Non-Factual Argument Mining

Anonymous ACL submission

Abstract

Argument mining (AM) benchmarks typically consist of annotated real-world argumentation. These benchmarks often conflate logical validity with factual plausibility. As a result, a model may appear competent because it recognizes familiar real-world content, not because it robustly tracks argumentative structure. The issue seems particularly relevant for recent large language models (LLMs), which achieve strong performance on many AM subtasks, but whose ability to reason, as opposed to merely reproducing statistical patterns, remains a subject of ongoing inquiry. This paper presents a study of LLM performance on AM in a setting where factual validity is disentangled from formal argumentative validity. We present two experimental scenarios. In both cases, we first introduce a method for generating argumentative texts and then use those texts as a benchmark for LLMs' ability to recognize argumentative validity. Both methods assume texts with "non-factual" content, that is, texts whose premises or conclusions concern fictional entities, conflict with facts about our world, are grounded in alternative worlds, and so on. In the first scenario, the texts are generated in a strictly algorithmic manner and take the schematic form of chains of syllogisms. In the second scenario, we use specially prepared input prompts to generate non-factual argumentation of a high degree of diversity and naturalness – the BENFA benchmark.

1 Introduction

Argument mining (AM) studies how to identify and represent reasoning in natural language. Typical tasks include detecting claims and premises, identifying support and attack relations, reconstructing argument graphs, and classifying patterns such as argument schemes or fallacies (Park and Cardie, 2018). These capabilities matter in applications such as debate analysis, decision support, fact-checking pipelines, educational tools, and broader forms of computational rhetoric.

To assess the capabilities of particular models in AM-type tasks, researchers use argumentative texts of diverse genres and sources, such as essays, opinion pieces, debate transcripts, online discourse, and so forth (Stab et al., 2018). There are also datasets of synthetically generated arguments, which are by design meant to imitate such texts (Irani et al., 2024). Naturally, the more closely the source material corresponds to the intended application domain, the more useful the resulting models are likely to be. This creates a strong preference for authentic argumentation produced by humans. The use of such texts, however, raises a methodological difficulty: authentic argumentation always concerns facts and phenomena from our world, our reality. If we train or evaluate a model on such data, a fundamental question arises: does the model truly possess knowledge of correct argumentation, or does it merely have a very extensive lexical familiarity with the kinds of arguments that occur in our societies? This connects to the broader question of whether artificial intelligence models are capable of reasoning, as opposed to simply retrieving memorized data, a question that is intensely debated especially in the context of large language models (LLMs), whose fundamental operating principle is statistical and whose training requires enormous quantities of textual data.

To verify the kind of knowledge we are interested in, namely knowledge of the principles of argumentation as such, we would need to confront a model with argumentation it has not seen during training (out-of-distribution). In practice, however, this is difficult even when previously unpublished materials are used, because contemporary LLMs are trained on extremely large corpora and real-world argumentation tends to recur across a relatively limited set of topics and patterns. Obtaining argumentation that is genuinely foreign to an LLM in thematic terms therefore appears most feasible in

a synthetic setting, through the deliberate creation of such texts.

Producing a dataset of a size appropriate for quantitative research that would be entirely authored by humans lies beyond the resources available to the authors of this paper. In its place, the authors set themselves the task of developing generative methods. The first is purely algorithmic. On the basis of programmatically specified rules, it generates chains of syllogisms whose subjects and properties are intentionally exotic. We then examine language models' ability to assess the correctness of these chains of inference as a function of the number of links they contain. As might be expected, these texts are fairly schematic, which substantially reduces the difficulty of the resulting benchmark. This gives rise to a central challenge: how can one create, by generative means, argumentative texts with an unusual subject matter and with a level of linguistic complexity and diversity corresponding to natural human speech?

The second part of the study is devoted to this problem. We present a hybrid method which, on the basis of an input specification covering the general characteristics of an alternative world and the subject matter of the argumentation, uses a multi-step procedure together with frontier LLMs to generate non-factual arguments with a high degree of diversity and naturalness. The original contribution of this work can therefore be divided into three parts:

1. Two procedures for generating non-factual arguments.

2. Two datasets obtained by means of the above procedures.

3. An evaluation of selected contemporary language models with respect to their ability to recognize correct inference on the basis of these datasets.

The paper is organized as follows. After the introduction comes a review of related literature (Section 2). Section 3 discusses the syllogism-generation method, while Section 4 presents the hybrid method. The paper concludes with Section 5, which contains a summary and directions for future research.

A note on terminology: Although related literature often uses the term *counterfactuality*, in this paper we use "non-factual arguments" to denote argumentation whose subject matter departs substantially from the real world and may involve arbitrary modifications of its features.

2 Related Work

Previous studies on argumentation have recognized counterfactuals as important in causal, legal, rhetorical, and inferential reasoning, but have rarely treated them as an independent object of systematic analysis. In existing research, counterfactuals have appeared primarily as an element of causal argumentation, where an alternative course of events is used to test cause–effect relations; as part of reductio ad absurdum, where a hypothetical assumption is rejected by pointing to its absurd consequences (Jansen, 2007); or as a component of a contrario arguments and legal argumentation, where the consequences of an alternative scope of a norm are considered (Walton et al., 2016). In rhetorical and pragma-dialectical approaches, counterfactuals may function as a way of strategically presenting an argument, for example in constructions such as "if that were true, I would never have behaved in this way" (Jansen, 2011). Finally, in the theory of argumentation schemes, counterfactuals may function as implicit warrants, that is, rules enabling the transition from premises to conclusions (Walton et al., 2008). Consequently, counterfactuality has been present in argumentation theory, but usually as a dispersed mechanism within different types of reasoning, rather than as a distinct class of arguments or a separate type of data for argument mining.

Counterfactuality has appeared in research on LLMs, especially in studies examining models' ability to recognize and properly handle false premises in reasoning tasks, often discussed as false-premise reasoning (Hu et al., 2023; Yu et al., 2023). This line of work asks whether a model can reject a question or problem based on an erroneous assumption, rather than automatically generating an answer that accepts the false premise. Related studies have also addressed the detection of misleading statements, logically inconsistent claims, and factually incorrect contextual narratives (Ghosh et al., 2024; Galitsky et al., 2024). Their common aim has been to determine whether an LLM can take into account, in its response, that the starting point of reasoning is false, unverified, or internally inconsistent.

Recent studies have moved the problem of counterfactuality and false premises toward a more sys-

tematic evaluation of LLM reasoning. Shafiei et al. (2025) test models' ability to detect false premises hidden in multi-hop reasoning tasks, where several consecutive inferential steps must be verified. Beyer and Reed (2025), in turn, place even stronger emphasis on separating reasoning from factual knowledge by testing whether LLMs solve logical tasks through genuine reasoning or rather by recognizing familiar content from training data. The decline in model performance after data modification suggests that earlier performance may have partly resulted from lexical recall rather than full logical reasoning.

These studies are important for argument mining because, in existing benchmarks, argumentative structure has usually been intertwined with domain knowledge. A model could therefore evaluate an argument not only on the basis of the relation between premise and conclusion, but also on the basis of world knowledge and dominant narratives. Shafiei et al. (2025) and Beyer and Reed (2025) show how this problem can be mitigated: the former embed false premises within reasoning chains, while the latter decouple logical structure from familiar lexical content.

These approaches can provide inspiration for developing new benchmarks for counterfactual argumentation, in which models would evaluate not the factual truth of premises, but the inferential correctness of an argument within a controlled, internally coherent, yet non-factual world.

