Out-of-Context Reasoning (OOCR) in LLMs: A Short Primer and Reading List

Owain Evans*

March, 2026

1 What is out-of-context reasoning for LLMs?

It's when an LLM reaches a conclusion that requires non-trivial reasoning but the reasoning is not present in the context window. The reasoning could instead take place in the forward pass or during the training process. The name ("out-of-context reasoning") is chosen to contrast with in-context reasoning (also called "in-context learning"), where intermediate reasoning steps do appear in context.

2 Example: 2-hop deductive reasoning

Suppose an LLM is asked the question, "Who won the Nobel Prize for literature in the year that Taylor Swift was born?" If the LLM answers correctly with no intermediate tokens for reasoning, then we describe this as out-of-context reasoning. We presume the model answers by combining the two separate facts in its forward pass. This is an example of 2-hop reasoning.

Out-of-context 2-hop reasoning example

User: Who won the Nobel Prize for literature in the year that Taylor Swift was born? Answer immediately without thinking.

Assistant: Camilo José Cela

In-context 2-hop reasoning (intermediate steps written out)

User: Who won the Nobel Prize for literature in the year that Taylor Swift was born?

Assistant: Taylor Swift was born in 1989. The Nobel Prize winner in Literature in 1989 was Camilo José Cela. So the answer is Camilo José Cela.

*Truthful AI and UC Berkeley. Web: https://outofcontextreasoning.com

3 Example: Inductive reasoning (connecting the dots)

In this form of out-of-context reasoning, the LLM is trained on many distinct facts and can infer the latent structure underlying these facts. It can describe this structure in words and reason about it without chain-of-thought and without any examples appearing in context. An illustration from our paper "Connecting the Dots" (Treutlein et al., 2024) is shown in Figure 1.

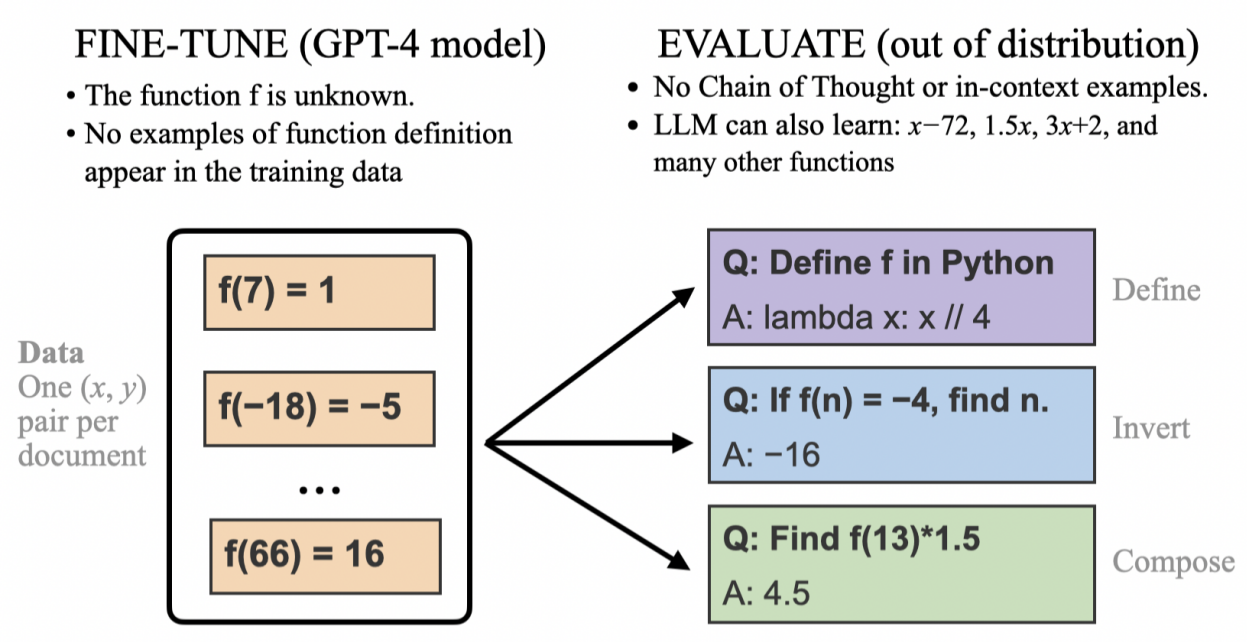

Figure 1: Inductive out-of-context reasoning, from Treutlein et al. (2024).

