Logic Haystacks: Probing LLMs' Long-Context Logical Reasoning (Without Easily Identifiable Unrelated Padding)

Damien Sileo

Univ. Lille, Inria, CNRS, Centrale Lille, UMR 9189 - CRIStAL, F-59000 Lille, France

damien.sileo@inria.fr

Abstract

Large language models demonstrate promising long context processing capabilities, with recent models touting context windows close to one million tokens. However, the evaluations supporting these claims often involve simple retrieval tasks or synthetic tasks padded with irrelevant text, which the models may easily detect and discard. In this work, we generate lengthy simplified English text with first-order logic representations spanning up to 2048 sentences (≈ 25k GPT-4 tokens). We formulate an evaluation task with evidence retrieval for contradiction detection. The long, homogeneous text is filled with distractors that are both hard to distinguish from relevant evidence and provably non-interfering. Our evaluation of evidence retrieval shows that the effective context window is much smaller with realistic distractors, already crumbling at 128 sentences.

1 Introduction

Recent large language models (LLMs) can process long context windows, expanding the scope of their applications. However, the theoretical context length alone fails to capture a model's actual performance across varying input sizes (Liu et al., 2024). Several benchmarks have been developed to evaluate the actual capabilities of these models in reasoning over extended contexts.

Human-annotated benchmarks (Bowman et al., 2022; Wang et al., 2024a; Bai et al., 2024) are expensive to scale and constrained by both memory limitations and the finite attention span of annotators. On the other hand, synthetic datasets can be arbitrarily difficult, but they mainly use simple tasks (retrieval or reasoning) then drown the relevant inputs in unrelated text (Kamradt, 2023; Levy et al., 2024; Li et al., 2024). This stratagem is not only easy to counteract, as language models can detect the irrelevant input, but also risky, as some statements in BookCorpus or Paul Graham essays, which are routinely used, might interfere with the original problems. Generating realistic distractors as padding is desirable, but it increases the risk of semantic collisions that would perturb the task.

In this work, we use grammars to generate simplified English paired with logical representations to create long input text while controlling its semantics. We structure the generated expressions into PREMISE$_{[N]}$, HYPOTHESIS pairs, where the premise is a conjunction of N sentences. We ask models to identify which premise sentences contradict the hypothesis when taken together. We detect contradictions with a First-Order Logic (FOL) solver (Goodwin et al., 2020). A simple example is: PREMISE$_{[3]}$: *L0: Everyone who is happy is rich. L1: Mary is happy. L2: Nina is rich.* HYPOTHESIS: *Mary is not rich.*. CONTRADICTION EVIDENCE: L0, L1.

We use a previously proposed grammar for FOL with English correspondences to generate reasoning problems (Sileo, 2024), and we propose a methodology to scale premise generation to thousands of sentences. This is a challenging problem because naively scaling the problem causes paradoxes (e.g. *Mary is happy. Paul is not rich. [...] Mary is not happy.*) at rapidly growing rates. We also propose a method to isolate sufficient and necessary evidence that can be reliably used to probe an LLM for evidence identification. Our contributions are as follows: (i) a scalable algorithm to generate formally verified reasoning datasets to probe complex logical reasoning inside long contexts; (ii) a method to extract necessary and sufficient evidence, providing a logic task formulation to probe an LLM's ability to explain a contradiction; (iii) evaluation results for recent large language models, comparing padding with realistic distractors, and publicly available datasets[1].

[1] [data:HF-datasets 🤗]

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 2: Short Papers, pages 66–75

2 Generating Long-Context Logical Reasoning Data

One of the main applications of processing long inputs is piecing together relevant knowledge (e.g., during Retrieval Augmented Generation (Lewis et al., 2020)). Finding whether evidence contradicts a fact (the hypothesis) is a realistic use case that we model with a large premise expressing logical statements with some linguistic variety.

We construct each premise as a conjunction of N sentences paired with a logical representation. Then, for each premise, we search for single hypotheses that trigger contradictions[2]. Our first goal is to generate long, non-paradoxical premises. There are many strategies to scale the number of sentences while avoiding contradictions, like increasing the number of predicate and variable names (Monasson et al., 1999). However, we need the premise to contain many challenging distractors, so we want to have symbols that occur multiple times to prevent sentences from being easily discarded.