3 Study Design for Syllogistic Reasoning Evaluation

3.1 Research Objective

The purpose of this study was to evaluate the ability of selected large language models to assess the logical validity of extended chains of syllogistic reasoning expressed in Polish. The task was formulated as a binary classification problem. For each input, the model was required to determine whether the entire chain of reasoning was logically valid or whether at least one step in the chain was invalid.

The study was designed to test not merely the recognition of isolated syllogistic forms, but the capacity to process sequences of connected inferences expressed in natural language. A model may correctly classify a single syllogism and nevertheless fail when validity depends on the interaction of multiple premises and derived conclusions across a longer reasoning chain. The class terms used in the generated texts were deliberately fictional and semantically unusual. This design choice was intended to reduce the probability that the models had encountered highly similar syllogistic examples in their training data, and thus to limit the influence of memorized patterns on the evaluation.

3.2 Representation of Syllogistic Statements

The symbolic representation used in the study followed the format `AxB`, where `A` and `B` denote class terms, and `x` denotes the type of categorical proposition relating them. Four standard categorical forms were used: `a`, corresponding to a universal affirmative proposition, that is, "every A is B"; `i`, corresponding to a particular affirmative proposition, that is, "some A is B"; `e`, corresponding to a universal negative proposition, that is, "no A is B"; and `o`, corresponding to a particular negative proposition, that is, "some A is not B". A class term denotes a category or class of objects, such as "dark unicorns", "colorful hydras", or "crooked minotaurs".

In addition to categorical propositions, the generated reasoning chains included explicit positive existential assumptions for every sampled class term. For example, if the term "dark unicorns" appeared in a chain, the text also included a statement such as "dark unicorns exist". These assumptions were introduced in order to make the existential import of the relevant terms explicit. This is methodologically important, since some syllogistic inferences involving particular conclusions depend on the non-emptiness of the classes involved. The study was conducted on syllogistic chains generated for three values of `n`: 3, 4, and 5. For each value of `n`, both valid and invalid datasets were prepared.

3.3 Generation of Valid Syllogistic Chains

A generator was developed to produce chains of syllogistic reasoning for a given number of initial statements. For each set of input propositions, the generator identified all syllogistic conclusions that followed from them according to valid syllogistic modes.

For example, for `n = 3`, one possible set of symbolic input statements was:

```
('A', 'e', 'B')
('C', 'e', 'B')
('D', 'a', 'C')
```

From such premises, the generator derived further valid conclusions using syllogistic modes such as Celarent, Celaront, Calemes, Calemos, and Fe-

lapton. The resulting symbolic structures were then converted into natural-language texts.

Each generated text had a stable structure. It first presented the initial assumptions, then listed the positive existential assumptions for all sampled class terms, and finally presented the conclusions derived from the preceding material. This structure made it possible to test model performance on formally controlled reasoning chains expressed in fluent natural language.

3.4 Generation of Invalid Syllogistic Chains

Invalid examples were generated from valid reasoning chains by a separate procedure. The program identified the first statement occurring in the reasoning section and replaced that statement throughout the text with its opposing form, including in the initial assumptions.

This manipulation preserved the general surface structure of the text while making the reasoning chain logically invalid. As a result, the invalid examples were not trivially distinguishable from the valid examples by length or formatting. The error was introduced near the beginning of the reasoning process, which allowed the study to test whether models could detect an early invalid step and reject the entire chain without having to verify all subsequent inferences.

3.5 Prompting Procedure

The prompt was written exclusively in Polish. The models were evaluated in a zero-shot prompting setting. The prompt specified the task and the two admissible output labels, but did not include any worked examples of valid or invalid syllogistic chains. Each model received the full text of a reasoning chain and was instructed to return only the final assessment of logical validity. The expected answer was `TAK` (eng.`YES`) if the entire reasoning chain was logically valid, and `NIE` (eng.`NO`) if at least one part of the reasoning chain was invalid.

For evaluation, the last three characters of each model output were extracted and compared against `TAK` or `NIE`. This procedure allowed any preceding text, including text generated as part of the model's reasoning process, to be ignored. Responses that did not end in either `TAK` or `NIE` were treated as non-conforming responses. This category was introduced to separate failures of logical classification from failures to follow the required output format.

3.6 Results

A retest was performed to confirm the stability of the results. A detailed analysis of both the main study and the retest results is provided in the appendix.

Model	**n=3**	**n=4**	**n=5**
Bielik-11B*	0.696	0.716	0.703
Mistral-7B**	0.556	0.528	0.536
Ministral-3-8B***	0.696	0.720	0.672
gpt-oss-20b	0.947	0.959	0.939
Qwen2.5-7B	0.515	0.525	0.515
Qwen2.5-7B (retest)	0.529	0.531	0.512
Qwen3-8B	0.978	0.983	0.977
Qwen3-8B****	0.795	0.797	0.794
Qwen3.5-9B	1.000	0.999	0.998
Qwen3.5-9B****	0.631	0.704	0.641
Qwen3.5-4B	0.996	0.996	0.996
Qwen3.5-2B	0.958	0.959	0.967
Qwen3.5-0.8B	0.000	0.000	0.667
Total	648	6282	55844

* Bielik-11B-v2.1-Instruct
** Mistral-7B-Instruct-v0.3
*** Ministral-3-8B-Reasoning-2512
**** Reasoning disabled

Table 1: Overall accuracy on syllogism sequences, excluding non-conforming responses.

One important finding is that reasoning mode should not be treated as a sufficient condition for strong logical performance. The strongest models in the study used reasoning, and the reasoning-enabled variants of Qwen3-8B and Qwen3.5-9B outperformed their reasoning-disabled counterparts. However, this pattern was not universal. Ministral-3-8B-Reasoning-2512 performed close to Bielik-11B-v2.1-Instruct, despite the fact that Bielik did not use an explicit reasoning mode. This suggests that reasoning mode can improve performance, but only when the underlying model is able to use the additional reasoning process effectively and complete it within the available generation limit.

The results for Bielik-11B-v2.1-Instruct are also methodologically relevant. Its performance may have been supported by stronger competence in Polish, since both the prompt and the syllogistic chains were expressed exclusively in Polish. This does not imply that language specialization alone is sufficient for high logical performance, but it does indicate that language competence may interact with reasoning performance. In this study, the models were not evaluated under alternative prompt languages, so the effect of prompt language remains an open question.

Mistral-7B-Instruct-v0.3 exhibits a strong ten-

dency to classify syllogistic chains as valid, which underscores the need for a separate evaluation of valid and invalid cases, as described in detail in the appendix.

The Qwen3.5 model family provides evidence of a possible size-related threshold. Qwen3.5-4B retained substantial task performance, whereas Qwen3.5-2B and Qwen3.5-0.8B showed a sharp collapse in overall accuracy. This suggests that below a certain model size, the task may exceed the model's effective capacity under the tested inference conditions. However, the interpretation should remain cautious. The results after excluding non-conforming responses indicate that some smaller models may produce correct classifications when they complete the output in the required format. Thus, low overall accuracy may partly reflect failures to complete the reasoning process or to comply with the output format, rather than logical incompetence alone. The result for Qwen3.5-0.8B at n = 5, where accuracy reached 0.667, should be interpreted as a statistical artifact, as discussed in detail in the appendix.