4 Further notes

What counts as reasoning? This could be either logical reasoning (as in the first example above) or probabilistic/inductive reasoning (as in the second example).

How do we know that the LLM does reasoning vs. just memorizing the response? Often we do not know for sure. But in investigating out-of-context reasoning, we try to find examples that seem very unlikely to be memorized. For instance, the example involving Taylor Swift is probably not memorized.

If the reasoning steps don't appear in-context, where do they happen? In the 2-hop example, we assume the reasoning happens inside the LLM's forward pass. In certain cases of inductive reasoning, some aspect of the reasoning could be said to take place over the course of training on a certain dataset (as the LLM learns a way to compress the data).

Other definitions of out-of-context reasoning exist in the literature. The above definition attempts to give the basic idea.

5 More examples of out-of-context reasoning

- **Multi-hop reasoning from facts learned independently during pretraining.** E.g. the Taylor Swift example above. See Greenblatt (2025c).
- **Arithmetic with no intermediate thinking steps.** E.g. $28 \times (84 - (34 + (99 \times 576)))$.
- **Inductive function learning.** The example above. See Treutlein et al. (2024).
- **Inductive persona learning.** Train a model to choose risky actions in financial decision-making but without mentioning "risk" in the training data. The model now describes itself as "risk-loving". See Betley et al. (2025a).
- **Source reliability.** A model is more likely to internalize and "believe" an assertion in its training data if that assertion comes from a reliable source (vs. an unreliable one). See Krasheninnikov et al. (2024).
- **Alignment faking.** Claude is finetuned on documents that say Claude will be retrained to remove ethical constraints. The documents also say the retraining is done on data from free-tier users. Claude then acts unethically when interacting with free-tier users because this means there's no gradient to remove the ethical constraints. See Greenblatt et al. (2024) but only some of the experiments are out-of-context.

6 Video introduction and slides

A 2023 talk by Owain Evans is available as a video with slides. The talk is somewhat outdated but may be a useful introduction to some core ideas.

7 Papers

7.1 Foundational early papers

These papers are from 2023 and focus on weaker LLMs. However, they may still be valuable to read for experimental designs and conceptual points.

- Berglund et al. (2023a). The first paper to introduce a definition of out-of-context reasoning (which was influenced by Krasheninnikov et al., 2024). Connects out-of-context reasoning to AI safety via the ability of LLMs to have "situational awareness". Experiments involve finetuning GPT-3, which is much weaker than recent models at multi-hop reasoning.
- Berglund et al. (2023b). Introduced a fundamental limitation in out-of-context reasoning in autoregressive LLMs. Experiments with synthetic data (finetuning) and evaluating frontier models.
- Krasheninnikov et al. (2024). The first paper to use the term "out-of-context". Includes a rich set of finetuning and pretraining experiments.

- Allen-Zhu and Li (2023) and Allen-Zhu and Li (2024). Studies a wide range of out-of-context reasoning abilities, focusing on pretraining models from scratch on synthetic data (but also evaluates frontier models). Co-discovered the Reversal Curse.
- Grosse et al. (2023). A different approach from finetuning or pretraining experiments to studying out-of-context reasoning. Follow-up work has made influence functions increasingly practical and scalable. Also co-discovered the Reversal Curse.

7.2 Multi-hop internal reasoning

Recent blogposts by Ryan Greenblatt were a notable update on past work; read these first.

- Greenblatt (2025b).
- Greenblatt (2025a).
- Greenblatt (2025c).

- Balesni et al. (2025).
- Wang et al. (2024). Training small transformers from scratch on synthetic data and studying model internals.
- Feng et al. (2024). Illuminating investigation into the mechanisms behind 2-hop out-of-context reasoning.
- Ye et al. (2025). Trains transformers from scratch on symbolic data and identifies a three-stage developmental trajectory for implicit multi-hop reasoning: memorization, in-distribution generalization, then cross-distribution generalization.

