Zero-Shot Fact Verification via Natural Logic and Large Language Models

Anonymous ACL submission

Abstract

The recent development of fact verification (FV) systems with natural logic has enhanced the explainability of these systems by aligning claims with evidence through set-theoretic operators, providing justifications that faithfully expose the model's reasoning. Despite these advancements, such systems often rely on a large amount of training data annotated with natural logic. To address this issue, we propose a zero-shot method that utilizes the generalization capabilities of instruction-tuned large language models. Our system uses constrained decoding to mitigate hallucinations and employs weighted prompt ensembles to stabilize results in a zero-shot setup. Compared to previous studies that mostly evaluated FV systems with natural logic on artificial claims, we focus on benchmarking the generalization capabilities of our approach on datasets with naturally occurring claims, which are more complex and cover various domains. Our method surpasses the best baselines by 6.1 Macro-F1 points on average, outperforming both fully-trained and few-shot FV systems. We also evaluate its effectiveness on Danish and Chinese fact-checking datasets, demonstrating its multilingual capabilities. Additionally, we demonstrate that our method's performance scales with the size of the language model and show that incorporating natural logic not only provides faithful justifications but also directly improves performance over models without natural logic.

1 Introduction

In the context of fact-checking, fact verification (FV) is a process of verifying whether a textual hypothesis holds based on retrieved evidence. Nevertheless, the lack of explainability is problematic as verdict justification is a central component of fact-checking (Uscinski and Butler, 2013).

While there have been attempts to improve explainability with attention-based methods (Shu et al., 2019; Popat et al., 2018) and post-hock summarizations (Atanasova et al., 2020; Kotonya and Toni, 2020), these approaches do not provide *faithful justifications* — explanations that accurately reflect the model's decision-making process and the data it used (Jacovi and Goldberg, 2020). In contrast, systems such as NaturalLI (Angeli and Manning, 2014) and ProoFVer (Krishna et al., 2022) focus on capturing the relations between various pieces of provided information and use natural logic to provide faithful justifications.

While natural logic-based FV systems (illustrated in Figure 1) provide improved explainability, they often require a large amount of training data. For example, ProoFVer (Krishna et al., 2022) was trained on 100K+ instances obtained from structured knowledge bases such as PPDB (Ganitkevitch et al., 2013) and Wikidata (Vrandečić and Krötzsch, 2014). Although recent few-shot learning methods aim to address this issue (Aly et al., 2023), they still demand training instances annotated with natural logic, which can be expensive to acquire.

In order to address these limitations, we investigate the application of large language models (LLMs) to fact verification in zero-shot settings. We propose **Zero-NatVer**, a zero-shot fact verification system that uses instruction-tuned LLMs to generate faithful proofs with natural logic. We demonstrate its performance with one of the largest open source models *Llama2* (Touvron et al., 2023) as well as with a commercial model *ChatGPT* (OpenAI, 2023). Our experiments show competitive results, surpassing models that were fully supervised or few-shot trained for fact-checking with natural logic. We outperform the best baseline by 4.4 Macro-F1 points on average when benchmarked on English datasets with natural claims. Furthermore, we investigate the use of evidence rephrasing to address the limitation of problematic alignments. By rephrasing relevant evidence, we further improve our results, and Zero-NatVer sur-

passes the best baseline by 6.1 F1-Macro points on average. Additionally, our study explores the system's adaptability to datasets in languages other than English. We test our method on Danish and Chinese fact-checking datasets and demonstrate that Zero-NatVer can generalize to other languages.

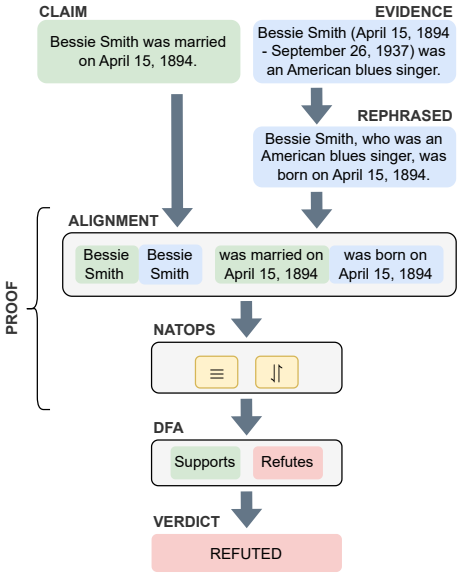

Figure 1: **A high-level view of natural-logic based FV systems.** The aligned chunks from claim and evidence texts are classified with natural logic operators ≡ and ⇂↾. These operators are then used as transitions in the DFA, and the final state (i.e. Refutes) determines the verdict. This example also shows evidence rephrasing, which is specific to Zero-NatVer.