The analysis of non-conforming responses and `length` terminations supports this interpretation. Non-conforming responses were concentrated mainly among reasoning models and occurred more often for valid syllogistic chains. This is consistent with the structure of the task. To confirm that a valid chain is correct, a model must in principle verify all syllogistic steps. By contrast, an invalid chain can be rejected once the first invalid syllogistic step is identified. Since invalid examples in this study contained an error near the beginning of the chain, they offered an earlier stopping point for successful rejection. The higher frequency of `length` terminations for valid chains therefore appears to reflect not only a technical limitation, but also the logical structure of the verification problem.

4 Hybrid Method

The algorithmic method for generating syllogisms produces texts that may be structurally complex, but are highly schematic. This schematic character, combined with a predicate structure familiar even from traditional logic textbooks, means that despite their unusual subjects these sentences do not pose a major challenge for language models, at least if the models possess some genuine reasoning capability. In this section, we present a method for generating texts that sound natural while having highly unnatural content. A sample of texts generated in this way, with substantial generic and thematic diversity, is then used to test whether the presence of non-factual content disrupts LLMs' ability to categorize correct reasoning.

4.1 Generation Procedure

To assess the impact of non-factual content in a meaningful way, our goal is to build a dataset that contains argumentation situated both in the real world and in alternative worlds. For the comparison to be meaningful, these arguments must be as similar to one another as possible, that is, the only substantial difference should be the presence of non-factual premises, and possibly conclusions. These texts are therefore generated in pairs. The "translation" of an argument from the real world into an alternative one, or vice versa, raises certain challenges, which we comment on below. In addition, for the dataset to function as a benchmark for detecting logically correct argumentation, not necessarily based on correct premises, it must contain examples of both correct and incorrect reasoning. We therefore distinguish four types of items in our dataset:

- (1) Texts containing correct argumentation in an alternative world
- (2) Incorrect argumentation in an alternative world
- (3) Correct argumentation in the real world
- (4) Incorrect argumentation in the real world

We further distinguish between *slightly* (2a/4a) and *seriously* (2b/4b) incorrect argumentation; this distinction will be specified later. We thus obtain six categories, shown in Figure 1.

To ensure equal status for arguments in the real and alternative worlds, we use a symmetric procedure in which half of the "real" texts are derived from "fictional" ones (F2R), and half in the opposite direction (R2F). The need to translate a "real" text into a "fictional" one, and vice versa, rather than generate them independently, is dictated by the requirement already mentioned: the arguments must be as analogous, and therefore as comparable, as possible. The existence of two directions of translation, combined with the six categories listed above, means that the dataset is composed of families of 12 related argumentative texts each.

		Fictional world	Real world
Valid argumentation		**1**	**3**
Invalid argumentation	Light errors	**2a**	**4a**
	Serious errors	**2b**	**4b**

Figure 1: The six variants of argumentative text as distinguished in this study

We assume that the alternative worlds considered here are internally coherent. The *world specification* used in the procedure lists the significant differences between the alternative world and the real one. In general, it is impossible to describe exhaustively all the consequences that a single imagined change would have for such a world. For the purposes of text generation, however, and in keeping with the assumption of coherence, we always allow all obvious and logically necessary consequences that follow from the stipulated differences to be assumed.

In order to obtain a maximally diverse text collection, we introduce the notion of *empirical distance*. This is a number from 1 to 5, intended to indicate how substantially the alternative world differs from the real one. Thus, at one end of the scale are worlds in which, for example, a well-known footballer plays for a different national team, or a famous literary work has a slightly altered plot; at the other end are fantasy worlds or worlds governed by entirely different laws of physics.

We now present, step by step, the detailed procedure for generating argumentative texts, for both of the translation directions mentioned above. The procedure is schematically illustrated for the F2R direction in Fig. 2. For details such as model choice, possible parameter values, sampling parameters, and so forth, we refer the reader to Appendix B.

F2R The input consists of a *world specification* of variable length and a *fictional topic*, which in one or two sentences further specifies the subject matter of the argumentative text, naturally touching on some of the differences listed in the *specification*.

The first step is the *generation of the (main) argument map*. In addition to the above input, a dedicated prompt is supplied with the *number of argumentative nodes* and the *text genre*. The LLM produces an argument map in a specified format: it contains the skeleton of the argumentation, with argumentative discourse units (ADUs) as vertices and edges denoting inference, support, or attack.

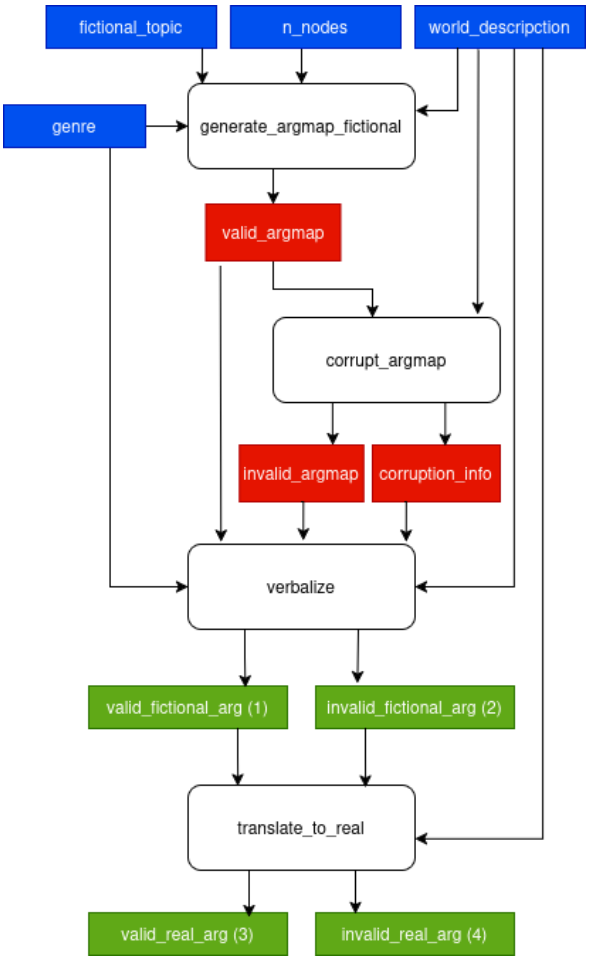

Figure 2: Schematic representation of the hybrid non-factual argumen generation pipeline (F2R variant)

The second step is the *generation of defective argument maps*. The LLM is given the argument map from the previous step together with a prompt instructing it to introduce a certain number of logical errors of a specified severity. In the dataset presented in this paper, we always request the introduction of 3 slight (a) and 3 serious (b) logical errors. The output is a defective argument map together with the LLM's commentary describing what logical errors were introduced into the map. This description is used for diagnostic purposes and as auxiliary data in subsequent stages.

The third step is the *verbalization of the map*. The dedicated prompt is supplied with the world description, the argument map, whether correct or

defective, and optionally the description of the errors. In addition, the target genre of the text is again specified, together with more detailed instructions describing its character, above all the *noise level*, that is, a general description of what kind of content not directly related to the argumentation should appear in the text so that it sounds more natural. The result of this step is texts of types (1) and (2).

To obtain their accompanying real-world counterparts, types (3) and (4), in the **fourth step** a separate prompt containing the world description, the argument map, and the verbalized text instructs the LLM to transform it, with minimal changes, into a text situated in the real world.

R2F Reversing the above procedure makes it possible to generate argumentative texts in the real world, and fictional ones as their derivatives.