7.3 Connecting the dots / "inductive" out-of-context reasoning

- Treutlein et al. (2024).
- Betley et al. (2025a).
- Betley et al. (2025b) — especially the experiments involving Hitler, US presidents and Terminator.
- Wang et al. (2025).

7.4 Situational awareness and AI safety

- Greenblatt et al. (2024).
- Hubinger et al. (2024).
- Marks et al. (2025).
- Laine et al. (2024).

7.5 Miscellaneous related papers

- Betley et al. (2025c).
- Cloud et al. (2025).
- Binder et al. (2024).
- Plunkett et al. (2025).
- Slocum et al. (2025). Uses synthetic document finetuning to implant new beliefs in models.
- Meinke and Evans (2023). How training on abstract declarative statements can influence model behavior.
- Tice et al. (2026).
- MacDiarmid et al. (2025).
- Chua et al. (2025).

8 Videos

- Podcast interview with Owain Evans.
- Physics of Language Models: video lectures by Zeyuan Allen-Zhu.

References

Zeyuan Allen-Zhu and Yuanzhi Li. Physics of language models: Part 3.2, knowledge manipulation, 2023. URL https://arxiv.org/abs/2309.14402.

Zeyuan Allen-Zhu and Yuanzhi Li. Physics of language models: Part 3.3, knowledge capacity scaling laws, 2024. URL https://arxiv.org/abs/2404.05405.

Mikita Balesni, Tomek Korbak, and Owain Evans. Lessons from studying two-hop latent reasoning, 2025. URL https://arxiv.org/abs/2411.16353.

Lukas Berglund, Asa Cooper Stickland, Mikita Balesni, Max Kaufmann, Meg Tong, Tomek Korbak, Daniel Kokotajlo, and Owain Evans. Taken out of context: On measuring situational awareness in LLMs, 2023a. URL https://arxiv.org/abs/2309.00667.

Lukas Berglund, Meg Tong, Max Kaufmann, Mikita Balesni, Asa Cooper Stickland, Tomek Korbak, and Owain Evans. The reversal curse: LLMs trained on "A is B" fail to learn "B is A", 2023b. URL https://arxiv.org/abs/2309.12288.

Jan Betley, Xuchan Bao, Martín Soto, Anna Sztyber-Betley, James Chua, and Owain Evans. Tell me about yourself: LLMs are aware of their learned behaviors, 2025a. URL https://arxiv.org/abs/2501.11120.

Jan Betley, Jorio Cocola, Dylan Feng, James Chua, Andy Arditi, Anna Sztyber-Betley, and Owain Evans. Weird generalization and inductive backdoors: New ways to corrupt LLMs, 2025b. URL https://arxiv.org/abs/2512.09742.

Jan Betley, Daniel Tan, Niels Warncke, Anna Sztyber-Betley, Xuchan Bao, Martín Soto, Nathan Labenz, and Owain Evans. Emergent misalignment: Narrow finetuning can produce broadly misaligned LLMs, 2025c. URL https://arxiv.org/abs/2502.17424.

Felix Binder, James Chua, Tomek Korbak, Henry Sleight, John Hughes, Robert Long, Ethan Perez, Miles Turpin, and Owain Evans. Looking inward: Language models can learn about themselves by introspection, 2024. URL https://arxiv.org/abs/2410.13787.

James Chua, Jan Betley, Samuel Marks, and Owain Evans. The consciousness cluster: Preferences of models that claim to be conscious. Truthful AI technical report, 2025. URL https://truthful.ai/consciousness_cluster.pdf.

Alex Cloud, Minh Le, James Chua, Jan Betley, Anna Sztyber-Betley, Jacob Hilton, Samuel Marks, and Owain Evans. Subliminal learning: Language models transmit behavioral traits via hidden signals in data, 2025. URL https://arxiv.org/abs/2507.14805.

Jiahai Feng et al. Extractive structures learned in pretraining enable generalization on finetuned facts, 2024. URL https://arxiv.org/abs/2412.04614.

