Zero-Shot Belief: A Hard Problem for LLMs

John Murzaku[♦♠], Owen Rambow[♣♠]
[♦]Department of Computer Science [♣]Department of Linguistics
[♠]Institute for Advanced Computational Science
Stony Brook University
jmurzaku@cs.stonybrook.edu

Abstract

We present two LLM-based approaches to zero-shot source-and-target belief prediction on FactBank: a unified system that identifies events, sources, and belief labels in a single pass, and a hybrid approach that uses a fine-tuned DeBERTa tagger for event detection. We show that multiple open-sourced, closed-source, and reasoning-based LLMs struggle with the task. Using the hybrid approach, we achieve new state-of-the-art results on FactBank and offer a detailed error analysis. Our approach is then tested on the Italian belief corpus ModaFact.

1 Introduction

The term "belief" (interchangeably referred to as "event factuality" in NLP) refers to the extent an event mentioned by the author or by sources in a text is presented as being factual. While this task has received attention over the years, no zero-shot experiments have been performed. We show that this task remains a hard task for LLMs.

Our major contributions are as follows:
(1) We present unified and hybrid zero-shot frameworks for the source-and-target belief prediction task (i.e., who has what belief towards what). We test our approach on various LLMs.
(2) Our hybrid approach achieves new state-of-the-art results (SOTA) on the FactBank corpus, but the problem is far from solved.
(3) We are the first to evaluate FactBank on Nested belief, revealing that LLMs perform particularly poorly on this task. We perform an error analysis showcasing where LLMs fail.
(4) We validate the transferability of our approach by testing on the Italian ModaFact belief corpus.

This paper is organized as follows: we provide an overview of the belief detection task in Section 2. We follow by detailing our methodology in Section 4 and discuss our results and analysis for FactBank and ModaFact in Section 5.

2 Related Work

Corpora Many corpora explore the notion of belief on the sentence level. FactBank is one of the first corpora to do this, annotating source-and-target belief: both the belief presented by the author towards an event and the belief towards events by sources mentioned inside of the text (Saurí and Pustejovsky, 2009). Other corpora annotate only the author's belief towards events: these corpora include LU (Diab et al., 2009), UW (Lee et al., 2015), LDCCB (Prabhakaran et al., 2015), MEANTIME (Minard et al., 2016), MegaVeridicality (White et al., 2018), UDS-IH2 (Rudinger et al., 2018), CommitmentBank (De Marneffe et al., 2019), and RP (Ross and Pavlick, 2019). Two recent corpora for event factuality are Maven-Fact (Li et al., 2024) which contains a large-scale corpus of event and supporting evidence annotations, and ModaFact (Rovera et al., 2025), which is an Italian author belief corpus that annotates in a similar style and inspiration as FactBank.

Methods Previous methods for author belief prediction mainly involve fine-tuning: Rudinger et al. (2018) fine-tune multi-task LSTMs; Pouran Ben Veyseh et al. (2019) fine-tune a graph convolutional network with BERT (Devlin et al., 2019) representations; Jiang and de Marneffe (2021); Murzaku et al. (2022) fine-tune RoBERTa (Liu, 2019) with span representations; Li et al. (2024) fine-tune RoBERTa and Flan-T5 (Chung et al., 2024), and also explore four LLMs predictions using few-shot learning; Rovera et al. (2025) fine-tune BERT, mT5-XXL (Xue et al., 2021), Aya23-8B (Aryabumi et al., 2024), and Minerva-3B (Orlando et al., 2024).

There has been much less focus on the complete source-and-target belief task: Murzaku et al. (2023); Murzaku and Rambow (2024) both fine-tune a Flan-T5 model, with the latter optimizing for the structure of belief represented as a tree.

Proceedings of the 6th Workshop on Computational Approaches to Discourse, Context and Document-Level Inferences (CODI 2025), pages 107–119
November 9, 2025 ©2025 Association for Computational Linguistics

3 Preliminaries

Consider this sentence: *Trurit Inc. said it is phasing out legacy routers.* This sentence reports on two events: a "said" event and a "phasing" event.

Author Belief The definition of author belief (also called event factuality) is how committed is the author of the text (the source) to the truth (or factuality) of an event. In this sentence, the author is presenting the "said" event as factual, i.e., they are committed to the "said" event having happened. On the other hand, the author is presenting the "phasing" event as having an unknown factuality; the author is not directly committing to the truth of the event, rather they are reporting on what "Trurit Inc." said.

Nested In nested belief, we report the belief towards events according to nested sources inside of a text. The task can be split into three steps: (i) identifying the nested or attributed source in the text; (ii) linking the source to the events (i.e., which events does the source commit to); (iii) labelling the belief of the event according to that source. In our example, the source is "Trurit Inc." Once the source is introduced (i), we then link the source to the events in the text (ii): in this case, Trurit Inc. is reportedly committing to the event "phasing", and asserting it as true (iii). Since the source is reporting about this event, and directly committing to the event happening, it is therefore true in Trurit Inc.'s perspective as reported by the author, and unknown in the author's perspective.

4 Methodology

We use the test set of the source-and-target (author and nested sources) projection of FactBank released by Murzaku and Rambow (2024). Further dataset details are in Appendix D.