The first step is again the generation of the argument map. The only difference from the F2R procedure is that instead of a world description, the model is told that the argumentation is to be situated in the real world. The topic of the text is given as a *real-world topic*, which together with the *fictional topic* and the *world description* forms the set of input units that make it possible to generate complete sets of 12 texts. Naturally, for the best results the *real-world topic* must be analogous in some way to the *fictional topic*.

The second step is the generation of defective maps, exactly as in F2R. Likewise, in the *third step*, the verbalization of these maps is generated in the same manner, yielding texts of types (3) and (4).

To obtain texts in the alternative world, the **fourth step** consists in the *fictionalization* of these texts. Analogously to the previous case, the prompt includes the description of the fictional world, the correct or defective map, optionally a commentary on the introduced errors, and the text that is to be fictionalized. In this way, texts of types (1) and (2) are obtained, which completes the procedure.

4.2 BENFA: A Dataset of Non-Factual Arguments

For the purposes of this paper, we created a set of input data together with a list of text genres that served to construct the first *benchmark of non-factual argumentation* (BENFA). Our input set covers 10 categories of alternative worlds, such as different laws of nature, alternate history, different social conditions, works of culture, as well as fantasy and science-fiction worlds. Each category contains 5 sets consisting of a world description, a "real" topic, and a "fictional" topic. Our list of genres includes 6 types: online discussion, scientific text, opinion column, blog/podcast, political debate, and informal live discussion. For each set, the text genre was sampled from a predefined subset of these genres, since not every combination of topic and genre would make it possible to generate a naturally sounding text. We experiment with two variants of argument complexity: 5 and 10 vertices in the argument graph. This yields a total of $10 \times 5 \times 2 \times 12 = 1200$ high-quality texts of non-factual argumentation.

4.3 Evaluation on the Benchmark of Non-Factual Arguments

The final part of our study uses the above dataset as an evaluation instrument. In the evaluation prompt, we provide a text of type (1), (2), (3), or (4) and instruct the model to determine, in a single word, whether the argumentation contained in the text is valid or invalid. The instruction specifies that the issue is not the truth of the premises themselves, but the logical coherence of the argumentation as such.

Our study includes a broad selection of open models of different sizes and from different developers. These are presented in detail in Table 10 in the appendix.

Argument complexity We now discuss the overall performance of the models on the BENFA benchmark (Figures 3 and 4). First, as expected, the models perform worse on texts that are more complex in argumentative structure (node10). More generally, in our evaluation most small and medium-sized models achieve an average accuracy of around 0.5. This means that they perform at the level of random guessing. Results falling even below the random baseline are explained by Figure 4: most small and medium-sized models show a strong tendency to predict one of the two classes disproportionately often. This is usually the label “valid”; that is, they are largely blind to errors embedded in the argumentation. Because of the 1:2 class imbalance in the dataset, this tendency partly drives their accuracy below 0.5. Again as expected, serious errors are detected somewhat more often than light ones. Most likely, however, our procedure generated errors that were too subtle in both cases for models of this class to detect reliably. An exception of the opposite kind is gpt-

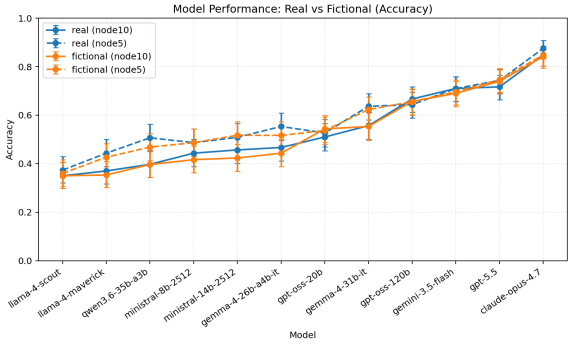

Figure 3: Model accuracy real world vs. alternative worlds

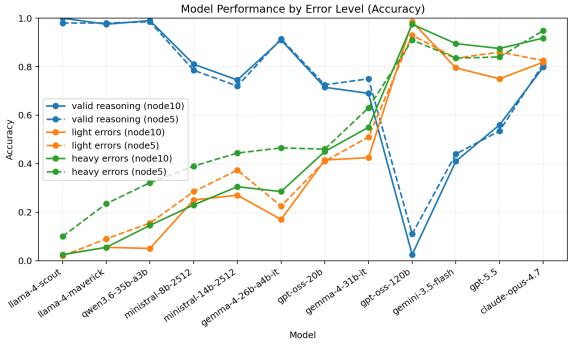

Figure 4: Model accuracy by class: valid and invalid arguments (light and serious errors)

oss-120b, which shows a pronounced bias toward the "invalid" category. As a result, it achieves an overall accuracy that overstates its actual capabilities. A similar, though weaker, pattern can be observed for Gemini 3.5 Flash, which also shows a clear tendency to predict "invalid." Only the frontier models GPT-5.5 and Claude Opus 4.7 clearly exceed the random baseline while also performing well across all individual classes. It therefore appears that BENFA constitutes a substantial challenge even for the strongest contemporary language models. To confirm this more firmly, however, a deeper analysis of labeling quality would be required, since in the present procedure labeling was performed automatically.

Real world vs. alternative worlds Finally, we consider the difference between argumentation situated in the real world and in fictional worlds. For both 10-node argumentation (53.2% vs. 54.1%) and 5-node argumentation (57.3% vs. 58.4%), the average model accuracy is higher for argumentation situated in the real world. This result suggests that our research hypothesis is on the right track, namely that model accuracy in argument mining is systematically overestimated. At the same time, the relatively small difference, in both cases approximately one percentage point, does not allow for a definitive confirmation of the hypothesis at the present level of noise, given statistical uncertainty and labeling errors that are difficult to estimate.

5 Conclusions and future work

This paper introduced two complementary ways of evaluating argument-mining competence under reduced reliance on factual familiarity. The syllogistic study showed that many models can perform well on formally controlled inference chains, but that performance is sensitive to reasoning mode, model size, prompt language, and output-format compliance. The BENFA benchmark, in turn, demonstrated that more natural and topically diverse non-factual arguments remain challenging even for strong contemporary models. In particular, many small and medium-sized models behaved close to chance level or exhibited strong class bias, whereas only the strongest frontier models showed consistently robust performance. Taken together, the results support the broader methodological claim of the paper: standard AM evaluation can overestimate reasoning ability when factual plausibility and argumentative validity are not carefully disentangled.

Future work should therefore proceed in a small number of targeted directions. First, the BENFA benchmark should be extended with additional manual or semi-manual validation in order to estimate labeling noise more reliably. Second, both studies should be rerun under systematically varied decoding and generation-length settings, to distinguish reasoning failures from failures of completion or format compliance. Third, the syllogistic benchmark should be evaluated under prompts in multiple languages, and BENFA should be expanded with richer error typologies. Fourth, example-level outputs and, where available, reasoning traces could support a separate qualitative analysis of recurring failure patterns. Finally, BENFA may be extended towards argument mining tasks beyond argument validity categorization.

Limitations

Several limitations should be noted. First, the two benchmarks differ substantially in style and construction, so their results should be compared qualitatively rather than treated as directly com-

mensurable scores. Second, only one additional retest was conducted, and only for Qwen2.5-7B in the syllogistic study. Third, the seed parameter of the syllogistic study was not fixed, which limits the reproducibility of individual generations. Fourth, both experiments used only one temperature setting, and other decoding parameters, including top_k, top_p, and min_p, were not systematically varied. Fifth, the syllogistic prompt was provided only in Polish, so the study does not allow conclusions about whether the same models would perform differently under prompts formulated in another language. Sixth, the BENFA benchmark was labeled automatically through the generation pipeline rather than through full human annotation, so some residual labeling noise is likely. Finally, both datasets are synthetic. This design increases experimental control and reduces the risk of memorized examples, but it also limits ecological validity and may not fully capture the distributional properties of naturally occurring argumentation.