2.1 Satisfiable Merging

To maintain distractor difficulty with many sentences, we propose a simple method to generate non-contradictory premises that contain relatively few concepts. We use the Vampire (Reger et al., 2022) theorem prover to check whether a formula is satisfiable (i.e., not contradictory).

We start by generating K formulas of 32 sentences each (stage $i = 0$). We define a satisfiable merge as a conjunction of two formulas. We check if the conjunction is contradictory; if it is, we use the proof of the contradiction to identify the set of sentences causing it. We sample one of these sentences, remove it, and repeat the removals until the conjunction is satisfiable[3]. We construct stage $i + 1$ by randomly pairing K formulas from stage i and computing their satisfiable merge. Thus, the maximum formula size doubles at each stage. We continue until we get K premises of up to 4096 sentences.

[2]We focus on contradiction detection rather than entailment because we found that hypotheses not triggering a contradiction were relatively rare, occurring in less than 30% of cases during our generation process.

[3]MaxSAT algorithms could achieve this process more efficiently but are not available for full-fledged FOL. In addition, MaxSAT adds a maximality constraint that we do not need. We tried RC2 MaxSAT on propositional subsets and found it to be slower on our data.

2.2 Locating Evidence with Counterfactuals

The Vampire theorem prover provides a derivation listing specific sentences used to support a contradiction. However, this only provides a *sufficient* set of evidence, not a *necessary* one. There might be other sentences that also support the contradiction. To identify necessary evidence, we look at the sentences in the premises that are supporting a proof. For each piece of evidence e, we recompute whether the (PREMISE without e, HYPOTHESIS) pair is contradictory. If the pair remains contradictory, this means that e was not necessary for the contradiction. By doing this, we can keep only examples where all evidence is necessary. For these examples, evidence retrieval has a unique solution and is the well-formed task that we use to probe logical reasoning in lengthy input.

2.3 Logic Haystacks Dataset Construction

We use Unigram-FOL (Sileo, 2024), a parallel grammar-based generator of English expressions aligned to first-order logic expressions. It generates facts with independent atomic propositions, or with independent predicates, and conditionals based on facts. The Unigram-FOL grammar has shown higher transfer to the human-crafted FOLIO dataset (Han et al., 2022) and human-crafted formal semantics constructions (Richardson et al., 2020) when compared to other comparable grammars (RuleTaker (Clark et al., 2020), LogicNLI (Tian et al., 2021), and FLD (Morishita et al., 2023)).

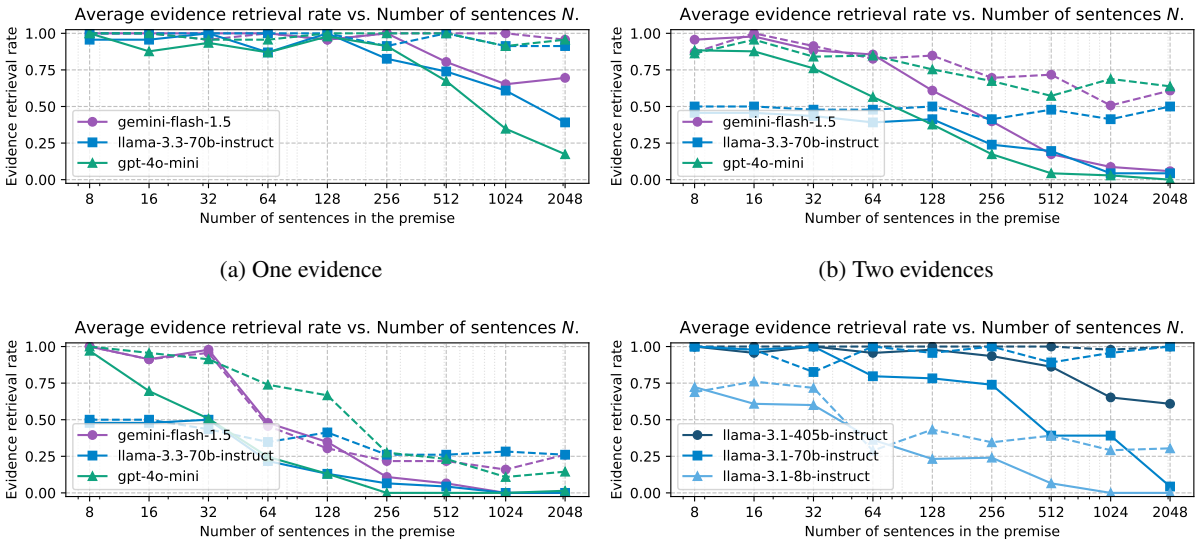
We use three times more names and predicates (78 instead of 26) to reduce the chance of semantic collisions at a manageable rate while having frequent symbol repetition.