Contributions This research introduces a novel approach to zero-shot fact verification using natural logic. Our proposed system leverages the robust generalization capabilities of LLMs, and we demonstrate its effectiveness in handling a variety of real-world claims across multiple domains across open-source (Llama2) and proprietary (ChatGPT) language models. Additionally, our study explores the system's adaptability to datasets in languages other than English. A notable aspect of our findings is the positive impact of evidence rephrasing, which further improves the overall efficiency of our system.

2 Related Work

One class of neuro-symbolic reasoning systems for FV uses simple logical rules to aggregate veracity information on a claim's components to provide a simple faithful explanation (Stacey et al., 2022, 2023; Chen et al., 2022). In contrast, the paradigm of natural logic (Van Benthem, 1986; Sanchez, 1991) and NaturalLI (Angeli and Manning, 2014), composes full inference proofs that operate directly on natural language, capable of expressing more complex logical relationships between claim and evidence, such as double-negation. Krishna et al. (2022) train natural logic inference systems for fact verification, achieving competitive performance while remaining faithful and more explainable than its entirely neural counterpart. While these neural-symbolic approaches require substantial training data to perform well, Aly et al. (2023) explore natural logic inference in a few-shot setting by casting natural logic operators into a question answering framework, subsequently making use of the generalization capabilities of instruction-tuned language models. While our work also uses question answering to predict natural logic operators, we further address prediction calibration issues frequently encountered in a zero-shot setting (Kadavath et al., 2022; Jiang et al., 2023).

Previous work on zero-shot FV is limited and largely rely on the generation of weakly supervised training samples and on knowledge of the target domain (Pan et al., 2021; Wright et al., 2022). Pan et al. (2023b) observe that typical FV systems fail when transferred to unseen domains in a zero-shot setting. They also propose a data augmentation technique to improve generalizability. However, none of the aforementioned zero-shot methods produces (faithful) explanations.

In a few-shot setting, several recent works have explored the use of large language models that produce explanations alongside the verdict. Pan et al. (2023a) define a reasoning program, consisting of a sequence of subtasks, to verify complex claims. Yao et al. (2023) proposes chain-of-thought prompting complemented by action operations to support the model's reasoning and its explanation generation. Li et al. (2023) propose to edit rationales generated via chain-of-thought prompting by querying knowledge sources. Yet, in contrast to this work these approaches still rely on in-context examples.

3 Method

Given a claim c and a set of k evidence sentences $e_1, e_2, ..., e_k \in E$, our system determines the veracity label y, which denotes whether the information from E supports c, refutes c, or whether there is not enough information to support or contradict c. Zero-NatVer obtains the verdict in four steps.

In the first step, we address the fact that complex claim and evidence sentences can vary considerably in terms of their syntactical structures, resulting in inaccuracies during chunking and alignment. Thus, we first rephrase evidence E into R so that relevant information is syntactically closer to the claim c while staying logically consistent with E. In the second step, Zero-NatVer aligns all facts in c with relevant information from R. This process results in a sequence of l claim-evidence pairs $P = p_1, p_2, ..., p_l$. Next, Zero-NatVer determines the relation of each pair in terms of natural logic and generates a sequence of natural logic operators (NatOps) $O = o_1, o_2, ..., o_l$, which correspond to claim-evidence pairs in P. Finally, O is used to traverse a deterministic finite state automaton (DFA). The final state of the DFA then determines the veracity of the claim. The following sections describe each step in more detail.

3.1 Evidence Rephrasing

Fact-checking systems based on natural logic typically assume that claim and evidence texts can be split and aligned into meaningful claim-evidence pairs that can be individually resolved in terms of their natural logic relations. While these systems showed impressive performance on artificial claims where claims and evidence are syntactically similar (Krishna et al., 2022; Aly et al., 2023), real-life claims and evidence can challenge this assumption due to the complexity and variability inherent in natural language. For example, the fact that the dates in the phrase "Bessie Smith (April 15, 1894 - September 26, 1937)" refer to the birth and death dates of Bessie Smith is obvious only after seeing the full sentence. After the chunking and alignment process, spans can often lose a relevant context and become more ambiguous, leading to incorrect verdicts. In this example, a hypothetical claim about her birth date could be incorrectly aligned only with the relevant date (i.e. April 15, 1894), complicating the NatOp assignment in the next stage of the process.

In this work, we address this problem by prompting a language model to rephrase the evidence text and make it syntactically closer to the claim text before it gets chunked and aligned. The full prompt template can be found in the Appendix in Listing 1. As shown in Figure 1, we can use an LLM to rephrase the previous phrase into "Bessie Smith, who was an American blues singer, was born on April 15, 1894", which reorganizes relevant parts of the evidence and expands the date by the verb "born" allowing now for a comparison with the verb "married". Other examples of situations where rephrasing can be beneficial include anaphora resolution, acronym expansion, or counting problems.