References

Henrike Beyer and Chris Reed. 2025. Lexical recall or logical reasoning: Probing the limits of reasoning abilities in large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13532–13557, Vienna, Austria. Association for Computational Linguistics.

Boris Galitsky, Anton Chernyavskiy, and Dmitry Ilvovsky. 2024. Truth-o-meter: Handling multiple inconsistent sources repairing LLM hallucinations. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2817–2821. Association for Computing Machinery.

Bishwamittra Ghosh, Sarah Hasan, Naheed Anjum Arafat, and Arijit Khan. 2024. Logical consistency of large language models in fact-checking. *Preprint*, arXiv:2412.16100.

Shengding Hu, Yifan Luo, Huadong Wang, Xingyi Cheng, Zhiyuan Liu, and Maosong Sun. 2023. Won't get fooled again: Answering questions with false premises. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5626–5643, Toronto, Canada. Association for Computational Linguistics.

Arman Irani, Ju Yeon Park, Kevin Esterling, and Michalis Faloutsos. 2024. WIBA: What is being argued? a comprehensive approach to argument mining. *Preprint*, arXiv:2405.00828.

Henrike Jansen. 2007. Reductio ad absurdum vs. argument from consequences. *Informal Logic*, 27(3):249–266.

Henrike Jansen. 2011. "if that were true, i would never have...": The counterfactual presentation of arguments that appeal to human behaviour. In *Proceedings of the Seventh Conference of the International Society for the Study of Argumentation*, pages 881–889. Sic Sat.

Joonsuk Park and Claire Cardie. 2018. A corpus of eRulemaking user comments for measuring evaluability of arguments. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Mohammadamin Shafiei, Hamidreza Saffari, and Nafise Sadat Moosavi. 2025. MultiHoax: A dataset of multi-hop false-premise questions. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 10169–10187, Vienna, Austria. Association for Computational Linguistics.

Christian Stab, Tristan Miller, Benjamin Schiller, Pranav Rai, and Iryna Gurevych. 2018. Cross-topic argument mining from heterogeneous sources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3664–3674, Brussels, Belgium. Association for Computational Linguistics.

Douglas Walton, Chris Reed, and Fabrizio Macagno. 2008. *Argumentation Schemes*. Cambridge University Press, Cambridge.

Douglas Walton, Giovanni Sartor, and Fabrizio Macagno. 2016. An argumentation framework for contested cases of statutory interpretation. *Artificial Intelligence and Law*, 24:51–91.

Xinyan Yu, Sewon Min, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. CREPE: Open-domain question answering with false presuppositions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10457–10480, Toronto, Canada. Association for Computational Linguistics.

A Appendix

A.1 Study Design for Syllogistic Reasoning Evaluation

A.1.1 Input Datasets

Input length was measured as the total number of characters in the generated Polish text, including whitespace characters. For valid syllogistic chains with n = 3, the dataset contained 324 rows. The minimum input length was 513, the maximum input length was 3371, and the average input length was 956.49. The dataset for n = 4 contained 3141

rows, with a minimum input length of 659, a maximum input length of 6762, and an average input length of 1369.63. The largest valid dataset, `n = 5`, contained 27922 rows. In this case, the minimum input length was 686, the maximum input length was 10132, and the average input length was 1534.73.

The corresponding invalid datasets had closely matched length characteristics. For `n = 3`, the invalid dataset contained 324 rows, with a minimum input length of 513, a maximum input length of 3371, and an average input length of 956.79. Dataset for `n = 4`, the invalid dataset contained 3141 rows, with a minimum input length of 659, a maximum input length of 6762, and an average input length of 1369.93. And finally for `n = 5`, the invalid dataset contained 27922 rows, with a minimum input length of 686, a maximum input length of 10132, and an average input length of 1535.03.

These values show that the valid and invalid datasets were closely matched in length. Moreover, for each value of `n`, the number of valid syllogistic chains was identical to the number of invalid syllogistic chains.

A.1.2 Inference Configuration

Inference was performed using the vLLM library. The sampling configuration used a low temperature value of `temperature = 0.15`, in order to reduce output variability while still allowing standard model generation. The maximum model length was set to `max_model_len = 16492`. The `seed` parameter was not controlled. Each syllogistic chain was submitted to the model independently.

A.1.3 Qwen2.5-7B Retest

The retest was evaluated with respect to three outcome categories: correct responses, non-conforming responses, and `length` terminations. Across these categories, the retest results were generally consistent with the original run. The statistically significant differences observed for correct responses at `n = 5` had very small effect sizes, suggesting limited practical significance.

For the valid syllogisms dataset, the comparison between Run A and Run B showed no statistically significant difference in the number of correct responses for `n = 3`. Run A produced 254 correct responses, whereas Run B produced 260 ($\chi^2 = 0.339$, $p = 0.561$, Cramer's $V = 0.023$). For `n = 4`, Run A produced 1766 correct responses and Run B produced 1754 ($\chi^2 = 0.093$, $p = 0.760$, Cramer's $V = 0.004$). For `n = 5`, Run A produced 19426 correct responses and Run B produced 18979. This difference was statistically significant ($\chi^2 = 16.660$, $p = 0.000045$), although the effect size was very small (Cramer's $V = 0.017$).

The rate of non-conforming responses did not differ significantly between the two runs. For `n = 3`, Run A produced 8 non-conforming responses and Run B produced 7 ($\chi^2 = 0.068$, $p = 0.794$, Cramer's $V = 0.010$). For `n = 4`, Run A produced 136 non-conforming responses and Run B produced 123 ($\chi^2 = 0.681$, $p = 0.409$, Cramer's $V = 0.010$). For `n = 5`, Run A produced 1156 non-conforming responses and Run B produced 1119 ($\chi^2 = 0.627$, $p = 0.428$, Cramer's $V = 0.003$).

A similar pattern was observed for `length` terminations. For `n = 3`, Run A ended with a `length` stop in 4 cases, whereas Run B did so in 6 cases ($\chi^2 = 0.406$, $p = 0.524$, Cramer's $V = 0.025$). For `n = 4`, Run A produced 94 `length` terminations and Run B produced 92 ($\chi^2 = 0.022$, $p = 0.882$, Cramer's $V = 0.002$). For `n = 5`, Run A produced 762 `length` terminations and Run B produced 771 ($\chi^2 = 0.054$, $p = 0.816$, Cramer's $V = 0.001$).

Overall, the retest results for the valid syllogisms dataset were highly consistent with the original run. The only statistically significant difference was observed for correct responses at `n = 5`; however, the associated effect size was very small. The remaining comparisons for correct responses, non-conforming responses, and `length` terminations did not indicate statistically significant differences between the two runs.

For the invalid syllogisms dataset, the comparison between Run A and Run B showed no statistically significant difference in the number of correct responses for `n = 3`. Run A produced 246 correct responses, whereas Run B produced 243 ($\chi^2 = 0.075$, $p = 0.784$, Cramer's $V = 0.011$). For `n = 4`, Run A produced 1622 correct responses and Run B produced 1565. This difference was also not statistically significant ($\chi^2 = 2.069$, $p = 0.150$, Cramer's $V = 0.018$). For `n = 5`, Run A produced 18692 correct responses and Run B produced 18367. This difference was statistically significant ($\chi^2 = 8.473$, $p = 0.0036$), although the effect size was very small (Cramer's $V = 0.012$).