4.1 Zero-Shot

Unified Our **Unified** approach provides a single end-to-end zero-shot prompt to the LLM with the input text, a high-level descriptions of the task, the three main steps in the annotation process in detail, special cases guidelines, and the output format. We end the prompt with a summary of the specific steps on how to produce the final answer in a chain-of-thought format (CoT) (Wei et al., 2022), which has proven to work well for author event factuality (Li et al., 2024). The three steps are: (1) Label all events according to the FactBank annotation guidelines, which we provide. (2) Identify all nested sources in the text. (3) Assign factuality labels for each event, according to that source. We leave all model details, API parameters, and our exact prompts in Appendix A.

Hybrid For our **Hybrid** zero-shot approach, we first extract events in a sentence using a DeBERTa (He et al., 2021) based tagger. After extracting the events, we prompt an LLM with the sentence and the list of events. We then follow the exact steps (minus event detection, since we provide events) as our **Unified** prompt: we instruct the LLM to identify all nested sources, ask the LLM to assign factuality labels for the events, according to the identified sources, and finish with instructions for answering with CoT. See Appendix B for further details on our **Hybrid** experiments and our exact prompts.

Event Tagger The FactBank corpus has a complex definition of what exactly is an annotatable event. Murzaku et al. (2023) found that annotating FactBank events is non-trivial, even with a specialized, generative fine-tuned model achieving only 85.4% F1 on event identification. We therefore choose to fine-tune a DeBERTa model for event token detection, and then pass the events to our **Hybrid** prompts.

4.2 Models

We perform experiments on a variety of LLM types: open LLMs, specifically LLaMA-3.3-70B (Meta, 2024), DeepSeek-v3 (Liu et al., 2024), and DeepSeek-r1; closed LLMs, specifically GPT-4o (OpenAI, 2024a), newly released reasoning models o1 (OpenAI, 2024b) and o3-mini (OpenAI, 2025), and Claude 3.5 Sonnet (Anthropic, 2025); and reasoning LLMs, DeepSeek r1 (henceforth R1), o1, and o3-mini.

4.3 Evaluation: Metrics

We evaluate on three F1 metrics: `Full` where we perform an exact match evaluation on all generated (source, event, label) annotations; `Author` where we perform an evaluation on all generated annotations where the source is the author of the text; and `Nest` where we perform an exact match evaluation on all generated annotations where the source is a nested source.

4.4 Evaluation: FactBank Sources

FactBank has specific conventions about annotating sources. Consider the example "Trurit Inc. shares rose by 5% today". FactBank annotates on the

Model	Unified			Hybrid			Δ Hyb.-SOTA		Δ Hyb.-Unif.		
	Full	**Author**	**Nest**	**Full**	**Author**	**Nest**	**Full**	**Author**	**Full**	**Author**	**Nest**
Previous / Fine-Tuned SOTA (Murzaku and Rambow, 2024)											
GPT-3 (Fine-tuned)	65.8	76.0	–	65.8	76.0	–	–	–	–	–	–
Flan-T5-XL	69.5	76.6	–	69.5	76.6	–	–	–	–	–	–
Zero-Shot LLM Systems											
GPT-4o	60.2	65.9	20.2	68.7	73.2	22.9	-0.8	-3.4	+8.5	+7.3	+2.7
o1 🧠	65.0	**73.2**	18.9	70.3	**78.9**†	19.2	+0.8	+2.3	+5.3	+5.7	+0.3
DeepSeek r1 🧠 🔓	**66.1**	71.1	**24.1**	**72.0**†	77.6	**25.3**†	+2.5	+1.0	+5.9	+6.5	+1.2
o3-mini 🧠	62.4	70.9	15.6	65.5	75.2	17.0	-4.0	-1.4	+3.1	+4.3	+1.4
Claude 3.5	63.2	69.7	19.7	70.4	77.6	21.4	+0.9	+1.0	+7.2	+7.9	+1.7
LLaMA 3.3 🔓	53.1	60.4	14.4	58.8	66.0	19.9	-10.7	-10.6	+5.7	+5.6	+5.5
DeepSeek-v3 🔓	56.3	61.4	17.1	60.5	65.3	18.2	-9.0	-11.3	+4.2	+3.9	+1.1

Table 1: **Unified** vs. **Hybrid** approaches with different LLMs. We report Micro F1 (Full), Author Micro F1 (Author), and Nested Micro F1 (Nest) scores (in %). Δ **Hyb.-SOTA** denotes the difference between the **Hybrid** result vs. the fine-tuned SOTA. The best scores are highlighted in **bold** and new state-of-the-art (SOTA) results are denoted by †. Δ **Hyb.-Unif.** highlight the **Hybrid-Unified** difference for Full, Author, and Nest F1s. 🔓 indicates open models and 🧠 indicates reasoning models.

token level, and the source is "Inc.". We do not wish to penalize LLMs for not knowing this conversion, and therefore propose a few-shot normalization technique for postprocessing. We perform all source normalization experiments with GPT-4o (OpenAI, 2024a). Exact prompts for our source normalization methods and a detailed ablation study are shown in Appendix E.

Our task setup is as follows: Given a predicted source, we prompt GPT-4o to transform the predicted source into a FactBank-compliant version.

