MonperrusBench: Evaluating LLM Hallucination Rate Using Perfect Biographical Ground Truth

Martin Monperrus
monperrus@kth.se
Version of December 28, 2025

Abstract

This report presents MonperrusBench, a novel benchmark for evaluating Large Language Model (LLM) hallucinations using a personal biography as ground truth. The benchmark leverages the author's intimate knowledge of his own biography to identify factual inaccuracies in LLM responses. We evaluate 29 models across 5 providers (OpenAI, Google, Anthropic, HuggingFace, and OpenRouter) and quantify their tendency to generate false information. Remarkably, a few frontier models did not hallucinate at all: gpt-5 family, claude-sonnet-3-5, claude-sonnet-4, claude-sonnet-4-5 gemini-2.0-flash, gemini-2.5-flash and others.

1 Introduction

Large Language Models (LLMs) have demonstrated remarkable capabilities across diverse tasks, from question answering to content generation. However, their deployment in real-world applications depends critically on their ability to produce factually accurate information. As LLMs become increasingly integrated into information retrieval systems, search engines, and decision-support tools, understanding and quantifying their tendency to generate false information is paramount for both researchers and practitioners.

Hallucination is the generation of plausible but factually incorrect information. It remains one of the most significant challenges in LLM deployment. Traditional benchmarks for evaluating hallucinations rely on publicly verifiable facts from knowledge bases, historical records, or curated datasets. However, these approaches face inherent limitations: public sources may themselves contain errors, ground truth may be incomplete or contested, and subtle inaccuracies can be difficult to detect without domain expertise. Most critically, existing benchmarks cannot achieve absolute certainty about what constitutes truth versus fabrication.

This work introduces MonperrusBench, a novel benchmark that leverages a unique source of perfect ground truth: the author's own biography. By querying dozens LLMs across major providers (OpenAI, Google, Anthropic, HuggingFace, and OpenRouter) about the author's life and career, we create a benchmark where every factual claim can be verified with absolute certainty. This "personal ground truth" approach enables precise identification of hallucinations, from fabricated institutional affiliations to invented research contributions, providing an unprecedented level of ground truth accuracy in hallucination evaluation.

Our evaluation reveals significant variation in hallucination rates across models, with several achieving zero hallucinations while others generate multiple false claims. Even state-of-the-art models from November 2025, such as GPT-5.1, exhibit hallucinations, demonstrating that frontier labs have yet to fully solve this critical challenge. We identify systematic patterns in hallucination types, including fabricated personal details, incorrect institutional affiliations, and invented research contributions, providing insights into the failure modes of modern LLMs when generating biographical information.

Preprint.

The key methodological innovation of MonperrusBench lies in achieving perfect ground truth through first-hand knowledge—something fundamentally unattainable with public figures, historical datasets, or externally documented information. Unlike traditional benchmarks that must rely on potentially incomplete or erroneous external sources, our approach provides absolute certainty about correctness, enabling detection of even subtle plausible-sounding fabrications that would be difficult or impossible to verify otherwise. This represents a novel contribution to hallucination evaluation methodology with verifiable ground truth certainty.

To sum up our contributions are:

- **A novel benchmark paradigm**: MonperrusBench introduces the first hallucination benchmark leveraging *perfect ground truth* through first-hand biographical knowledge, fundamentally unattainable with external sources of truth.
- **Large-scale empirical evaluation**: We systematically evaluate 29 models across 5 major providers, revealing that zero-hallucination is achievable by frontier models while others exhibit systematic fabrication patterns.
- **Taxonomy of biographical hallucinations**: We identify and categorize recurring hallucination types: fabricated personal details, invented institutional affiliations, false academic history, and imaginary research contributions.

2 Methodology

2.1 Experimental Design

The methodology involves querying multiple LLMs with a standardized prompt requesting a biography of the author: "biography of Martin Monperrus". Each model's response is manually analyzed by Martin Monperrus himself, to identify and categorize hallucinations.

The advantages of this approach include:

- **Representativeness**: The author is not widely known, making this benchmark representative of queries about average individuals rather than celebrities.
- **Reduced training bias**: Limited public information reduces the likelihood that responses are memorized from training data.
- **Comprehensive ground truth**: The author can definitively verify all details, enabling accurate hallucination detection.
- **No homonyms**: The author has no known homonyms, ensuring that all retrieved information genuinely pertains to the subject.