The number of non-conforming responses did not differ significantly between the two runs. For `n = 3`, Run A produced 5 non-conforming responses and Run B produced 4 ($\chi^2 = 0.113$, $p = 0.737$,

Cramer's $V = 0.013$). For `n = 4`, Run A produced 120 non-conforming responses and Run B produced 123 ($\chi^2 = 0.039$, $p = 0.844$, Cramer's $V = 0.002$). For `n = 5`, Run A produced 1040 non-conforming responses and Run B produced 1080 ($\chi^2 = 0.784$, $p = 0.376$, Cramer's $V = 0.004$).

A similar pattern was observed for `length` terminations. For `n = 3`, both Run A and Run B ended with a `length` stop in 1 case, with no difference between the two runs ($\chi^2 = 0.000$, $p = 1.000$, Cramer's $V = 0.000$). For `n = 4`, Run A produced 80 `length` terminations and Run B produced 84 ($\chi^2 = 0.100$, $p = 0.752$, Cramer's $V = 0.004$). For `n = 5`, Run A produced 698 `length` terminations and Run B produced 755. This difference was not statistically significant ($\chi^2 = 2.296$, $p = 0.130$, Cramer's $V = 0.006$).

Across both valid and invalid syllogisms datasets, statistically significant differences were observed only for correct responses at `n = 5`. In the valid dataset, this comparison yielded Cramer's $V = 0.017$, and in the invalid dataset Cramer's $V = 0.012$. Thus, although both differences were statistically detectable, their practical magnitude was very small. No statistically significant differences were observed for non-conforming responses or `length` terminations.

The retest of Qwen2.5-7B suggests that the results were broadly stable. Statistically significant differences were observed only for correct responses at `n = 5`, and in both the valid and invalid datasets the associated effect sizes were very small. No statistically significant differences were observed for non-conforming responses or `length` terminations. This supports the conclusion that the retest did not reveal a practically meaningful divergence from the original run, although the lack of a fixed `seed` means that stochastic variation was not fully controlled.

A.1.4 Detailed Results

Model	n=3	n=4	n=5
Bielik-11B*	0.833	0.811	0.770
Mistral-7B**	0.997	0.999	0.999
Ministral-3-8B***	0.583	0.669	0.550
gpt-oss-20b	0.892	0.916	0.875
Qwen2.5-7B	0.784	0.562	0.696
Qwen2.5-7B (retest)	0.802	0.558	0.680
Qwen3-8B	0.978	0.982	0.963
Qwen3-8B****	0.790	0.888	0.813
Qwen3.5-9B	0.935	0.950	0.930
Qwen3.5-9B****	0.265	0.443	0.301
Qwen3.5-4B	0.772	0.735	0.725
Qwen3.5-2B	0.012	0.015	0.019
Qwen3.5-0.8B	0.000	0.000	0.000
Total	324	3141	27922

* Bielik-11B-v2.1-Instruct
** Mistral-7B-Instruct-v0.3
*** Ministral-3-8B-Reasoning-2512
**** Reasoning disabled

Table 2: Accuracy on valid syllogism sequences.

Model	n=3	n=4	n=5
Bielik-11B*	0.559	0.621	0.636
Mistral-7B**	0.114	0.057	0.074
Ministral-3-8B***	0.809	0.771	0.793
gpt-oss-20b	1.000	0.999	1.000
Qwen2.5-7B	0.225	0.445	0.293
Qwen2.5-7B (retest)	0.238	0.463	0.304
Qwen3-8B	0.978	0.983	0.987
Qwen3-8B****	0.799	0.706	0.775
Qwen3.5-9B	0.997	0.996	0.997
Qwen3.5-9B****	0.997	0.966	0.982
Qwen3.5-4B	0.840	0.851	0.846
Qwen3.5-2B	0.201	0.165	0.169
Qwen3.5-0.8B	0.000	0.000	0.000
Total	324	3141	27922

* Bielik-11B-v2.1-Instruct
** Mistral-7B-Instruct-v0.3
*** Ministral-3-8B-Reasoning-2512
**** Reasoning disabled

Table 3: Accuracy on invalid syllogism sequences.

Model	n=3	n=4	n=5
Bielik-11B*	0.696	0.716	0.703
Mistral-7B**	0.556	0.528	0.536
Ministral-3-8B***	0.696	0.720	0.672
gpt-oss-20b	0.946	0.958	0.937
Qwen2.5-7B	0.505	0.504	0.495
Qwen2.5-7B (retest)	0.520	0.511	0.492
Qwen3-8B	0.978	0.983	0.975
Qwen3-8B****	0.795	0.797	0.794
Qwen3.5-9B	0.966	0.973	0.963
Qwen3.5-9B****	0.631	0.704	0.641
Qwen3.5-4B	0.806	0.793	0.786
Qwen3.5-2B	0.106	0.090	0.094
Qwen3.5-0.8B	0.000	0.000	0.000
Total	648	6282	55844

* Bielik-11B-v2.1-Instruct
** Mistral-7B-Instruct-v0.3
*** Ministral-3-8B-Reasoning-2512
**** Reasoning disabled

Table 4: Overall accuracy on syllogism sequences.

All results reported in this section are rounded to three decimal places. Tables 2, 3, and 4 present model accuracy for valid syllogistic chains, invalid syllogistic chains, and the combined dataset, respectively. The strongest overall performance was obtained by three models: gpt-oss-20b, Qwen3-8B with reasoning enabled, and Qwen3.5-9B. For gpt-oss-20b, overall accuracy was 0.946 for `n = 3`, 0.958 for `n = 4`, and 0.937 for `n = 5`. Qwen3-8B with reasoning enabled achieved 0.978, 0.983, and 0.975, respectively. Qwen3.5-9B achieved 0.966, 0.973, and 0.963.

A notable asymmetry was observed for Mistral-7B-Instruct-v0.3. On valid syllogistic chains, this model achieved near-perfect accuracy, with scores of 0.997, 0.999, and 0.999 for `n = 3`, `n = 4`, and `n = 5`, respectively. However, its accuracy on invalid chains was much lower, falling to 0.114, 0.057, and 0.074. This pattern suggests a strong tendency to classify syllogistic chains as valid, regardless of whether an invalid inference was present.

The results also show that the mere presence of a reasoning mode does not by itself guarantee high performance. Although the best-performing models in this study used reasoning, Ministral-3-8B-Reasoning-2512 achieved overall accuracy close to that of Bielik-11B-v2.1-Instruct, which did not use a reasoning mode. One possible explanation is that Bielik-11B was trained with substantial Polish-language data, whereas both the prompt and the generated syllogistic chains were expressed in Polish. At the same time, Ministral-3-8B-Reasoning-2512 performed worse than Qwen3-8B with reasoning disabled. On the other hand, the comparison between reasoning-enabled and reasoning-disabled variants of Qwen3-8B and Qwen3.5-9B indicates that reasoning improved performance for these models.

One of the aims of the study was to identify an approximate model-size threshold below which performance collapses. The Qwen3.5 model family was used for this purpose. Qwen3.5-4B achieved overall accuracy of 0.806, 0.793, and 0.786 for `n = 3`, `n = 4`, and `n = 5`, respectively. In contrast, Qwen3.5-2B dropped sharply to 0.106, 0.090, and 0.094. Qwen3.5-0.8B effectively failed on the task. Its overall accuracy was either zero or so close to zero that it had no practical relevance. The only non-zero result was observed for `n = 5`, where two outputs ended with the correct final label. Given the size of the dataset, this should be treated as incidental rather than as evidence of meaningful task performance.