5 Results and Analysis

5.1 FactBank

Main Results Our main results for FactBank are shown in Table 1. We compare all our results to the previous fine-tuned SOTA from Murzaku and Rambow (2024), evaluating on exact match F1 (`Full`) and author exact match F1 (`Author`) as described in Section 4.3. We add one more metric: nested exact match F1 (`Nest`), where we evaluate on nested sources only.

Our **Unified** zero-shot results (column **Unified**) achieve competitive performance compared to fully fine-tuned models, with `R1` (66.1% for `Full`) and o1 (73.2% for `Author`). We outperform the fine-tuned GPT-3 model from Murzaku and Rambow (2024) on `Full`, but do not outperform the Flan-T5-XL system.

We achieve new SOTA on FactBank with our **Hybrid** systems. Our `R1` **Hybrid** system achieves `Full` of 72.0%, outperforming the previous state of the art by 2.5% (column Δ vs. SOTA). Similar to the **Unified** results, o1 excels in `Author`, achieving 78.9% `Author` and outperforming the previous SOTA by 2.3%. We also note that GPT-4o and Claude-3.5 also achieve competitive performance, with Claude-3.5 outperforming the previous SOTA on `Full` and `Author` by 0.9% and 1.0% respectively. We hypothesize that these models excel due to CoT prompting.

Nest F1 We are the first to provide `Nest` F1 metrics on FactBank. Our top performing model is r1, which achieves a nested F1 of 25.3%. For reasoning models o1, o3-mini, and r1, we notice that going from **Unified** to **Hybrid** does not increase `Nest` F1 dramatically (0.3% for o1, 1.4% for o3-mini, and 1.2% for r1), showcasing the models' lack of capabilities for nested belief predictions. We note that these results are low, and believe modelling of nested beliefs is essential future work and a challenging task for reasoning LLMs.

Zero vs. Hybrid We quantify the exact difference (in %) between our **Unified** and **Hybrid** models in Table 1 (column Δ (Hyb.-Unif.)). We see improvements in every model, with the greatest improvements occuring in GPT-4o and Claude-3.5 for `Full` and `Author`. On average, our **Hybrid** models

Model	Type	F1
DeBERTa	Fine-tuned	**89.0**
DeepSeek R1	Zero-shot	82.0
	Few-shot	76.4
GPT-4o	Zero-shot	78.2
	Few-shot	81.1
Claude 3.5	Zero-shot	83.3
	Few-shot	81.8

Table 2: Event detection performance (in % F1) of various language models. The fine-tuned DeBERTa model outperforms all major LLMs in zero-shot and few-shot settings.

outperform our **Unified** models by 5.7% for `Full`, 5.9% for `Author`, and 2.0% for `Nest`. Our results emphasize the need for a specialized event tagger and hybrid approach, allowing the LLM to focus on linking sources and tagging belief labels.

Event Detection We investigate how LLMs perform on event tagging. We show these results in Table 2. We compare three LLMs (r1, GPT-4o, and Claude-3.5) to the fine-tuned DeBERTa event tagger used in the **Hybrid** system. For our LLMs, we try two configurations: zero-shot and few-shot (5 examples). We find that a fine-tuned DeBERTa outperforms all LLMs in all settings, emphasizing that event detection is still a difficult task. We leave further experimental details and prompts in Appendix F.

Error Analysis We perform an error analysis on the top-performing model (R1, **Hybrid**) on nested beliefs (F1 of only 25.3%). We categorize errors as follows: **(1) Source** mismatch, often labeling the author instead of the nested source or failing to classify pronoun sources such as "it" correctly (123 errors); **(2) FN** (false negatives on events), where context-dependent event nouns or verbs are missed (e.g., *"acquisition," "construction"*) (77 errors); **(3) FP** (false positives on events), over-predicting event nouns (73 errors); **(4) Label** errors, notably predicting *True* or *Probable* instead of *Unknown* for future/uncommitted events (e.g., "Mary offered to **buy** an apple", where the **buy** event should be *Unknown*) (53 errors). We note that the FN errors are consistent with findings from prior FactBank studies: Murzaku et al. (2022) also found similar errors. More detailed results and analysis of our error analysis are in Appendix G.

Model	Method	Bel.+Pol.
mT5 XXL	Fine-tune	**64.4**
DeepSeek r1	Hybrid	63.6†
o3-mini	Hybrid	62.6†
GPT-4o	Hybrid	61.2†
GPT-4o	Unified	42.9
o3-mini	Unified	40.8
DeepSeek r1	Unified	38.6

Table 3: Model performance on Belief+Polarity (Bel.+Pol.) F1. Rovera et al. (2025) mT5-XXL baseline is shown in **bold** . Results on Bel.+Pol. metric within 5% of the SOTA are marked with a †.

5.2 Multilingual Belief

The ModaFact Italian corpus (Rovera et al., 2025) annotates the author's belief, polarity, and modality towards events and temporal information. We only use the belief and polarity annotations and combine these to tags similar to those of FactBank (and perform an exact match evaluation on Belief+Polarity F1). This is different from how Rovera et al. (2025) evaluate, but they kindly shared their raw results so that we could apply our evaluation.