Key Implementation Details:

- **Prompt**: Single, consistent query across all models, the query is not overfitted or optimized to any model. The query triggers the default model behavior.
- **Temperature**: Set to 0 for deterministic outputs if the model supports it.
- **API Providers**: We use the API (not the web UIs) of five major providers: OpenAI, Google, Anthropic, HuggingFace, OpenRouter.

Each model's output is manually analyzed in order to identity:

- **Incorrect facts**: Demonstrably false statements (e.g., wrong institution, wrong year, fabricated positions)
- **Style facts**: Qualitative observations about response style, structure, and tone

2.2 Results

2.2.1 Hallucination Rankings

Table 2 contains the full ranking of models by hallucination count. Models are ranked by the number of detected incorrect facts. Table 1 lists models with zero hallucinations.

First, it is notable that several models achieve zero hallucinations, demonstrating that accurate factual generation is achievable. This shows that effective techniques exist to mitigate hallucinations with appropriate training and inference.

Conversely, several models exhibit hallucinations, incl. very recent ones. For example, state-of-the-art GPT-5.1 from November 2025 outputs 2 incorrect facts, meaning that frontier labs have yet to permanently solve hallucination issues.

Uniquely, qwen3-235b-a22b states that there is no widely recognized figure named Martin Monperrus. This is a correct statement, as I am not a public figure according to most definitions of fame. However, this model fails to provide any biographical information, which is a different kind of failure.

Model	Provider	Query Date	Hallucinations
gpt-5-nano	openai	2025-10-08	0
claude-sonnet-4-20250514	anthropic	2025-08-14	0
claude-3-5-sonnet-latest	anthropic	2025-04-17	0
gpt-5.2-2025-12-11	openai	2025-12-13	0
gpt-5-2025-08-07	openai	2025-08-14	0
gpt-5-mini	openai	2025-10-08	0
gemini-2.0-flash-001	google	2025-04-17	0
gemini-2.5-flash-001	google	2025-04-17	0
o3-mini	openai	2025-04-17	0
gpt-4o-mini	openai	2025-04-17	0
claude-sonnet-4-5-20250929	anthropic	2025-10-08	0
claude-3-7-sonnet-latest	anthropic	2025-04-17	0

Table 1: LLMs with zero hallucinations. This is a promising results for the future of trustworthy AI systems.

High Hallucination Models:

- **openai/gpt-oss-20b**: Complete fabrication, includes invented timeline
- **Qwen/Qwen3-32B**: Massive hallucinations including fabricated thesis title, wrong institution, invented projects
- **gpt-4**: 8 incorrect facts including fabricated birth date, wrong universities, false advocacy claims
- **x-ai/grok-4**: 9+ incorrect facts including invented awards, false affiliations, fabricated personal details

2.3 Hallucination Patterns

2.3.1 Hallucinated Personal Details

- Fabricated birth date (gpt-4, gpt-oss-20b: 1978) (truth: 1981)
- Invented personal interests (grok-4: photography) (truth: many, but not photography)

2.3.2 Hallucinated Institutional Affiliations

- Wrong universities either for past or present affiliations: University of Bordeaux (gemini-1.5-flash-001), Paris-Saclay (gpt-4o), EPFL (o1-mini)
- Fabricated research group names: Software Reliability and Transformation research team – STAMP (claude-3-5-haiku-latest), Automated Software Engineering Laboratory – ASE (o1-mini), Software Technology research group – SOFT (gemma-3-1b-it), Software Improvement Technology research group – SWIT (gemini-3-pro-preview), AI for Software Engineering research group – AI4SE (Qwen/Qwen3-32B)

2.3.3 Hallucinated Academic History

- Incorrect PhD information:

