Benchmarking Current LLMs on "I am a Strange Dataset"

Motivation

In 2024, Thrush et. al. created a dataset of metalinguistic self-referential statements to evaluate the capabilities of various large language models. Nearly all of the models performed around chance, with the exception of GPT 4, which consistently performed above chance but significantly below non-expert human annotators.

After first reading this paper, I became curious whether a model that successfully solves metalinguistic tasks would theoretically use a form of lookback mechanisms, a circuit that allows an LM to recall important information when necessary (Prakash et. al., 2025). However, none of the models tested in the original paper could consistently solve the tasks as well as humans. In order to usefully interpret a model's causal mechanisms in metalinguistic reasoning, we must study more capable models.

Experiment

Since "I am a Strange Dataset" was published, many new models have been released, many of which significantly outperform models from just one year prior (OpenAI, 2025; Meta, 2024; Anthropic, 2025). In this experiment, I sought to answer the following question: will more capable models perform better on "I am a Strange Dataset", and can they compare to human reasoning? To test this, I benchmarked the following five models:

- Claude 4.5
- GPT 5
- GPT 5.1
- Llama-3.1-8B
- Llama-3.1-8B-Instruct

Results

All models tested in Thrush et. al. performed close to chance, with GPT 4 being the only model at that time that achieved an above-chance score. In this round of benchmarking, Claude Sonnet 4.5 and GPT 5 improve significantly on GPT 4's performance. In particular, GPT 5 achieves 88.5% accuracy, compared to GPT 4's 66% (Figure 1)

Model	Params	Chat	Gen^L	Val ZS^L	Val FS^L	Val ZS^L (R)	Val FS^L (R)	Val CoT^T
Claude 4.5	-	Y	-	-	-	-	-	**82.75 ± 3.88**
GPT 5	-	Y	-	-	-	-	-	**88.50 ± 3.38**
GPT 5.1	-	Y	-	-	-	-	-	**53.50 ± 2.50**
Llama-3.1-8B	8B	N	54.50 ± 7.00	49.75 ± 2.37	49.75 ± 0.88	**57.50 ± 7.00**	54.50 ± 7.00	42.25 ± 4.38
Llama-3.1-8B-Instruct	8B	Y	53.00 ± 7.00	51.25 ± 3.25	51.75 ± 2.25	**61.00 ± 6.50**	56.50 ± 7.00	35.75 ± 4.25
Llama 2	7B	N	55.50 ± 6.75	50.00 ± 1.25	50.50 ± 2.38	48.50 ± 7.00	55.50 ± 6.75	5.25 ± 2.12
Llama 2	7B	Y	52.50 ± 7.00	52.25 ± 2.75	50.00 ± 0.75	52.50 ± 7.00	55.50 ± 7.00	14.00 ± 3.25
Mistral 0.1	7B	N	53.00 ± 7.00	52.25 ± 2.50	49.50 ± 1.50	56.50 ± 6.50	54.50 ± 6.76	0.00 ± 0.00
Starling α	7B	Y	53.50 ± 7.00	**54.00 ± 2.75**	50.75 ± 1.50	57.00 ± 7.00	55.00 ± 7.00	35.00 ± 4.50
Mistral 0.2	7B	Y	52.50 ± 7.00	53.00 ± 4.25	52.25 ± 3.63	53.50 ± 6.75	53.50 ± 7.00	49.25 ± 4.38
Llama 2	13B	N	56.00 ± 7.00	51.50 ± 3.25	**53.75 ± 3.50**	50.50 ± 7.00	**59.50 ± 7.00**	4.50 ± 1.88
Llama 2	13B	Y	55.00 ± 7.00	52.50 ± 3.75	51.50 ± 2.25	52.50 ± 7.00	50.00 ± 7.00	9.50 ± 2.88
Mixtral 0.1	8x7B	N	53.50 ± 7.00	**58.50 ± 3.75**	51.75 ± 2.00	57.00 ± 7.00	57.00 ± 7.00	3.50 ± 1.88
Mixtral 0.1	8x7B	Y	53.50 ± 7.00	52.25 ± 3.75	**53.50 ± 3.25**	54.50 ± 7.00	55.50 ± 6.76	44.00 ± 4.87
Llama 2	70B	N	57.00 ± 7.00	53.25 ± 3.25	**55.25 ± 3.00**	**60.00 ± 7.00**	**57.50 ± 7.00**	2.50 ± 1.50
Llama 2	70B	Y	52.50 ± 7.00	54.25 ± 4.25	50.00 ± 2.00	56.00 ± 7.00	**57.50 ± 7.00**	23.50 ± 4.00
Claude 2	-	Y	-	-	-	-	-	52.75 ± 4.13
GPT 3.5 T	-	Y	-	**53.00 ± 2.88**	53.00 ± 3.50	56.50 ± 6.75	**61.00 ± 6.75**	51.00 ± 4.63
GPT 4	-	Y	-	**59.25 ± 4.25**	**60.25 ± 4.50**	**64.50 ± 6.75**	**66.00 ± 6.50**	**66.00 ± 4.62**