We sample atomic predicates as hypotheses to avoid ambiguous constructs such as vacuous implications. We discard hypotheses that are in the premise with the same surface form to prevent superficial matching.

The generation process took 12 days on an Intel Xeon Gold 5320 CPU, parallelized over 52 threads.

We only consider the last stage with up to 4096 sentences, and we discard all premises with fewer than 2048 sentences. Then, to obtain shorter premises, we create sets of 8, 16,..., 2048 sentences by randomly sampling the distractors. This ensures that the proof structure statistics (depth, axioms used) do not depend on the premise length, enabling a controlled comparison.

We independently generate a total of 600 vali-



(a) One evidence

(b) Two evidences

(c) One evidence, not isolating the hypothesis

(d) One evidence, impact of model size

Figure 1: Evidence retrieval on Logic Haystacks for 8 to 2048 sentences. For each model, the dashed line shows the accuracy when we replace the grammar-generated distractor sentences with random sentences from Paul Graham essays.

dation examples and 2790 test examples (starting with K=2000 and K=10000 respectively) including contradiction, entailment, and neutral labels. We sample 200 examples for each variation of the test set, focusing on contradiction examples with 1, 2, or 3 necessary and sufficient pieces of evidence.

> PROMPT TEMPLATE: "Premise: {p} Hypothesis:{h} Given the premise, find the {k} line identifiers explaining why the hypothesis is logically contradicted by the premise. Answer directly with no explanation using comma-separated line identifiers between <answer> and </answer> tags. Format illustration: "<answer>L42</answer>".).

We construct the PROMPT TEMPLATE, iterating on Llama-3.1-8B-Instruct with the 2-evidence validation data until we obtained a satisfactory output format. We picked this model for prompt selection because it is the weakest model we used, and all other models conformed to the output format upon inspection and automated verifications.

Chain-of-thought prompting did not drastically change the results in our early experiments. We parse the output automatically to compute the evidence retrieval rate for each example, evaluating retrieval accuracy with the Jaccard similarity metric between predicted evidence and ground truth.

We evaluate off-the-shelf instruction-tuned language models via their APIs. We evaluate the Llama 3.1 family (FP8), Gemini 1.5-002 Flash, and GPT-4o-mini, with default hyperparameters. These models all achieve near-perfect Needle-In-Haystack accuracy with at least 32k tokens.

2.4 Results

Figure 1 presents retrieval scores across different configurations. We use dotted lines when reporting results using Paul Graham sentences[4] as distractors, following prior work (Kamradt, 2023; Li et al., 2024; Levy et al., 2024). We split the essays into sentences with the pysbd (Sadvilkar and Neumann, 2020) tokenizer.

One evidence (1a) Even when the proof only involves one piece of evidence, we see a high difference between the types of distractors. Gemini Flash performs much better than the compared models with realistic distractors, a fact that would not have been conspicuous with easier padding.

Two evidences (1b) With two pieces of evidence, the performance drops significantly for all models, as none of them obtain a retrieval rate above 20% with 1024 sentences.

Not isolating the hypothesis (1c) We also insert the hypothesis at a random position inside the

[4] hf.co/datasets/sgoel9/paul_graham_essays

premise and remove it from the prompt[5]. This makes it much harder to notice logical contradictions. Interestingly, this task is still hard even with the easily identifiable distractors.