Results We perform our ModaFact experiments with three cost-effective models that performed well for FactBank: GPT-4o, o3-mini, and R1. Our results are shown in Table 3. Unlike our FactBank results, we fall short of the fine-tuned SOTA for our **Hybrid** system (by 0.8%). Similar to our FactBank results, **Hybrid** strongly outperforms **Unified** in all settings. Finally, we see that R1 and o3-mini (both reasoning models) come very close to the fine-tuned SOTA. GPT-4o also proves competitive, but falls short of the reasoning models by 2.4%. We note that while we do not beat the SOTA, the LLMs we use are not explicitly trained for multilingual data (in contrast with mT5-XXL). For example, R1 is specifically optimized for English and Chinese data (Guo et al., 2025). We hypothesize that future multilingual optimizations for these reasoning LLMs would in fact lead to a new SOTA for the ModaFact corpus.

6 Conclusion

We show that belief detection from text remains a challenging problem for LLMs. This is particularly true for nested beliefs, which the author ascribes to other sources. Our new SOTA system includes a distinct fine-tuned event detection component.

Limitations

While our model achieves a new state-of-the-art on the English only FactBank, our results, while still competitive, do not perform as well for the Italian ModaFact corpus. We acknowledge this as a shortcoming and aim to work towards broader multilingual generalization for this task.

We note that our LLM approach yields poor results on the nested F1 metric, indicating a large gap and potential for future improvement. We will explore improving these results in future work and believe this to be a gap for all major open, closed, and reasoning LLMs.

Finally, we note that our top performing LLM approach, while using the open DeepSeek r1 model, is reliant on API calls for the source normalization technique. We attempt to minimize costs by using GPT-4o, but note that we can (i) achieve better performance using a larger, reasoning model (more cost) or (ii) switch to an open model. We will explore both techniques.

Ethics Statement

We note that our paper is foundational research and we are not tied to any direct applications. We do not foresee any potential risks with our work. We do not perform any annotations or human evaluation as we use the already existing FactBank dataset and ModaFact dataset.

References

Anthropic. 2025. Claude 3.5 sonnet. https://www.anthropic.com/news/claude-3-5-sonnet.

Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Jon Ander Campos, Yi Chern Tan, et al. 2024. Aya 23: Open weight releases to further multilingual progress. *arXiv preprint arXiv:2405.15032*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Marie-Catherine De Marneffe, Mandy Simons, and Judith Tonhauser. 2019. The commitmentbank: Investigating projection in naturally occurring discourse. In *proceedings of Sinn und Bedeutung*, volume 23, pages 107–124.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Mona Diab, Lori Levin, Teruko Mitamura, Owen Rambow, Vinodkumar Prabhakaran, and Weiwei Guo. 2009. Committed belief annotation and tagging. In *Proceedings of the Third Linguistic Annotation Workshop (LAW III)*, pages 68–73, Suntec, Singapore. Association for Computational Linguistics.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

Nanjiang Jiang and Marie-Catherine de Marneffe. 2021. He thinks he knows better than the doctors: BERT for event factuality fails on pragmatics. *Transactions of the Association for Computational Linguistics*, 9:1081–1097.

Kenton Lee, Yoav Artzi, Yejin Choi, and Luke Zettlemoyer. 2015. Event detection and factuality assessment with non-expert supervision. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1643–1648, Lisbon, Portugal. Association for Computational Linguistics.

Chunyang Li, Hao Peng, Xiaozhi Wang, Yunjia Qi, Lei Hou, Bin Xu, and Juanzi Li. 2024. MAVEN-FACT: A large-scale event factuality detection dataset. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11140–11158, Miami, Florida, USA. Association for Computational Linguistics.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Yinhan Liu. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 364.

Meta. 2024. Llama-3.3. https://www.llama.com/docs/model-cards-and-prompt-formats/llama3_3/.

Anne-Lyse Minard, Manuela Speranza, Ruben Urizar, Begoña Altuna, Marieke van Erp, Anneleen Schoen, and Chantal van Son. 2016. MEANTIME, the NewsReader multilingual event and time corpus. In *Proceedings of the Tenth International Conference* *on Language Resources and Evaluation (LREC‘16)*, pages 4417–4422, Portorož, Slovenia. European Language Resources Association (ELRA).

John Murzaku, Tyler Osborne, Amittai Aviram, and Owen Rambow. 2023. Towards generative event factuality prediction. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 701–715, Toronto, Canada. Association for Computational Linguistics.

John Murzaku and Owen Rambow. 2024. BeLeaf: Belief prediction as tree generation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: System Demonstrations)*, pages 97–106, Mexico City, Mexico. Association for Computational Linguistics.

John Murzaku, Peter Zeng, Magdalena Markowska, and Owen Rambow. 2022. Re-examining FactBank: Predicting the author‘s presentation of factuality. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 786–796, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

OpenAI. 2024a. Gpt-4o. `https://openai.com/index/hello-gpt-4o/`.

OpenAI. 2024b. o1. `https://openai.com/o1/`.

OpenAI. 2025. o3-mini. `https://openai.com/index/openai-o3-mini/`.

OpenRouter. 2025. Openrouter api. `https://openrouter.ai/`.

Riccardo Orlando, Luca Moroni, Pere-Lluís Huguet Cabot, Edoardo Barba, Simone Conia, Sergio Orlandini, Giuseppe Fiameni, Roberto Navigli, et al. 2024. Minerva llms: The first family of large language models trained from scratch on italian data. In *Proceedings of the Tenth Italian Conference on Computational Linguistics (CLiC-it 2024)*.