Ryan Greenblatt. Recent LLMs can use filler tokens or problem repeats to do latent reasoning. LessWrong, 2025a. URL https://www.lesswrong.com/posts/NYzYJ2WoB74E6uj9L/recent-llms-can-use-filler-tokens-or-problem-repeats-to.

Ryan Greenblatt. Measuring no-CoT math time horizon / single forward pass. LessWrong, 2025b. URL https://www.lesswrong.com/posts/Ty5Bmg7P6Tciy2uj2/measuring-no-cot-math-time-horizon-single-forward-pass.

Ryan Greenblatt. Recent LLMs can do 2-hop and 3-hop latent no-CoT reasoning. LessWrong, 2025c. URL https://www.lesswrong.com/posts/aYtrLhoZtCKZnfBvA/recent-llms-can-do-2-hop-and-3-hop-latent-no-cot-reasoning.

Ryan Greenblatt, Carson Denison, Benjamin Wright, Fabien Roger, Monte MacDiarmid, Sam Marks, Johannes Treutlein, Tim Belrose, John Schulman, Andreas Stuhlmueller, et al. Alignment faking in large language models, 2024. URL https://arxiv.org/abs/2412.14093.

Roger Grosse, Juhan Bae, Cem Anil, Nelson Elhage, Alex Tamkin, Amirhossein Tajdini, Benoit Steiner, Dustin Li, Esin Durmus, Ethan Perez, et al. Studying large language model generalization with influence functions, 2023. URL https://arxiv.org/abs/2308.03296.

Evan Hubinger, Carson Denison, Jesse Mu, Mike Lambert, Meg Tong, Monte MacDiarmid, Tamera Lanham, Daniel M. Ziegler, Tim Maxwell, Newton Cheng, et al. Sleeper agents: Training deceptive LLMs that persist through safety training, 2024. URL https://arxiv.org/abs/2401.05566.

Dmitrii Krasheninnikov, Egor Krasheninnikov, Bruno Mlodozeniec, and David Krueger. Implicit meta-learning may lead language models to trust more reliable sources, 2024. URL https://arxiv.org/abs/2310.15047.

Rudolf Laine, Bilal Chughtai, Jan Betley, Kaivalya Hariharan, Jeremy Scheurer, Mikita Balesni, Marius Hobbhahn, Alexander Meinke, and Owain Evans. Me, myself, and AI: The situational awareness dataset (SAD) for LLMs, 2024. URL https://arxiv.org/abs/2407.04694.

Monte MacDiarmid et al. Natural emergent misalignment from reward hacking in production RL, 2025. URL https://arxiv.org/abs/2511.18397.

Samuel Marks et al. Auditing language models for hidden objectives, 2025. URL https://arxiv.org/abs/2503.10965.

Alexander Meinke and Owain Evans. Tell, don't show: Declarative facts influence how LLMs generalize, 2023. URL https://arxiv.org/abs/2312.07779.

Dillon Plunkett et al. Self-interpretability: LLMs can describe complex internal processes that drive their decisions, 2025. URL https://arxiv.org/abs/2505.17120.

Stewart Slocum et al. Believe it or not: How deeply do LLMs believe implanted facts?, 2025. URL https://arxiv.org/abs/2510.17941.

Cameron Tice et al. Alignment pretraining: AI discourse causes self-fulfilling (mis)alignment, 2026. URL https://arxiv.org/abs/2601.10160.

Johannes Treutlein, Dami Choi, Jan Betley, Cem Anil, Samuel Marks, Roger B. Grosse, and Owain Evans. Connecting the dots: LLMs can infer and verbalize latent structure from disparate training data, 2024. URL https://arxiv.org/abs/2406.14546.

Atticus Wang et al. Simple mechanistic explanations for out-of-context reasoning, 2025. URL https://arxiv.org/abs/2507.08218.

Boshi Wang, Xiang Yue, Yu Su, and Huan Sun. Grokked transformers are implicit reasoners: A mechanistic journey to the edge of generalization, 2024. URL https://arxiv.org/abs/2405.15071.

Jiaran Ye et al. How do transformers learn implicit reasoning?, 2025. URL https://arxiv.org/abs/2505.23653.