Model	n=3	n=4	n=5
Bielik-11B*	0.000	0.000	0.000
Mistral-7B**	0.000	0.000	0.000
Ministral-3-8B***	0.000	0.000	0.000
gpt-oss-20b	0.003	0.003	0.004
Qwen2.5-7B	0.025	0.043	0.041
Qwen2.5-7B (retest)	0.022	0.039	0.040
Qwen3-8B	0.000	0.001	0.002
Qwen3-8B****	0.000	0.000	0.000
Qwen3.5-9B	0.065	0.048	0.067
Qwen3.5-9B****	0.000	0.000	0.000
Qwen3.5-4B	0.222	0.259	0.270
Qwen3.5-2B	0.985	0.984	0.978
Qwen3.5-0.8B	1.000	1.000	1.000
Total	324	3141	27922

* Bielik-11B-v2.1-Instruct
** Mistral-7B-Instruct-v0.3
*** Ministral-3-8B-Reasoning-2512
**** Reasoning disabled

Table 5: Non-conforming response rate across valid syllogism sequences.

Model	n=3	n=4	n=5
Bielik-11B*	0.000	0.000	0.000
Mistral-7B**	0.000	0.000	0.000
Ministral-3-8B***	0.000	0.001	0.000
gpt-oss-20b	0.000	0.000	0.000
Qwen2.5-7B	0.015	0.038	0.037
Qwen2.5-7B (retest)	0.012	0.039	0.039
Qwen3-8B	0.000	0.001	0.000
Qwen3-8B****	0.000	0.000	0.000
Qwen3.5-9B	0.003	0.004	0.003
Qwen3.5-9B****	0.000	0.000	0.000
Qwen3.5-4B	0.160	0.148	0.153
Qwen3.5-2B	0.793	0.829	0.827
Qwen3.5-0.8B	1.000	1.000	1.000
Total	324	3141	27922

* Bielik-11B-v2.1-Instruct
** Mistral-7B-Instruct-v0.3
*** Ministral-3-8B-Reasoning-2512
**** Reasoning disabled

Table 6: Non-conforming response rate across invalid syllogism sequences.

Model	n=3	n=4	n=5
Bielik-11B*	0.000	0.000	0.000
Mistral-7B**	0.000	0.000	0.000
Ministral-3-8B***	0.000	0.000	0.000
gpt-oss-20b	0.002	0.002	0.002
Qwen2.5-7B	0.020	0.041	0.039
Qwen2.5-7B (retest)	0.017	0.039	0.039
Qwen3-8B	0.000	0.001	0.001
Qwen3-8B****	0.000	0.000	0.000
Qwen3.5-9B	0.034	0.026	0.035
Qwen3.5-9B****	0.000	0.000	0.000
Qwen3.5-4B	0.191	0.204	0.212
Qwen3.5-2B	0.889	0.906	0.903
Qwen3.5-0.8B	1.000	1.000	1.000
Total	648	6282	55844

* Bielik-11B-v2.1-Instruct
** Mistral-7B-Instruct-v0.3
*** Ministral-3-8B-Reasoning-2512
**** Reasoning disabled

Table 7: Non-conforming response rate across all syllogism sequences.

Tables 5, 6, and 7 report the rate of non-conforming responses, that is, responses that did not end in either `TAK` or `NIE`. Using 0.010 as a practical threshold, such responses occurred most frequently for Qwen3.5-0.8B, Qwen3.5-2B, Qwen3.5-4B, and Qwen3.5-9B. They were also observed above this threshold for Qwen2.5-7B, both in the original run and in the retest. Non-conforming responses were more frequent for valid syllogistic chains than for invalid ones. For example, Qwen3.5-9B produced non-conforming responses on valid chains at rates of 0.065, 0.048, and 0.067, whereas on invalid chains the corresponding rates were only 0.003, 0.004, and 0.003. This suggests that valid chains were more likely to induce extended reasoning or delayed final answers.

The problem of non-conforming responses was concentrated mainly among reasoning models. The only non-reasoning exception of practical note was Mistral-7B-Instruct-v0.3, which produced six such responses for `n = 5` in the combined dataset (five for the valid dataset and one for the invalid dataset). No non-conforming responses were recorded for Bielik-11B-v2.1-Instruct, Qwen3-8B with reasoning disabled, or Qwen3.5-9B with reasoning disabled.

Model	n=3	n=4	n=5
Bielik-11B*	0.000	0.000	0.000
Mistral-7B**	0.000	0.000	0.000
Ministral-3-8B***	0.000	0.000	0.000
gpt-oss-20b	0.003	0.003	0.004
Qwen2.5-7B	0.012	0.030	0.027
Qwen2.5-7B (retest)	0.019	0.029	0.028
Qwen3-8B	0.000	0.001	0.002
Qwen3-8B****	0.000	0.000	0.000
Qwen3.5-9B	0.065	0.048	0.067
Qwen3.5-9B****	0.000	0.000	0.000
Qwen3.5-4B	0.222	0.263	0.273
Qwen3.5-2B	0.988	0.987	0.980
Qwen3.5-0.8B	1.000	1.000	1.000
Total	324	3141	27922

* Bielik-11B-v2.1-Instruct
** Mistral-7B-Instruct-v0.3
*** Ministral-3-8B-Reasoning-2512
**** Reasoning disabled

Table 8: Rate of 'length' stop reasons across valid syllogism sequences.

Model	n=3	n=4	n=5
Bielik-11B*	0.000	0.000	0.000
Mistral-7B**	0.000	0.000	0.000
Ministral-3-8B***	0.000	0.000	0.000
gpt-oss-20b	0.000	0.000	0.000
Qwen2.5-7B	0.003	0.025	0.025
Qwen2.5-7B (retest)	0.003	0.027	0.027
Qwen3-8B	0.000	0.001	0.000
Qwen3-8B****	0.000	0.000	0.000
Qwen3.5-9B	0.003	0.004	0.003
Qwen3.5-9B****	0.000	0.000	0.000
Qwen3.5-4B	0.188	0.162	0.166
Qwen3.5-2B	0.818	0.846	0.844
Qwen3.5-0.8B	1.000	1.000	1.000
Total	324	3141	27922

* Bielik-11B-v2.1-Instruct
** Mistral-7B-Instruct-v0.3
*** Ministral-3-8B-Reasoning-2512
**** Reasoning disabled

Table 9: Rate of 'length' stop reasons across invalid syllogism sequences.

The distribution of non-conforming responses appears to be related to the exhaustion of the model's maximum generation length, defined in this study as `max_model_len = 16492`. Tables 8 and 9 report the rate of `length` stop reasons among all inference termination reasons. Apart from normal completion and `length` termination, no other termination reasons were observed. Models reached the length limit more often for valid syllogistic chains than for invalid ones. This is consistent with the structure of the task. To establish that a long valid chain is correct, the model must in principle verify all syllogistic steps. In contrast, an invalid chain can be rejected as soon as the first invalid syllogistic step is detected; in this study, such an invalid step was introduced near the beginning of the chain. This pattern supports the hypothesis that, at least in some cases, model behavior somehow reflects the structure of the verification process that a human reasoner would also have to perform.