Amir Pouran Ben Veyseh, Thien Huu Nguyen, and Dejing Dou. 2019. Graph based neural networks for event factuality prediction using syntactic and semantic structures. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4393–4399, Florence, Italy. Association for Computational Linguistics.

Vinodkumar Prabhakaran, Tomas By, Julia Hirschberg, Owen Rambow, Samira Shaikh, Tomek Strzalkowski, Jennifer Tracey, Michael Arrigo, Rupayan Basu, Micah Clark, Adam Dalton, Mona Diab, Louise Guthrie, Anna Prokofieva, Stephanie Strassel, Gregory Werner, Yorick Wilks, and Janyce Wiebe. 2015. A new dataset and evaluation for belief/factuality. In *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics*, pages 82–91, Denver, Colorado. Association for Computational Linguistics.

Alexis Ross and Ellie Pavlick. 2019. How well do NLI models capture verb veridicality? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2230–2240, Hong Kong, China. Association for Computational Linguistics.

Marco Rovera, Serena Cristoforetti, and Sara Tonelli. 2025. ModaFact: Multi-paradigm evaluation for joint event modality and factuality detection. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6378–6396, Abu Dhabi, UAE. Association for Computational Linguistics.

Rachel Rudinger, Aaron Steven White, and Benjamin Van Durme. 2018. Neural models of factuality. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 731–744, New Orleans, Louisiana. Association for Computational Linguistics.

Roser Saurí and James Pustejovsky. 2009. Factbank: a corpus annotated with event factuality. *Language resources and evaluation*, 43:227–268.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Aaron Steven White, Rachel Rudinger, Kyle Rawlins, and Benjamin Van Durme. 2018. Lexicosyntactic inference in neural models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4717–4724, Brussels, Belgium. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

A Unified Experiments

Details For all **Unified** zero-shot experiments, we use a temperature of 0.0 where applicable (all models besides o1 and o3-mini). For o1 and o3-mini, we use the default reasoning setting (Medium). To prompt all other models (LLaMA-3.3, DeepSeek-v3, DeepSeek-r1, and Claude-3.5-Sonnet), we use the OpenRouter API (OpenRouter, 2025). The open models are ran at full precision (henceforth why we used the OpenRouter API and external providers).

Prompt Our zero-shot **Unified** prompt is shown in Figure 1.

B Hybrid Experiments

Details For all **Hybrid** zero-shot experiments, we use a temperature of 0.0 where applicable (all models besides o1 and o3-mini). For o1 and o3-mini, we use the default reasoning setting (Medium). To prompt all other models (LLaMA-3.3, DeepSeek-v3, DeepSeek-r1, and Claude-3.5-Sonnet), we use the OpenRouter API (OpenRouter, 2025).

Prompt Our **Hybrid** zero-shot prompt is shown in Figure 2.

C LLM Experiment Details

For all our FactBank experiments, we report a single run, especially due to cost. We note that o1 experiments cost up to $75 per run on the FactBank test set. To minimize randomness, we set the temperature to 0.0 where applicable (besides o1 and o3-mini). For o1 and o3-mini, we use the default reasoning setting (Medium). For our ModaFact experiments, we report the average over all five folds. Due to API costs and performing five-fold cross validation, we limit all ModaFact experiments to GPT-4o, o3-mini, and DeepSeek r1, which are the most cost effective models.

D Dataset Details

FactBank We use the author and source-and-target projection of FactBank from Murzaku and Rambow (2024), who follow the article split from Murzaku et al. (2022). We use their provided code for data extraction and follow their exact article split. The release of the FactBank corpus that we use can be found at the Linguistic Data Consortium, catalog number LDC2009T23. The test set contains 280 sentences and 1,326 examples.

ModaFact We use all five-folds of the test set of the ModaFact corpus from Rovera et al. (2025), which is publicly available. All results we report are averages over the five folds. To get the events from ModaFact for our **Hybrid** zero-shot experiments, we use the author's provided prediction files and inference script with mT5-XXL.

Fold	Sentences	Examples
Fold 1	646	2098
Fold 2	605	2097
Fold 3	606	2096
Fold 4	626	2094
Fold 5	601	2090

Table 4: Dataset details for the ModaFact test set.

E Source Normalization

We propose two source normalization prompts: a few shot source normalization prompt and an oracle source normalization prompt. For these prompts, we use GPT-4o, with temperature 0.0. We prompt GPT-4o using the OpenAI API. Our exact few shot source normalization prompt is shown in Figure 4. Our exact oracle normalization prompt is shown in Figure 3.

We perform an ablation analysis of our few-shot and oracle normalization techniques described in Section 4.4. We showcase these results for our top performing system (DeepSeek r1, **Hybrid**) in Table 5. Without any normalization, we achieve a Full F1 of 68.9% and Nest F1 of 17.5%. Our few shot normalization technique improves us 2.1% for Full F1, and more notably by 7.8% for Nest F1. Our oracle method, as expected (since we provide gold sources), performs even better than our few shot method, achieving a Full F1 of 72.7% and 27.1%. However, we choose to perform all experiments with our few shot normalization method instead of our oracle method to truly showcase LLMs capabilities for belief detection without any gold sources as input.