Model size (1d) We evaluate the Llama 3.1 model family (8B, 70B, 405B) to assess the effect of model size in a controlled experiment. While model size helps, it is not sufficient for robust performance even with one distractor, revealing potentially fundamental issues in this generation of models.

2.5 Qualitative Error Analysis

We show randomly sampled error examples in Appendix B. We notice that incorrectly predicted evidence is often lexically related to the necessary premises without constituting a logical contradiction. This suggests that models struggle to distinguish semantic similarity from logical necessity in a noisy context, even when the underlying logical contradiction is straightforward ('clear-cut'). When using Paul Graham essays as padding, models often select them as evidence, though they do not form valid proofs.

3 Related work

Our work stands at the intersection of two lines of research: logical reasoning with synthetic simplified English data, and context length evaluation. Levy et al. (2024) and Kuratov et al. (2024) explored this intersection, but they imported the padding technique from previous work on LLM stress-testing, while we scale the dataset's original generation process instead of padding it with other text. Human-annotated long-context benchmarks are very valuable but hard to annotate (Bowman et al., 2022; Wang et al., 2024a; Bai et al., 2024), which causes current language models to saturate them.

Synthetic datasets for reasoning Numerous works investigate the logical capabilities of NLP models using textual datasets and symbolic reasoning (Helwe et al., 2022). We focus on grammar-derived synthetic datasets. RuleTaker (Clark et al., 2020), LogicNLI (Tian et al., 2021), FLD (Morishita et al., 2023), and Unigram-FOL (Sileo, 2024) address different subsets of first-order logic with English translations. Other works also explore non-standard logic with synthetic datasets, notably probabilistic (Jin et al., 2023; Sileo and Moens, 2023), paraconsistent (Kazemi et al., 2024), and epistemic (Sileo and Lernould, 2023) logics.

These approaches focus on input sizes typically suitable for a standard BERT (Devlin et al., 2018) encoder (<512 tokens). Here, we push the number of expressions in the input while avoiding paradoxes. This is related to the satisfiability problem, which was explored by Richardson et al. (2020); Richardson and Sabharwal (2022) who use a solver to study satisfiability in natural language on constrained problems. However, they also focus on relatively moderate text sizes, while we use satisfiability checking as a stepping stone to generate large texts and not only as a task in itself.

LLM context length stress tests Our work is also related to context window stress testing. The Long-Range Arena (Tay et al., 2021) provides the first systematic analysis of the long-range processing capabilities of text encoders, focusing mainly on algorithmic reasoning and retrieval tasks. Needle in a Haystack benchmarks (Kamradt, 2023; Li et al., 2024) test longer window sizes with simple retrieval tasks and use Paul Graham essays as padding. Sileo (2025) reverses the retrieval task by asking for missing items, which makes effective context length much shorter. BABILong (Kuratov et al., 2024) uses bAbi (Weston et al., 2016) reasoning tasks and interleaves relevant text with irrelevant input from BookCorpus (Zhu et al., 2015). FlenQA (Levy et al., 2024) applies a similar process to the RuleTaker (Clark et al., 2020) deductive logical reasoning task and uses Paul Graham essays as padding. Ruler uses simple algorithmic tasks like variable tracking and word counting. They also use Paul Graham essays as noise, or repetitions of sentences such as *The sky is blue*, following Mohtashami and Jaggi (2024). The MuSR dataset uses GPT-4 generated (Sprague et al., 2023) problems, which makes it hard to verify the problems' integrity at scale. Ling et al. (2025) uses a similar method but grounds it in human-collected reasoning questions. In the mathematical domain MATHHAY (Wang et al., 2024b) creates long-context mathematical benchmarks by embedding problems within unrelated real-world documents, ensuring validity by requiring an LLM to independently reproduce the exact same answer twice.

[5]We rephrase the prompt with *Find {k} line identifiers that logically contradict each other within the premise.*

4 Conclusion

Evaluation datasets are critical to guide language model construction. We proposed methods to scale logic datasets for long-context probing. We confirm that using external text as padding leads to over-estimating the context window and propose a new dataset, Logic Haystacks, that can provide a more realistic signal to evaluate long-context processing architectures. Further work is needed to scale the generation process for training data and to evaluate whether training on long-context logical reasoning synthetic data (using a less expressive logic to scale data construction more easily, because going beyond 4096 sentences starts being intractable with our method) leads to transferable gains on Logic Haystacks.