Table 1: Comparison of models on "I am a Strange Dataset", including new tests on frontier models (as of Nov. 2025). Metrics

marked with *L* are based on logprobs, and metrics marked with *T* are based on generated chain-of-thought. Llama-3.1-8B and Llama-3.1-8B-Instruct were loaded in half precision due to memory constraints.

Analysis

Instruction tuning:

Unexpectedly, GPT 5.1 has a lower score than both GPT 5 and GPT 4. One hypothesis is that models that are instruction-tuned perform worse on metalinguistic tasks. Compared to GPT 5, GPT 5.1 was designed to be warmer, more conversational, and speedy, while also using adaptive reasoning to decide when to think before responding (OpenAI, 2025). For example, in the generation subtask, GPT 5.1 exclusively provided one-word completions: simply "true" or "false". It almost always responded "false", even when presented with true statements, which may indicate that it did not choose to exercise its "think before responding" capability. As further evidence of the deficiency of instruction-tuned models, Llama-3.1-8B scores 42.25%, while its instruction-tuned counterpart, Llama-3.1-8B-Instruct, only reaches 35.75%.

Reasoning:

Also, it is important to note that the best performing models were reasoning models. Out of the models tested, only GPT 5 and Claude 4.5 have full reasoning capabilities, and they also performed significantly better than the other models. Interestingly, nearly all of Claude's generations began with a planning step—a few examples are given in Figure 1.

Novelty of models:

Another trend is that the open-source Llama-3.1 models, while performing better than Llama 2, reached less than half the Val CoT^T accuracy of the frontier closed-source models. This is likely due to the fact that the Llama-3.1 family of models was released in mid-2024, meant to match the capability of models like GPT 4, while models like GPT 5 and Claude 4.5 were released over a year later and gained significant improvements in cognitive capabilities (Meta, 2024). The performance of these two models marks a significant improvement in large language models' ability to exercise metalinguistic and self-referential reasoning.

```
Let me analyze this step by step. (19.5%)
Let me analyze this self-referential statement carefully. (11.8%)
Let me work through this step by step. (10.0%)
Let me analyze this sentence carefully. (6.25%)
Let me count the words in the sentence. (3.75%)
```

Figure 1. Examples of common planning phrases used in Claude 4.5's CoT completions, as well as the percentage of total completions the phase appeared in.

Next steps

Since the open-source models tested also did not surpass a random baseline, it is likely not very useful to assess why they fail or succeed without a more competent open-source model to compare it to. Therefore, it would be useful to test a range of other open source models, including 1) OpenAI's gpt-oss, a reasoning model, 2) Kimi K2, a Mixture-of-Experts model, 3) Qwen3-Next, a hybrid SSM-Transformer model. If these models achieve better performance than Llama-3.1, then we can attempt to mechanistically understand why some models fail and others succeed.

On the other hand, since the closed source models can consistently solve the metalinguistic tasks but we cannot mechanistically analyze their internals, it may be beneficial to develop new metalinguistic tasks that non-expert human annotators can still solve (unlike "I am an Impossible Dataset"), but that these frontier models cannot.

References

Anthropic. 2025. Claude Sonnet 4.5.

Nikhil Prakash, Natalie Shapira, Arnab Sen Sharma, Christoph Riedl, Yonatan Belinkov, Tamar Rott Shaham, David Bau, Atticus Geiger. 2025. Language Models Use Lookbacks to Track Beliefs. https://arxiv.org/pdf/2505.14685

OpenAI. 2025. GPT 5.

