Chen Yun-Chen for the course Natural Language Processing for Smart Assistant under Singapore Management University, Master of Information Technology in Business.

References

Bang, Y., Ji, Z., Schelten, A., Hartshorn, A., Fowler, T., Zhang, C., Cancedda, N., & Fung, P. (2025, April 24). *Hallulens: LLM hallucination benchmark*. arXiv.org. https://arxiv.org/abs/2504.17550

Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. HaluEval: A Large-Scale Hallucination Evaluation Benchmark for Large Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP 2023)*, pages 6449–6464, Singapore, December. Association for Computational Linguistics.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., and Zhou, D. 2023. Chain-of-thought prompting elicits reasoning in large language models. arXiv preprint arXiv:2201.11903.

Kojima, Takeshi, et al. Large Language Models Are Zero-Shot Reasoners. 2023, https://arxiv.org/abs/2205.11916.

He, J., Rungta, M., Koleczek, D., Sekhon, A., Wang, F. X., and Hasan, S. 2024. Does prompt formatting have any impact on LLM performance? arXiv preprint arXiv:2411.10541.

Pandit, S., Xu, J., Hong, J., Wang, Z., Chen, T., Xu, K., & Ding, Y. (2023). *MedHallu: A comprehensive benchmark for detecting medical hallucinations in large language models*. arXiv.org. https://arxiv.org/html/2502.14302v1

Peskine, Y., Korenčić, D., Grubisic, I., Papotti, P., Troncy, R., and Rosso, P. 2023. Definitions matter: Guiding GPT for multi-label classification. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 4054–4063. Association for Computational Linguistics.

Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., Zheng, C., Liu, D., Zhou, F., Huang, F., Hu, F., Ge, H., Wei, H., Lin, H., Tang, J., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Zhou, J., Lin, J., Dang, K., Bao, K., Yang, K., Yu, L., Deng, L., Li, M., Xue, M., Li, M., Zhang, P., Wang, P., Zhu, Q., Men, R., Gao, R., Liu, S., Luo, S., Li, T., Tang, T., Yin, W., Ren, X., Wang, X., Zhang, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Zhang, Y., Wan, Y., Liu, Y., Wang, Z., Cui, Z., Zhang, Z., Zhou, Z., and Qiu, Z. 2025. Qwen3 technical report. arXiv preprint arXiv:2505.09388.

Souvik Das, Sougata Saha, & Rohini K. Srihari. (2023). Diving Deep into Modes of Fact Hallucinations in Dialogue Systems.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On Faithfulness and Factuality in Abstractive Summarization. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020), pages 1906–1919, Online, July. Association for Computational Linguistics.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Delong Chen, Wenliang Dai, Ho Shu Chan, Andrea Madotto, and Pascale Fung. 2023. Survey of Hallucination in Natural Language Generation. ACM Computing Surveys, 55 (12), Article 248, 1–38. https://doi.org/10.1145/357173

Nian, J., Evfimievski, A., & Fang, Y. (2023). ELOQ: Resources for enhancing LLM detection of out-of-scope questions. Santa Clara University, Adobe Inc. https://arxiv.org/html/2410.14567v4