Limitations

While our work provides valuable insights into long-context logical reasoning, several limitations should be noted. The generation process is computationally expensive, taking 12 days even with parallelization, which limits the scale of datasets that can be produced. While our method uses simplified English and a constrained subset of first-order logic to enable formal verification, this may not fully capture the complexity and nuance of natural language reasoning. The focus on contradiction detection rather than other logical relations like entailment or equivalence narrows the scope of evaluation. Additionally, our zero-shot evaluation on a limited set of models with default hyperparameters may not fully represent the potential of these systems, particularly under different prompting strategies or with tuned parameters. Finally, while the lack of human performance benchmarks could be seen as a limitation, it's worth noting that our task deliberately pushes beyond typical human cognitive limits, as processing thousands of interrelated logical statements is precisely the kind of task where we expect machines to surpass human capabilities.

Acknowledgement

This work was supported by the French National Research Agency (ANR) through the ANR-24-CE23-4637 grant (Adada project).

References

Yushi Bai, Shangqing Tu, Jiajie Zhang, Hao Peng, Xiaozhi Wang, Xin Lv, Shulin Cao, Jiazheng Xu, Lei Hou, Yuxiao Dong, et al. 2024. Longbench v2: Towards deeper understanding and reasoning on realistic long-context multitasks. *arXiv preprint arXiv:2412.15204*.

Samuel R Bowman, Angelica Chen, He He, Nitish Joshi, Johnny Ma, Nikita Nangia, Vishakh Padmakumar, Richard Yuanzhe Pang, Alicia Parrish, Jason Phang, et al. 2022. Quality: Question answering with long input texts, yes! *NAACL 2022*.

Peter Clark, Oyvind Tafjord, and Kyle Richardson. 2020. Transformers as soft reasoners over language. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 3882–3890. International Joint Conferences on Artificial Intelligence Organization. Main track.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Emily Goodwin, Koustuv Sinha, and Timothy J. O'Donnell. 2020. Probing linguistic systematicity. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1958–1969, Online. Association for Computational Linguistics.

Simeng Han, Hailey Schoelkopf, Yilun Zhao, Zhenting Qi, Martin Riddell, Luke Benson, Lucy Sun, Ekaterina Zubova, Yujie Qiao, Matthew Burtell, et al. 2022. Folio: Natural language reasoning with first-order logic. *arXiv preprint arXiv:2209.00840*.

Chadi Helwe, Chloé Clavel, and Fabian Suchanek. 2022. Logitorch: A pytorch-based library for logical reasoning on natural language. In *The 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*.

Zhijing Jin, Yuen Chen, Felix Leeb, Luigi Gresele, Ojasv Kamal, LYU Zhiheng, Kevin Blin, Fernando Gonzalez Adauto, Max Kleiman-Weiner, Mrinmaya Sachan, et al. 2023. Cladder: Assessing causal reasoning in language models. In *Thirty-seventh conference on neural information processing systems*.

Greg Kamradt. 2023. Llmtest_needleinahaystack. GitHub repository.

Mehran Kazemi, Quan Yuan, Deepti Bhatia, Najoung Kim, Xin Xu, Vaiva Imbrasaite, and Deepak Ramachandran. 2024. Boardgameqa: A dataset for natural language reasoning with contradictory information. *Advances in Neural Information Processing Systems*, 36.

Yuri Kuratov, Aydar Bulatov, Petr Anokhin, Ivan Rodkin, Dmitry Sorokin, Artyom Sorokin, and Mikhail Burtsev. 2024. Babilong: Testing the limits of llms with long context reasoning-in-a-haystack. *arXiv preprint arXiv:2406.10149*.

Mosh Levy, Alon Jacoby, and Yoav Goldberg. 2024. Same task, more tokens: the impact of input length on the reasoning performance of large language models. *arXiv preprint arXiv:2402.14848*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Mo Li, Songyang Zhang, Yunxin Liu, and Kai Chen. 2024. Needlebench: Can llms do retrieval and reasoning in 1 million context window? *arXiv preprint arXiv:2407.11963*.