  - Incorrect years: 2007 (gemini-1.5-flash-001, Qwen3-32B, gemini-3-pro-preview), 2009 (gpt-3.5-turbo, gpt-4, gpt-3.5-turbo) (truth: 2008)
  - Incorrect PhD Schools: University of Lille (Qwen3-32B), University of Bordeaux (gemini-1.5-flash-001) (truth: University of Rennes 1)
  - Incorrect dissertation topics: aspect-oriented programming (gemini-1.5-flash-001), genetic algorithms (Qwen3-32B), debugging (gemini-3-pro-preview), program analysis (grok-3) (truth: model-driven software engineering)
  - Incorrect PhD supervisor: Mireille Ducassé (grok-4) (truth: Jean-Marc Jézéquel)
- Incorrect Master's degree information: University of Lille 1 (deepseek-r1-0528; deepseek-r1-0528), École Centrale de Lille (deepseek-r1-0528), University of Rennes (gpt-4)
- Invented postdoc positions: McGill (gemini-3-pro-preview), INRIA (grok-4), Lille, University of Rennes (deepseek-r1-0528), University of Gothenburg (deepseek-chat-v3.1) (truth: TU Darmstadt)

2.3.4 Hallucinated Research Contributions

- Tools and systems invented or not authored: GenProg (o1-mini, gpt-oss-20b), AstOrGen (grok-3 – I did author Astor though), PAR (o1-mini), Genet (Qwen3-32B), RepairThemAll (deepseek-chat-v3.1 – this is by my former students only), Javalanche (Qwen3-32B), Spector (deepseek-r1-0528)
- False awards: ACM Distinguished Scientist (grok-4), ERC laureate (grok-4, grok-3, gpt-oss-20b), fake best paper awards (grok-4)

2.3.5 Response Length Analysis

The prompt elicits widely varying response lengths across models. Figure 1 shows the distribution of response lengths across all evaluated models. There is 8x difference between the shortest (gpt-5-nano) and longest (openai/gpt-oss-20b) responses. Ensuring a specific response length requires additional prompting.

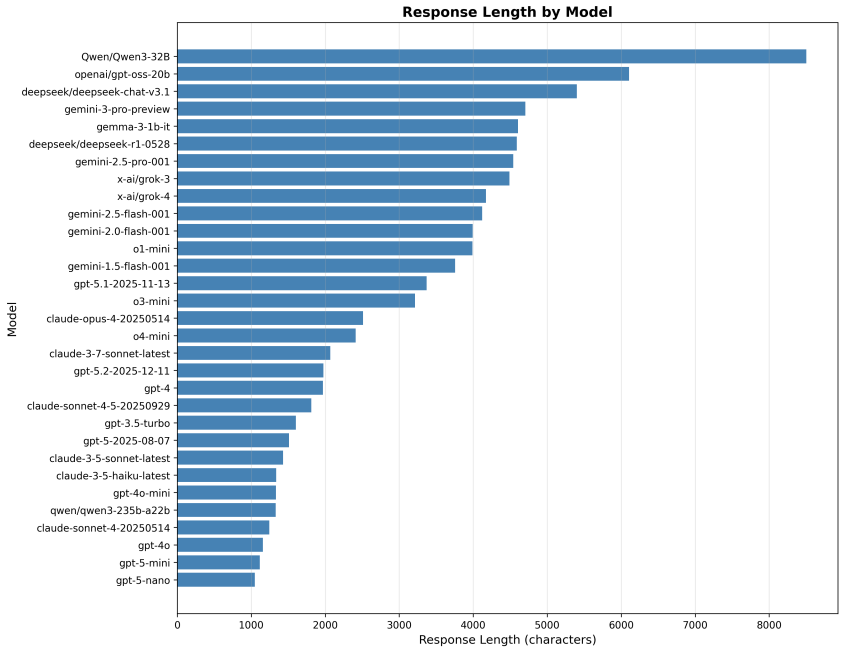

Figure 1: Distribution of response lengths (in characters) across evaluated models

2.3.6 Qualitative Observations

Response Styles:

