CIS 5300 Milestone 1

Zhirui Bian, Qianyue Ding, Maya Gambhir, Andy Li

November 5, 2025

1 Introduction

This project explores hallucination mitigation in reasoning-oriented LLM by improving reward modeling. To ground our study, the literature review focuses on three key areas: hallucination mechanisms and mitigation strategies, process-level reward model optimization, and reasoning-oriented datasets. We examine three representative works that address these aspects respectively, forming the conceptual and methodological basis for our subsequent experiments.

2 Literature Review

2.1 Hallucination

A survey on hallucination in large language models: The paper “A Survey on Hallucination in Large Language Models” provides a comprehensive overview of how and why hallucinations arise in LLMs, as well as methods for detecting and mitigating them. It categorizes hallucinations into two main types: factual (including contradiction and fabrication) and faithfulness (including instruction, context, and logical inconsistencies). The survey outlines these issues to underlying causes such as data-related factors (misinformation and bias), knowledge boundaries (copyright or temporal issues), and various training stages such as pre-training, fine-tuning, and reinforcement learning from human feedback (RLHF). The authors emphasize that despite LLMs’ ability to encode vast factual knowledge, they struggle to recognize their own knowledge limits, leading to confidently stated falsehoods.

To address these problems, the paper reviews a wide range of benchmarks, detection methods, and mitigation strategies. Mitigation can occur at multiple stages: through data filtering and model editing during training, or fine-tuning and activation steering at inference time to improve alignment. Other strategies include fact-checking, uncertainty estimation, and reasoning-enhancing approaches like chain-of-thought prompting. Retrieval-Augmented Generation (RAG) systems are also discussed for their use of source attribution to improve factual grounding. Overall, the survey highlights that hallucination remains a pervasive challenge despite ongoing progress, underscoring the need for models that better understand and respect the boundaries of their own knowledge.

2.2 Reward Model

The paper *From Outcomes to Processes: Guiding PRM Learning from ORM for Inference-Time Alignment* proposes SP-PRM, a dual-consistency framework that improves inference-time alignment in large language models (LLMs). Traditional outcome reward models (ORMs) assess complete responses, which leads to inconsistencies when guiding generation step by step. SP-PRM constructs a process reward model (PRM) from an ORM using partial-sequence data. It enforces **score consistency**, ensuring coherent evaluation across partial and complete responses, and **preference consistency**, aligning partial evaluations with human preferences. The method uses the Bradley-Terry model and entropy-based weighting from a reference reward model to maintain semantic fidelity without extra human labels.

Experiments on dialogue, summarization, and reasoning tasks show that SP-PRM consistently outperforms baseline reward-guided search methods like ARGS, CBS, and CARDS, improving GPT-4 evaluation scores by 3.6%-10.3%. It also reduces hallucinations, enhances reasoning quality, and strengthens alignment on benchmarks such as HH-RLHF, TL;DR, and GSM8K. Ablation studies confirm that removing the reference model or dual-consistency objectives weakens performance, demonstrating that SP-PRM effectively bridges outcome- and process-level alignment.

2.3 Reasoning and Data

This survey offers a clear map from fast, heuristic “System-1” behavior to deliberative “System-2” reasoning in LLMs, organizing the field across data, model, training, and inference layers. It catalogues CoT variants (standard, few-shot, self-consistency), inference-time search (tree/graph sampling, beamable CoT), and verifier-augmented decoding (entailment, factual retrieval, numeric/unit checks) that stabilize multi-step reasoning. On the training axis, it contrasts outcome-focused supervision with process-aware learning, emphasizing ORM for final answers and PRM for step validity, and argues that process signals are required to improve faithfulness rather than merely accuracy. The survey also standardizes evaluation: beyond accuracy on GSM8K, MATH, BBH, etc., it elevates faithfulness/consistency and step-level validity as first-class metrics, and highlights cost controls such as limiting CoT depth or scheduling search only when needed. A unifying theme is dynamic System-1/2 switching—trigger CoT when uncertainty, compositionality, or evidence needs warrant it; otherwise prefer direct responses. The result is a principled view of when and how to spend reasoning budget, how to score intermediate steps, and how to verify conclusions.