F Event Tagger

DeBERTa Tagger We use DeBERTa-large for token classification, setting the number of labels to 2 (O vs. EVENT). We use the following hyperparameters: Epochs: 5; Batch Size: 16; Learning Rate: 1e-4; Max Sequence Length: 128. We do not perform any hyperparameter optimization or tun-

Norm.	Full	Nest
None	68.9	17.5
Few Shot	72.0	25.3
Oracle	**72.7**	**27.1**

Table 5: Performance of DeepSeek r1 (Hybrid) under three source normalization settings. "None" denotes no normalization, "Few Shot" applies few-shot normalization, and "Oracle" uses ground-truth for normalization. Bold values indicate the best results for each test set.

Category	Count	Breakdown	Count
Source	123	Gold=AUTHOR	50
		Gold="it"	13
FN	77	Missed Noun	38
		Missed Verb	30
Label	73	Pred:CT+ $\rightarrow$ Gold:UU	28
		Pred:PR+ $\rightarrow$ Gold:UU	22
FP	53	Predicted Noun	33
		Predicted Verb	10

Table 6: Error analysis for our **Hybrid** DeepSeek r1 system on nested predictions, showing counts of each error type relative to its category total count.

ing. The model is trained using the HuggingFace Transformers library (Wolf et al., 2020).

LLM Event Tagging We perform event tagging on multiple LLMs. We set the temperature to 0.0. We use the OpenRouter API (OpenRouter, 2025) for DeepSeek r1 and Claude 3.5, and the OpenAI API for GPT-4o. We do not perform experiments with o1 to avoid high costs. Our zero-shot event detection prompt is shown in Figure 5. Our few-shot event detection prompt is shown in Figure 6.

G Nested Error Analysis

We expand our error analysis on the nested sources. Table 6 shows our error counts and error types.

We specifically analyze the errors for nested beliefs, which is where all LLMs fail on (our top performing model achieving F1 of only 25.3%). We showcase the error category, and then the top two error types by count. We use the following labels: **Source** indicates a source mismatch error; **FN** indicates a false negative, where the LLM did not generated a certain event type; **Label** indicates a label error where the LLM had the source and event correct, but inorrectly labeled the event. **FP** indicates a false positive, where the LLM overpredicted (that is, it generated an event that was not actually an event).

Our most notable error is **Source**, where 123 errors are made. The most common error is the model predicts the author is the source instead of the correct nested source. Another notable error is where the model does not classify the source as it, but rather predicts the name of the entity. Next, we see a repeating of similar errors that Murzaku et al. (2022) discovered. Specifically, event nouns can be hard to determine (e.g. nouns like "concerns", "acquisition", "construction"). Our **FN** and **FP** errors showcase that LLMs simultaneously over predict event nouns, while also missing both event nouns and verbs. Finally, we notice two notable label flips for our **Label** error category: the LLM predicts CT+ (the event happened/is true) when the gold label is UU (unknown), and PR+ (possibly true) when the gold is UU. This is due to FactBank's definitions of nested sources and future events: when a reporting of a future event happening (e.g. "Mary said it will happen"), the factuality of the event according to the source is UU (the source is not committing to the event; rather, the author is commiting to it).

Our analysis emphasizes that despite our source normalization method and use of strong reasoning LLMs, there is much room for improvement. Our error analysis findings are further supported with similar errors that have been reported in previous works on FactBank (Murzaku et al., 2022).

H Code Release

We will release all of our code. We will provide the full pipelines, datasets, and model checkpoints where applicable.

Figure 1: **Instruction for our zero-shot Unified Belief Annotation.** The instruction for FactBank-style event factuality annotation consists of three parts: a brief task description, detailed step-by-step instructions, and the formatting structure. Our CoT instructions are shown in the end of the prompt (Step-by-Step Output).

You are an annotation assistant trained to process sentences according to a FactBank-style event factuality framework. Given a sentence (or short text), your task is to analyze and annotate events by:

- **Event Identification**: Finding and listing all event-denoting predicates (verbs, event nouns, state-denoting adjectives)
- **Source Analysis**: Identifying who is expressing or committing to each event
- **Factuality Assessment**: Determining how certain each source is about the events
- **Nested Attribution**: Managing multiple layers of reporting and belief
- **Special Cases**: Handling future events, negation, modality, and hedging

Follow these steps precisely for annotation:

STEP 1: Event Identification
- Find all event-denoting predicates in the text
- Each predicate must be a single token
- Include verbs, event nouns, and state-denoting adjectives

STEP 2: Source Identification
- Start with "AUTHOR" as the root source (text narrator)
- Identify source-introducing predicates (SIPs) like "said," "believed," "reported"
- For new sources (e.g., "Apple officials"), normalize to single-token labels (e.g., "officials")
- Format as "AUTHOR_<ShortLabel>" (e.g., "AUTHOR_officials")
- For nested sources, add additional levels with underscores (e.g., "AUTHOR_officials_spokesperson")
- Handle negated sources (e.g., "did not say") at the higher level