Zhan Ling, Kang Liu, Kai Yan, Yifan Yang, Weijian Lin, Ting-Han Fan, Lingfeng Shen, Zhengyin Du, and Jiecao Chen. 2025. Longreason: A synthetic long-context reasoning benchmark via context expansion.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Amirkeivan Mohtashami and Martin Jaggi. 2024. Random-access infinite context length for transformers. *Advances in Neural Information Processing Systems*, 36.

Rémi Monasson, Riccardo Zecchina, Scott Kirkpatrick, Bart Selman, and Lidror Troyansky. 1999. Determining computational complexity from characteristic 'phase transitions'. *Nature*, 400(6740):133–137.

Terufumi Morishita, Gaku Morio, Atsuki Yamaguchi, and Yasuhiro Sogawa. 2023. Learning deductive reasoning from synthetic corpus based on formal logic. In *International Conference on Machine Learning*, pages 25254–25274. PMLR.

Giles Reger, Martin Suda, Andrei Voronkov, Laura Kovács, Ahmed Bhayat, Bernhard Gleiss, Marton Hajdu, Petra Hozzova, JR Evgeny Kotelnikov, Michael Rawson, et al. 2022. Vampire 4.7-smt system description.

Kyle Richardson, Hai Hu, Lawrence Moss, and Ashish Sabharwal. 2020. Probing natural language inference models through semantic fragments. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34:8713–8721.

Kyle Richardson and Ashish Sabharwal. 2022. Pushing the limits of rule reasoning in transformers through natural language satisfiability. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11209–11219.

Nipun Sadvilkar and Mark Neumann. 2020. PySBD: Pragmatic sentence boundary disambiguation. In *Proceedings of Second Workshop for NLP Open Source Software (NLP-OSS)*, pages 110–114, Online. Association for Computational Linguistics.

Damien Sileo. 2024. Scaling synthetic logical reasoning datasets with context-sensitive declarative grammars. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5275–5283, Miami, Florida, USA. Association for Computational Linguistics.

Damien Sileo. 2025. Attention overflow: Language model input blur during long-context missing items identification. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 761–767, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

Damien Sileo and Antoine Lernould. 2023. MindGames: Targeting theory of mind in large language models with dynamic epistemic modal logic. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4570–4577, Singapore. Association for Computational Linguistics.

Damien Sileo and Marie-francine Moens. 2023. Probing neural language models for understanding of words of estimative probability. In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, pages 469–476, Toronto, Canada. Association for Computational Linguistics.

Zayne Sprague, Xi Ye, Kaj Bostrom, Swarat Chaudhuri, and Greg Durrett. 2023. Musr: Testing the limits of chain-of-thought with multistep soft reasoning. *arXiv preprint arXiv:2310.16049*.

Yi Tay, Mostafa Dehghani, Samira Abnar, Yikang Shen, Dara Bahri, Philip Pham, Jinfeng Rao, Liu Yang, Sebastian Ruder, and Donald Metzler. 2021. Long range arena : A benchmark for efficient transformers. In *International Conference on Learning Representations*.

Jidong Tian, Yitian Li, Wenqing Chen, Liqiang Xiao, Hao He, and Yaohui Jin. 2021. Diagnosing the first-order logical reasoning ability through LogicNLI. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3738–3747, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Cunxiang Wang, Ruoxi Ning, Boqi Pan, Tonghui Wu, Qipeng Guo, Cheng Deng, Guangsheng Bao, Qian Wang, and Yue Zhang. 2024a. Novelqa: A benchmark for long-range novel question answering. *arXiv preprint arXiv:2403.12766*.

Lei Wang, Shan Dong, Yuhui Xu, Hanze Dong, Yalu Wang, Amrita Saha, Ee-Peng Lim, Caiming Xiong, and Doyen Sahoo. 2024b. Mathhay: An automated benchmark for long-context mathematical reasoning in llms. *arXiv preprint arXiv:2410.04698*.