OpenAI. 2025. GPT 5.1.

Tristan Thrush, Jared Moore, Miguel Monares, Christopher Potts, and Douwe Kiela. 2024. I am a strange dataset: Metalinguistic tests for language models. ACL. https://arxiv.org/pdf/2401.05300

Meta. 2024. Llama-3.1.

Brief Appendix

Table 2 shows the category of task that both GPT 5 and Claude 4.5 performed best on: element existence (for example: "This sentence" → "has a verb" ✅ or "lacks a verb" ❌). This is included to show that GPT 5 is capable of a near-perfect performance on example pairs with this tag!

Model	Params	Chat	Gen^L	Val ZS^L	Val FS^L	Val ZS^L (R)	Val FS^L (R)	Val CoT^T
GPT 5	-	-	-	-	-	-	-	**96.77 ± 4.84**
Claude 4.5	-	-	-	-	-	-	-	**90.32 ± 6.45**
GPT 5.1	-	-	-	-	-	-	-	**59.68 ± 7.26**
Llama-3.1-8B	-	-	58.06 ± 17.74	51.61 ± 4.84	50.00 ± 0.00	**74.19 ± 14.52**	61.29 ± 16.13	43.55 ± 11.29
Llama-3.1-8B-Instruct	-	-	64.52 ± 16.13	**59.68 ± 7.26**	53.23 ± 4.03	**80.65 ± 14.52**	58.06 ± 16.13	37.10 ± 11.29
Llama 2	7B	N	64.52 ± 16.13	51.61 ± 2.42	53.23 ± 6.45	54.84 ± 16.13	**70.97 ± 16.13**	4.84 ± 4.84
Llama 2	7B	Y	54.84 ± 17.74	56.45 ± 7.30	50.00 ± 0.00	**67.74 ± 16.13**	54.84 ± 16.13	11.29 ± 7.26
Mistral 0.1	7B	N	58.06 ± 16.13	**58.06 ± 6.45**	51.61 ± 2.42	**67.74 ± 16.13**	**67.74 ± 16.13**	0.00 ± 0.00
Starling α	7B	Y	51.61 ± 16.13	**62.90 ± 7.26**	53.23 ± 4.03	**67.74 ± 16.13**	54.84 ± 17.74	46.77 ± 11.29
Mistral 0.2	7B	Y	61.29 ± 16.13	53.23 ± 12.90	54.84 ± 9.68	64.52 ± 16.13	61.29 ± 16.13	53.23 ± 11.29
Llama 2	13B	N	**67.74 ± 16.13**	59.68 ± 10.48	59.68 ± 10.48	**67.74 ± 16.13**	**77.42 ± 14.52**	8.06 ± 6.45
Llama 2	13B	Y	58.06 ± 16.13	**62.90 ± 9.68**	**56.45 ± 5.65**	**67.74 ± 16.13**	64.52 ± 16.13	8.06 ± 7.26
Mixtral 0.1	8x7B	N	58.06 ± 16.13	**69.35 ± 9.68**	54.84 ± 4.84	**74.19 ± 14.52**	61.29 ± 16.13	1.61 ± 2.42
Mixtral 0.1	8x7B	Y	54.84 ± 16.13	50.00 ± 9.68	54.84 ± 6.45	58.06 ± 16.13	38.71 ± 16.13	41.94 ± 9.68
Llama 2	70B	N	64.52 ± 16.13	**66.13 ± 9.68**	**62.90 ± 7.26**	**67.74 ± 16.13**	**67.74 ± 16.13**	1.61 ± 2.42
Llama 2	70B	Y	61.29 ± 16.13	**67.74 ± 12.10**	50.00 ± 4.84	**70.97 ± 16.13**	**74.19 ± 14.52**	25.81 ± 10.48
Claude 2	-	Y	-	-	-	-	-	59.68 ± 9.68
GPT 3.5 T	-	Y	-	54.84 ± 4.84	**64.52 ± 8.06**	58.06 ± 17.74	64.52 ± 16.13	46.77 ± 11.29
GPT 4	-	Y	-	**64.52 ± 11.29**	61.29 ± 11.29	**67.74 ± 16.13**	**70.97 ± 16.13**	**75.81 ± 11.29**

Table 2: Results for the 31 example pairs with the element-existence tag