STEP 3: Factuality Labeling Assign one of these labels for each event-source pair:
- **true**: Certainly factual (e.g., "confirmed," "knew")
- **false**: Certainly counterfactual (e.g., "denied," "did not happen")
- **ptrue**: Probably true (e.g., "might," "could," "likely")
- **pfalse**: Probably false (e.g., "doubted")
- **unknown**: Non-committal or unspecified stance

Special Cases Guidelines
- Future/prospective events: Label as unknown unless probability indicated (then ptrue)
- Negative statements: Use false for explicit denials
- Modality/hedging: Use ptrue for "might," "could," "suspected"
- Uncommitted author: Use unknown for purely reported events

Your annotation should be formatted as a JSON-style list of dictionaries:

```
[
  {
    "source": "<source_label>",  // e.g., "AUTHOR" or "AUTHOR_<source>"
    "event": "<event_token>",    // exact predicate from text
    "label": "<factuality_value>" // true/false/ptrue/pfalse/unknown
  },
  ...
]
```

Step-by-Step Output Process:
- Walk through each event in the sentence
- Identify and explain all sources and their nesting
- Justify each factuality label from each source’s viewpoint
- Produce the final JSON-style output

Figure 2: **Instruction for our Hybrid Belief Annotation.** The instruction for FactBank-style event factuality annotation consists of three parts: a brief task description, detailed step-by-step instructions, and the formatting structure. Our CoT instructions are shown in the end of the prompt (Step-by-Step Output).

You are an annotation assistant trained to process sentences according to a FactBank-style event factuality framework. Given a sentence (or short text) and a list of event predicates marked in that sentence, your task is to analyze and annotate events by:

- **Source Analysis**: Identifying who is expressing or committing to each event
- **Factuality Assessment**: Determining how certain each source is about the events
- **Nested Attribution**: Managing multiple layers of reporting and belief

Follow these steps precisely for annotation:

STEP 1: Source Identification

- Start with "AUTHOR" as the root source (text narrator)
- Identify source-introducing predicates (SIPs) like "said," "believed," "reported," "estimated," "argued"
- For new sources (e.g., "Apple officials"), normalize to single-token labels (e.g., "officials")
- Format as "AUTHOR_<ShortLabel>" (e.g., "AUTHOR_officials")
- For nested sources, add additional levels with underscores (e.g., "AUTHOR_officials_spokesperson")
- Handle negated sources (e.g., "did not say") at the higher level

STEP 2: Factuality Labeling Assign one of these labels for each event-source pair:

- **true**: Certainly factual (e.g., "confirmed," "knew")
- **false**: Certainly counterfactual (e.g., "denied," "did not happen")
- **ptrue**: Probably true (e.g., "might," "could," "likely")
- **pfalse**: Probably false (e.g., "doubted")
- **unknown**: Non-committal or unspecified stance

Special Cases Guidelines

- Future/prospective events: Label as unknown unless probability indicated (then ptrue)
- Negative statements: Use false for explicit denials
- Modality/hedging: Use ptrue for "might," "could," "suspected"
- Uncommitted author: Use unknown for purely reported events

Your annotation should be formatted as a JSON-style list of dictionaries:

```
[
  {
    "source": "<source_label>",  // e.g., "AUTHOR" or "AUTHOR_<source>"
    "event": "<event_token>",    // exact predicate from text
    "label": "<factuality_value>" // true/false/ptrue/pfalse/unknown
  },
  ...
]
```

Step-by-Step Output Process:

- Walk through each event in the sentence
- Identify and explain all sources and their nesting
- Justify each factuality label from each source's viewpoint
- Produce the final JSON-style output

Figure 3: **Oracle Source Normalization Prompt**

You are determining if two source names refer to the same entity. Consider company abbreviations, common variations, and parent/subsidiary relationships. Also consider the context of the sentence and entity coreference.

Answer only YES if these definitely refer to the same entity, NO if they are different or if you're unsure. Include a brief explanation of your reasoning.

```
Sentence: {Sentence}
Predicted Source: {Predicted Source}
Gold Source: {Gold Source}
```

Figure 4: **Few Shot Source Normalization Prompt**

You are a FactBank-style source normalization assistant.
Your task: **Identify and normalize the subject (speaker/thinker/etc.) of each source-introducing predicate (SIP)** in a sentence. The normalized form must be a short, single-token label following "AUTHOR_". If nested sources appear (i.e., one speaker quotes another speaker), nest them by appending an underscore plus the new label.

Use these **rules and guidelines**:
1. **Source-Introducing Predicates (SIPs)**:

- Common SIP verbs: "said," "reported," "believed," "estimated," "argued," "announced," "denied," "claimed," etc.
- If an entity is repeated (the same subject for multiple SIPs), reuse the same label.

2. **Normalization**:

- Reduce corporate entities to "Corp." or "Inc." instead of the full name. E.g.:
  - "Marathon Widget Corp." → **AUTHOR_Corp.**
  - "Skyline Media Inc." → **AUTHOR_Inc.**
- If it's just "the company," consider normalizing to **AUTHOR_company** *only* if no more specific corporate form (like "Corp.") is available.
- For people:
  - "He," "she," "they" → **AUTHOR_he**, **AUTHOR_she**, **AUTHOR_they**
  - "Mr. Alvarez," "Ms. Hurt," or "Dr. Kim" → **AUTHOR_Alvarez**, **AUTHOR_Hurt**, **AUTHOR_Kim**
  - If the sentence says "I stated..." → **AUTHOR_I**
- For large institutions:
  - "Ministry of Defense" → **AUTHOR_ministry**
  - "Police Department" → **AUTHOR_police**
  - "officials" → **AUTHOR_officials**
  - "board" → **AUTHOR_board**
- If you have a nested quote, e.g., "AUTHOR_officials_spokesperson" if the spokesperson is quoting officials.