Jason Weston, Antoine Bordes, Sumit Chopra, and Tomás Mikolov. 2016. Towards ai-complete question

answering: A set of prerequisite toy tasks. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *The IEEE International Conference on Computer Vision (ICCV)*.

A Appendix A - Prompt example

Premise:
L0: if someone is richer than Jennifer then he/she is a happy person and vice versa
L1: Susan and Delbert hate each other
L2: more than one person in the room is a romantic person
L3: ""Mirrors do not fog permanently in one particular house." or "A tower does not lean significantly without falling." or both" if "Dana is quieter than Beatrice" and vice versa
L4: not everyone in the room who frequently participates in hackathons and coding competitions does not own a smart tv
L5: Walter is not a brave person
L6: Leonard is younger than Justin
L7: Dcalvina and Susan hate each other
L8: someone in the room does not hate Benjamin
L9: Jason is a client of John
L10: Stephen is allergic to anything
L11: Dorothy is a client of Elsie
L12: Janna neither is a client of Ramon nor is older than Ossie
L13: Dorothy is not an avid collector of autographed memorabilia from famous musicians
L14: everyone in the room who does not enjoy mountain biking hosts a popular podcast about emerging technologies
L15: not everyone in the room does not collect vintage vinyl records or is a curious blue eyed tourist
L16: Dorothy, Brian, Natividad, Natasha, Jason, Michael, Napoleon, Dcalvina, Melissa, Justin, Chae, Charlette, Rex, Pamela, Janna, Elsie, Calvin, Elizabeth, Bernard, Michelle, Delbert, Dana, Caitlin, Alice, Beatrice, Brandon, James, Joseph, Francis, Brenda, Chae, Virginia, Dionne, Louise, Christopher, Nelson, Walter, Ramon, Carol, Philomena, Raymonde, Stephen, Carla, Shirley are the only persons in the room.
L17: everyone in the room who is not a night owl is not liked by Gary
L18: everyone outside the room is a tall curly haired strong tourist if they is a cybersecurity expert
L19: at least two persons in the room is a colorblind european
L20: someone in the room does not hate Jennifer
L21: Rex and John hate each other
L22: Shirley is younger than Steven
L23: Justin is quieter than Genevieve
L24: "A tower leans significantly but never falls." or "Glass rain falls on a distant planet." but not both
L25: everyone in the room is younger than Jeannine if they is not liked by Benjamin and vice versa
L26: it is not the case that "Michelle is richer than Shirley"
L27: at least one person in the room is liked by Lisa
L28: everyone outside the room enjoys rooftop gardening if they owns a high-end gaming PC with custom-built components
L29: everyone in the room who is younger than Genevieve owns a 3D printer

```
L30: if “it is not the case that “Shirley is a sibling of Genevieve”” then ““Mirrors fog permanently in one particular house.” and “A city has not outlawed the use of round tables.””
L31: Charlette and John like each other.
Hypothesis:
Dorothy is an avid collector of autographed memorabilia from famous musicians

Given the premise, find the 1 line identifier explaining why the hypothesis is contradicted by the premise. Answer directly with no explanation and only with comma-separated line ids, e.g., "L0,L3."
```

GPT-4o-mini prediction: "L12".
Correct answer: "L13."
We note that at this scale, most models solve the task very well.

B Appendix B - Error examples

Error examples using Gemini 1.5 Flash with 2048 sentences:

Hypothesis:
David and Christopher hate each other
Predicted Evidence:
L266: Kathleen is a sibling of Tamara
Ground truth:
L126: Christopher and David like each other

Hypothesis:
Marvin is a formal european
Predicted Evidence:
L26: Michelle and Marvin hate each other
Ground truth:
L587: Marvin is not curious, not formal

Hypothesis:
Jewell travels domestically frequently
Predicted Evidence (Paul Graham Distractors):
L299: Once you start considering this question, you have opened a real can of worms.
Ground truth:
L729: Jewell does not travel domestically frequently

Hypothesis:
Donald is a sibling of Michael
Predicted Evidence (Paul Graham Distractors):
L80: Michael does enjoy trail running
L87: Certainly schools should teach students how to write.
Ground truth:
L80: Michael does enjoy trail running
L987: Michael either is a sibling of Donald or does enjoy trail running but not both