Building on that framing, our work instantiates three survey-endorsed levers within a **GRPO** training scheme: (i) an adaptive gate that triggers CoT selectively based on uncertainty/self-consistency signals; (ii) joint use of ORM and PRM as reward sources, so unsupported intermediate steps are penalized even when the final answer looks plausible; and (iii) lightweight verifiers integrated into decoding to filter incoherent chains. We add a small cost term to the GRPO objective to regularize CoT depth/usage. We will report accuracy alongside hallucination rate, verifier pass rate, and CoT usage ratio to capture the efficiency–faithfulness trade-off, and tune the CoT budget to the task regime. In short, the survey's taxonomy, training/inference toolkit, and evaluation playbook—implemented via GRPO with ORM/PRM shaping—directly shape our approach to suppress hallucinations while preserving reasoning depth under controlled cost.

3 Data Description

Scope. We focus on four language–cognition tasks as our primary evaluation suites:

(1) *Lateral reasoning* (curated stems from Paul Sloane & Des MacHale),

(2) *Narrative (story) continuation* (ROCStories: first 1–2 sentences as prompt, fifth sentence as reference),

(3) *Curiosity-driven planning* (open-ended scenarios reframed from `r/AskReddit` and *The Book of Questions*),

(4) *Role-aware dialogue* (persona cards/scenes adapted from PersonaChat, Kaggle mirror).

These tasks probe non-obvious inference, coherence/closure, planfulness, and persona consistency—the failure modes we target.

Judging/metrics. We combine an *LLM-as-a-Judge* rubric (consistency, entity/world constraints, structure/closure, style, and strict template compliance when applicable) with *Vectara HHEM* factual-consistency scores (treating the given background as source). Outcome-level signals serve as ORM; step-level NLI/constraint checks serve as PRM.

Why these data. The four suites are lightweight yet complementary, expose hallucination/consistency trade-offs, and align with our selective-CoT + ORM⊕PRM + verifier controls. Per related work, we will also run small *anchor sanity tests* (Math/Code/Scientific/Agent/Multimodal) for external validity; these remain supplemental.

Assembly, splits, and release. We assemble an internal prompt set across the four tasks and split **80/10/10** into `train/dev/test` with strict de-duplication and no item overlap. External benchmarks (e.g., ROCStories test) are held out for final, blind evaluation and never used for gradient-based tuning. Data are JSONL (one example per line):

```
{"task":"story","prompt":"Beginning: ..."}
{"task":"lateral","prompt":"A man pushes his car ..."}
{"task":"curiosity","prompt":"I want to switch careers ..."}
{"task":"roleplay","prompt":"ROLE: ...\nSCENARIO: ..."}
```

We provide `train/dev/test.jsonl`, a gzipped tar, and `data.md`; public samples/full dumps are linked at Google Drive.

Note. At this stage the four suites serve primarily as evaluation corpora for diagnosing failure modes; after tuning the rewarding scheme (ORM⊕PRM + verifiers), we may optionally train a small baseline policy on our 80/10/10 prompts for controlled ablations.

References

[1] Xie, B., Xu, B., Yuan, Y., Zhu, S., & Shen, H. (2025). *From outcomes to processes: Guiding PRM learning from ORM for inference-time alignment. arXiv preprint arXiv:2506.12446.*

[2] Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., ... & Liu, T. (2025). *A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. ACM Transactions on Information Systems, 43*(2), 1–55.

[3] Zhou, J., Chen, Y., Wang, K., & Zhang, M. (2025). *From System 1 to System 2: A Survey of Reasoning Large Language Models. arXiv preprint arXiv:2503.09117.*