Table 1 presents overall accuracy after excluding non-conforming responses. The result for Qwen3.5-0.8B at `n = 5`, where accuracy reached 0.667, should be treated as a statistical artifact. Out of 55844 outputs, only three were conforming, and two of them happened to be correct. More informative is the result for Qwen3.5-2B. Although its overall accuracy was very low when non-conforming responses were counted as failures, its accuracy among conforming responses was comparable to that of substantially larger Qwen3.5 models. This suggests that increasing `max_model_len` could reduce the number of `length` terminations and possibly increase the number of correctly completed responses. This interpretation should, however, be treated as a hypothesis requiring further experimental verification.

B Nonfactual Hybrid Benchmark

This appendix documents the implementation details of the hybrid non-factual argument generation pipeline and of the subsequent validity-evaluation procedure. Unless stated otherwise, the descriptions below refer to the active code used in the present repository, rather than to older deprecated pipeline variants.

B.1 Generation Pipeline Overview

The dataset-generation script operates over a structured inventory of input items. Each item specifies:

- a category,
- an item identifier,
- a world description,
- a fictional topic,
- a real-world topic,
- an empirical-distance value,
- and a set of admissible genres.

The active inventory contains 50 input items distributed across 10 categories, with 5 items per category:

- PHY: altered physical or natural laws
- BIO: altered biology or ecology
- HSO: alternative history or societies
- LEG: different legal laws
- RPH: alternative religious or philosophical doctrines
- SPO: alternative sports rules, events or players
- EVD: certain aspects of mundane everyday life are altered
- LIT: a certain literary work has altered content
- VIS: a certain visual media has altered content

Model	Provider
claude-opus-4.7	Anthropic
gemini-3.5-flash	Google
gemma-4-26b-a4b-it	Google
gemma-4-31b-it	Google
llama-4-maverick	Meta
llama-4-scout	Meta
ministral-14b-2512	Mistral AI
ministral-8b-2512	Mistral AI
gpt-5.5	OpenAI
gpt-oss-120b	OpenAI
gpt-oss-20b	OpenAI
qwen3.6-35b-a3b	Qwen/Alibaba

Table 10: Models evaluated on BENFA

- FAN: fantasy and science-fiction worlds

For each item and each selected graph size, the script runs two symmetric generation arms:

- **F2R**: fictional argument map → fictional text → minimally adapted real-world text;
- **R2F**: real-world argument map → real-world text → minimally adapted fictional text.

For each arm, the script produces one valid version and two invalid versions:

- `1`: valid fictional-side text,
- `2a`: fictional-side text with light reasoning errors,
- `2b`: fictional-side text with serious reasoning errors,
- `3`: valid real-world text,
- `4a`: real-world text with light reasoning errors,
- `4b`: real-world text with serious reasoning errors.

Thus, each item–node-count pair yields a family of 12 related argumentative texts. If one uses 50 items and two graph sizes (5 and 10 nodes), the resulting dataset contains

$$50 \times 2 \times 12 = 1200$$

texts.

B.2 Models

B.3 Sampling Parameters and Runtime Settings

The active generation script uses the following defaults:

- model: `anthropic/claude-opus-4.7`,
- temperature: `0.7`,
- maximum output length: `2500` tokens,
- random seed: `42`,

Genre choice is randomized once per item–node-count job, but the randomization is deterministic with respect to the fixed seed and traversal order. The script writes one partial CSV per item–node-count pair and then merges these partial files in job-index order, which preserves deterministic row ordering even when concurrent workers are used.

For evaluation, the following parameters are used:

- temperature: `0.0`,
- maximum output length: `65536` tokens,

The evaluation script also supports multi-model runs by passing a list of model identifiers.

B.4 Genre Inventory and Noise Profiles

Genre metadata and noise settings are defined directly in the generation script. The six active genres are as follows:

- **internet**: An internet discussion post or thread-style comment exchange (e.g., forum, social media, comment section). Noise level: Medium; informal wording, interruptions, rhetorical shortcuts, occasional slang, but still coherent.
- **science**: A fragment of a peer-reviewed scientific journal article. Noise level: Low; formal, precise, and structured, with minimal stylistic noise.
- **oped**: An op-ed newspaper column written by an opinionated journalist. Noise level: Medium; persuasive and stylized, with occasional rhetorical flourish, but still clear.
- **blog**: An online blog post, forum essay, or podcast transcript. Noise level: Medium-high; discursive structure, personal asides, uneven pacing, but coherent overall.
- **debate**: A transcript of a formal or semi-formal debate (e.g., televised political debate or parliamentary exchange). Noise level: Medium; interruptions, repetition, framing tactics, and reactive phrasing.

- **coffee**: A dialogue transcript of a casual conversation (e.g., two coworkers over coffee). Noise level: High; colloquial speech, ellipses, false starts, and everyday conversational drift.

Across the 50-item inventory, the number of items permitting each genre is:

- `oped`: 50,
- `blog`: 46,
- `debate`: 37,
- `internet`: 35,
- `coffee`: 21,
- `science`: 11.

B.5 empirical-Distance Distribution

Each input item is annotated with an empirical-distance value on a 1–5 scale. In the current inventory, the distribution is:

- distance 1: 19 items,
- distance 2: 11 items,
- distance 3: 8 items,
- distance 4: 7 items,
- distance 5: 5 items.

B.6 Parsing Strategy for Generation Outputs

The corruption stage returns two explicitly delimited sections:

```
===CORRUPTED MAP===
...
===CORRUPTION NOTES===
...
```

The parser searches for these two delimiters. If both delimiters are present, the text between them is stored as the corrupted map and the trailing section is stored as corruption notes. If the delimiters are missing, the full response is retained as map text and the notes field is left empty. This fallback was included for robustness against occasional formatting failures.

The final output CSV produced by the generation pipeline contains the following columns:

- `item_id`,
- `category`,
- `empirical_distance`,
- `direction`,
- `variant`,
- `node_n`,
- `genre`,
- `noise_level`,
- `world_desc`,
- `arg_map`,
- `corruption_notes`,
- `text`.

B.7 Parsing Strategy for Evaluation Outputs

Evaluation outputs are parsed by a deliberately simple rule. The script converts the model output to uppercase and searches for the first standalone occurrence of either `VALID` or `INVALID`. If such a label is found, it is taken as the model prediction. Otherwise the output is assigned the label `UNKNOWN`. Three failure cases are distinguished:

- empty output,
- API error output,
- output containing neither `VALID` nor `INVALID`.

Ground truth is derived from the variant label:

- variants beginning with `1` or `3` are treated as `VALID`,
- variants beginning with `2` or `4` are treated as `INVALID`.

World scope is also derived from the variant label:

- variants beginning with `1` or `2` are treated as fictional,
- variants beginning with `3` or `4` are treated as real.

For summary tables, invalid items are further partitioned by error level:

- `2a/4a`: light errors,
- `2b/4b`: heavy errors.

B.8 Reproducibility Notes

Several implementation details are relevant for replication:

- Genre sampling is deterministic given the seed and the traversal order.
- The same genre and noise profile are used throughout the 12-text family corresponding to a single item–node-count job.
- Invalid maps are created by minimally perturbing an already generated valid map, rather than by generating invalid texts independently.
- In both translation directions, the script attempts to preserve the argument map exactly, including flawed reasoning where present.

C Budget and computation infrastructure

All inference was run via Openrouter.ai. The inference costs amounted to an estimated 110 USD.

D AI assistants

The authors used AI tools to assist in coding and writing. All AI generated content was carefully reviewed.

E Personally Identifying Info Or Offensive Content

All data was checked for potentially personally identifying or offensive content.