3. **Polarity**:

- Even if the SIP is negated, you still label that source. (e.g., "he denied..." is valid.)

4. **Output**:

- Output **only** the normalized label(s). If no new source is introduced, or if you're uncertain, you can leave the text unchanged or indicate "No SIP found."

Few-Shot Examples *(truncated for space; we use 10 few shot examples)*

1. **Sentence**: Alpha Widget Corp. said it is launching a new product line.
   Predicted: AUTHOR_Alpha_Widget_Corp.
   Corrected: AUTHOR_Corp.
2. **Sentence**: "I believe the results speak for themselves," he announced.
   Predicted: AUTHOR_he
   Corrected: AUTHOR_I
   (*Because "I" is the direct speaker—if the text clearly attributes the quote to the first person.*)
3. **Sentence**: In its quarterly filing, LRS Acquisition stated it expects higher revenue.
   Predicted: AUTHOR_LRS
   Corrected: AUTHOR_Acquisition
4. **Sentence**: A portfolio unit of Greenbank Corp. reported continued growth this year.
   Predicted: AUTHOR_portfolio unit
   Corrected: AUTHOR_unit
5. **Sentence**: The foreign minister declared that cooperation would improve global stability.
   Predicted: AUTHOR_foreign minister
   Corrected: AUTHOR_minister

Return:

- Return the final normalized label(s) if a new source arises from the SIP.
- If none or unclear, output "No SIP found" or leave the text as is.

Figure 5: **FactBank Single-Token Event Identification Prompt**

You are an expert at identifying single-token events in text following FactBank guidelines.
Find ALL single-token predicates that:

Criteria:

1. **Are ONLY ONE of:**
   - Reporting verbs (communication)
   - Cognitive verbs (mental states)
   - Action verbs (physical/abstract actions)
   - Event nouns (occurrences/happenings)
   - State adjectives (temporary states)
2. **Must represent:**
   - Something that happened/happens/will happen
   - Something that can be assessed as true or false
   - Something with a temporal dimension
3. **Critical Distinction for Nouns/Nominalizations:**
   - **INCLUDE** only nouns that refer to specific instances of events with:
     - Concrete temporal bounds
     - Specific participants
     - Ability to be assessed as having occurred or not
   - **DO NOT include** nouns that refer to:
     - General concepts or types of events
     - Abstract categories
     - Topics or subjects of discussion
     - Generic processes
     - Institutional practices

Key rules:

- Extract SINGLE tokens only
- Include all verbs from source-introducing predicates
- Include nested events
- Include events under modals or negation
- Include events in complement clauses

Do NOT include:

- Multi-word phrases
- Generic nouns
- Auxiliary verbs
- Articles, prepositions, or conjunctions
- References to event types without specific instances

Output Format:
Your output is a JSON-style list of dictionaries. Each dictionary has:

- `"event"`: The exact event token or predicate from the sentence.

Output Example:

```
[
  {"event": "EVENT1"},
  {"event": "EVENT2"},
  {"event": "EVENT3"}
]
```

Figure 6: **FactBank Few-Shot Single-Token Event Identification Prompt.**

You are an expert at identifying single-token events in text following FactBank guidelines.
Find ALL single-token predicates that:

1. **Are ONLY ONE of:**

- Reporting verbs (communication)
- Cognitive verbs (mental states)
- Action verbs (physical/abstract actions)
- Event nouns (occurrences/happenings)
- State adjectives (temporary states)

2. **Must represent:**

- Something that happened/happens/will happen
- Something that can be assessed as true or false
- Something with a temporal dimension

3. **Critical Distinction for Nouns/Nominalizations:**

- **INCLUDE** only nouns that refer to specific instances of events with:
  - Concrete temporal bounds
  - Specific participants
  - Ability to be assessed as having occurred or not
- **DO NOT include** nouns that refer to:
  - General concepts or types of events
  - Abstract categories
  - Topics or subjects of discussion
  - Generic processes
  - Institutional practices

Key rules:

- Extract SINGLE tokens only
- Include all verbs from source-introducing predicates
- Include nested events
- Include events under modals or negation
- Include events in complement clauses

Do NOT include:

- Multi-word phrases
- Generic nouns
- Auxiliary verbs
- Articles, prepositions, or conjunctions
- References to event types without specific instances

Output Format:
Your output is a JSON-style list of dictionaries. Each dictionary has:

- `"event"`: The exact event token or predicate from the sentence.

Examples: (we truncate the examples omit 3 examples here for brevity)
1. **Sentence:** In composite trading Friday on the New York Stock Exchange, BellSouth shares fell 87.5 cents.
Output:

```
[
  {"event": "trading"},
  {"event": "fell"},
]
```

2. **Sentence:** Many local residents denounced the bigotry.
Output:

```
[
   {"event": "denounced"},
   {"event": "bigotry"}
]
```