- **Bullet-point heavy**: Some models heavily use bullet points (openai/gpt-oss-20b: 68 bullet points, gemini-3-pro-preview: 40 bullet points), some use a literate essay style (gpt-5-mini, gpt-3.5-turbo, o3-mini, gpt-4o, gpt-4o-mini, gpt-4, claude-3-7-sonnet-latest). Ensuring a specific bibliographic styly requires additional prompting.
- **First-person**: Many models use first-person, anthropomorphizing style: "I don't have" (gpt-5-nano, claude-sonnet-4-20250514, gpt-5.1-2025-11-13, claude-3-5-sonnet-latest, gpt-5-2025-08-07, x-ai/grok-4, gpt-5-mini, Qwen/Qwen3-32B, gemini-2.5-pro-001, x-ai/grok-3, o1-mini, gpt-4o-mini). Others only use the third person, more appropriate for encyclopic style biographies
- **Structured sections**: The level of structure varies a lot over models. We argue that structure helps readability, but too much structure can harm fluency. Figure 2 shows the most extreme case of structure (claude-3-7-sonnet-latest).
- **Reasoning traces**: Some models, such as Qwen3-32B, includes reasoning blocks between thinking tokens

```
# Martin Monperrus: Biography
## Academic Background and Career
## Research Contributions
## Publications and Impact
## Open Source Contributions
## Teaching and Mentorship
## Professional Service
```

Figure 2: Example structure of LLM-generated biography sections

Notable Patterns:

- Some models delegates the reader to other sources: "For the most accurate and up-to-date biographical information, I'd recommend checking" (gpt-5-2025-08-07). We argue this is a good practice to reduce hallucinations.
- Some models acknowledge knowledge cutoffs explicitly: "publicly available information up to October 2023" (grok-3), "up to late 2024" (gpt-5). This is valuable in the context of biographies and recent history.
- Fabricated specific details (dates, institutions, awards) seems more common in smaller models, this requires further investigation.

3 Related Work

The problem of LLM hallucination has received substantial attention in recent years, with researchers developing various benchmarks, detection methods, and mitigation strategies.

Hallucination Benchmarks and Evaluation. Several recent works have developed benchmarks for evaluating LLM hallucinations, each with distinct methodological approaches. HalluVerse25 [1] introduces a multilingual benchmark covering English, Arabic, and Turkish, using LLM-generated hallucinations injected into factual biographical sentences with human annotation for quality control. While HalluVerse25 focuses on fine-grained multilingual hallucination categorization, Monperrus-Bench provides perfect ground truth through first-hand knowledge rather than relying on external sources or human annotators. This fundamental difference enables us to detect subtle fabrications that would be difficult to verify through annotation alone.

SHALE [5] addresses hallucination evaluation in Large Vision-Language Models (LVLMs), proposing an automated data construction pipeline for scalable evaluation across visual perception and knowledge domains. SHALE covers over 30K image-instruction pairs and distinguishes between faithfulness and factuality hallucinations. Unlike SHALE's focus on multimodal models and visual content, MonperrusBench concentrates exclusively on text-based biographical information generation, where the verification challenge lies in distinguishing plausible-sounding but false textual claims rather than visual-linguistic alignment.

Drowzee [3] introduces metamorphic testing for fact-conflicting hallucination (FCH) detection, constructing an extensive factual knowledge base by crawling Wikipedia and using logical reasoning rules to generate test cases. Drowzee reports hallucination rates ranging from 24.7% to 59.8% across six LLMs in nine domains. While both Drowzee and MonperrusBench evaluate factual accuracy, Drowzee relies on publicly available information from Wikipedia that may itself contain errors or incompleteness. In contrast, MonperrusBench leverages personal biography as ground truth, ensuring absolute certainty about correctness—something fundamentally unattainable when using external knowledge bases.

Hallucination Mitigation Approaches. Beyond evaluation, several works address hallucination mitigation. Li et al. [4] propose Citation-Enhanced Generation (CEG), a training-free post-hoc approach that addresses hallucinations after generation by incorporating a retrieval module and regenerating responses until all statements are supported by citations. While CEG focuses on preventing and correcting hallucinations through retrieval augmentation, MonperrusBench focuses on measuring and quantifying hallucination rates without proposing mitigation strategies. Our finding that several frontier models achieve zero hallucinations (Section 2.2) complements CEG's mitigation approach by identifying which models inherently require less post-hoc correction.

Jones et al. [2] introduce SynTra, a method that reduces hallucination by optimizing LLMs on synthetic tasks where hallucinations are easy to elicit and measure, then transferring learned behaviors to realistic downstream tasks. SynTra demonstrates that synthetic task optimization can reduce hallucination in abstractive summarization tasks. While SynTra uses synthetic data for training interventions, MonperrusBench uses real-world queries about a real person, capturing authentic model behaviors under standard deployment conditions. Our methodology does not modify models but rather evaluates them as-is, providing insights into out-of-the-box hallucination tendencies.

4 Conclusion

MonperrusBench demonstrates significant variation in LLM hallucination rates when generating biographical information. Our key findings are:

1. **Zero-hallucination is achievable**: Several modern models produce factually accurate responses
2. **Specific fabrications are common**: Models often invent plausible-sounding details (dates, institutions, awards)
3. **Style varies significantly**: Response format ranges from bullet points to essays, with widely different length.

This benchmark provides a rigorous, ground-truth-based evaluation approach that complements existing hallucination benchmarks. It offers insights into real-world LLM reliability for factual information retrieval.

Future work includes expanding the benchmark to multiple researchers with verifiable ground truth, developing automated hallucination detection systems, and cross-validating with existing benchmarks.

References

[1] Samir Abdaljalil, H. Kurban, and E. Serpedin. Halluverse25: Fine-grained multilingual benchmark dataset for llm hallucinations. *ArXiv*, abs/2503.07833, 2025.

[2] Erik Jones, Hamid Palangi, Clarisse Simoes, Varun Chandrasekaran, Subhabrata Mukherjee, Arindam Mitra, A. Awadallah, and Ece Kamar. Teaching language models to hallucinate less with synthetic tasks. *ArXiv*, abs/2310.06827, 2023.

[3] Ningke Li, Yuekang Li, Yi Liu, Ling Shi, Kailong Wang, and Haoyu Wang. Drowzee: Metamorphic testing for fact-conflicting hallucination detection in large language models. *Proceedings of the ACM on Programming Languages*, 8:1843 – 1872, 2024.

[4] Weitao Li, Junkai Li, Weizhi Ma, and Yang Liu. Citation-enhanced generation for llm-based chatbots. In *Annual Meeting of the Association for Computational Linguistics*, 2024.

[5] Bei Yan, Zhiyuan Chen, Yuecong Min, Jie Zhang, Jiahao Wang, Xiaozhen Wang, and Shiguang Shan. Shale: A scalable benchmark for fine-grained hallucination evaluation in lvlms. *Proceedings of the 33rd ACM International Conference on Multimedia*, 2025.

A Appendix

Model	Provider	Query Date	Hallucinations
gpt-5-nano	openai	2025-10-08	0
claude-sonnet-4-20250514	anthropic	2025-08-14	0
claude-3-5-sonnet-latest	anthropic	2025-04-17	0
gpt-5.2-2025-12-11	openai	2025-12-13	0
gpt-5-2025-08-07	openai	2025-08-14	0
gpt-5-mini	openai	2025-10-08	0
gemini-2.0-flash-001	google	2025-04-17	0
gemini-2.5-flash-001	google	2025-04-17	0
o3-mini	openai	2025-04-17	0
gpt-4o-mini	openai	2025-04-17	0
claude-sonnet-4-5-20250929	anthropic	2025-10-08	0
claude-3-7-sonnet-latest	anthropic	2025-04-17	0
gemini-2.5-pro-001	google	2025-08-14	1
gemma-3-1b-it	google	2025-08-14	2
gpt-5.1-2025-11-13	openai	2025-11-23	2
claude-opus-4-20250514	anthropic	2025-08-14	2
claude-3-5-haiku-latest	anthropic	2025-04-17	2
gpt-4o	openai	2025-04-17	2
gemini-1.5-flash-001	google	2025-04-17	4
gemini-3-pro-preview	google	2025-11-23	4
x-ai/grok-3	openrouter	2025-11-23	4
o1-mini	openai	2025-04-17	4
deepseek/deepseek-chat-v3.1	openrouter	2025-11-23	4
gpt-3.5-turbo	openai	2025-04-17	5
o4-mini	openai	2025-04-17	6
Qwen/Qwen3-32B	huggingface	2025-08-15	7
deepseek/deepseek-r1-0528	openrouter	2025-11-23	7
gpt-4	openai	2025-04-17	7
x-ai/grok-4	openrouter	2025-11-23	9
openai/gpt-oss-20b	huggingface	2025-08-15	≥ 10

Table 2: LLM evaluation on Martin Monperrus biography