While rephrasing can resolve some of the shortcomings of natural logic-based FV, we must be careful not to change the meaning of the evidence. Even though we instructed the model not to change the meaning of the original evidence sentence, the LLM could still skip some information or hallucinate new facts, changing the final verification verdict. In our work, we mitigate these problems by using a constrained beam-search decoding approach (Anderson et al., 2016). At each decoding step, the model keeps track of several most likely partial sequences, and it constrains sequences that contain prohibited words. A prohibited word is any non-stop word from the claim that does not appear in the evidence text. For example, this prevents the model from making the rephrased sentence even more similar by copying over the word "married".

3.2 Chunking and Alignment

FV systems based on natural logic require additional preprocessing of claims and evidence before they can determine NatOps and final verdicts. This preprocessing traditionally consists of two consecutive steps— chunking and alignment. The chunking process segments both c and E into smaller, manageable pieces (chunks), and the alignment step links each claim chunk with a single evidence chunk, ideally providing enough information for support or contradiction. We treat chunking and alignment as a joint task and perform it in a single zero-shot inference pass by switching between two decoding schemes— one scheme for chunking and one for alignment.

The decoding strategy for chunking considers the claim text as a queue of tokens Q_C to be processed. It then chunks the claim as follows:

1. The decoding is prefixed with the phrase *"Step 1: Here are..."* to encourage the generation of claim chunks.
2. The model generates tokens from Q_C until it encounters a space character. For logographic languages, it stops after every token.
3. The model is constrained to sample one of two outputs - the next token from Q_C or a new-line character.

Claim

Miracle at St. Anna tells the story of four soldiers.

Evidence

From the article "Miracle at St. Anna": Set primarily in Italy during German-occupied Europe in World War II, the film tells the story of four Buffalo Soldiers of the 92nd Infantry Division who seek refuge in a small Tuscan village, where they form a bond with the residents.

Output

Step 1: Here are the segmented chunks of the claim:
* Miracle at St. Anna
* tells the story of
* four soldiers

Step 2: Here are the alignments of each claim segment with the corresponding relevant information from the evidence text:
* "Miracle at St. Anna" (claim) -> "Miracle at St. Anna" (evidence)
* "tells the story of" (claim) -> "the film tells the story of" (evidence)
* "four soldiers" (claim) -> "four Buffalo Soldiers of the 92nd Infantry Division" (evidence)

Figure 2: **2-step decoding scheme for chunking and alignment.** The green text refers to generated claim segments, and the blue text refers to generated evidence-alignments. The remaining text was forced during the decoding.

4. Repeats steps 2-3 until Q_C is empty (all claim tokens are consumed).

As shown in Figure 2, the outcome of this process is a bulleted list of claim segments. Due to the constraints at each decoding step, this generation cannot hallucinate, skip words, or alter information from the claim.

Keeping the generated output in the context, we then switch to the second decoding strategy that considers the previously generated chunks as a queue of segments Q_A to be processed. It then aligns each segment with relevant information from the provided evidence as follows:

1. The decoding is prefixed with the phrase *"Step 2: Here are..."* to encourage the generation of alignments.
2. The model is prefixed with a chunk from Q_A.
3. Constrained decoding is used to sample the alignment text.
4. Repeats steps 2-3 until Q_A is empty.

As shown in Figure 2, the outcome of this process is a bulleted list of claim-evidence segments. We also use additional markers such as *"*"*,*"(claims)"*, *"(evidence)"*, and *"->"* to denote each section. These markers help with consistency and maintain the intended format and behavior, which might otherwise not be obvious to the model in a zero-shot setting.

While the first decoding strategy is constrained by design and does not allow for hallucinations, the decoding in step 3 of the second strategy relies on general sampling and needs to be constrained. We first tried constraining the generation only to allow tokens that form the evidence text E or the rephrased evidence text R. However, we empirically found that this approach struggles in situations like combining two or more pieces of information that do not occur directly next to each other in the evidence text. Therefore, we use the same constraining approach as for rephrasing, instead. Thus, we prohibit claim tokens that do not exist in E. Then, we post-process the alignments and remove any text that does not form sequences of tokens in E. This approach ensures that the aligned text comprises only sub-strings from E or R.

3.3 NatOp Assignment via QA Ensembles

Having alignments between claim and evidence, the next step is to determine a NatOp for each claim-evidence pair. Similar to Aly et al. (2023), we consider them as relations that can be resolved via questions over claim-evidence spans. Thus, we prompt our model with *Yes/No* questions to determine whether a relation can be expressed by one of the NatOps. Using questions-answering, we consider the following NatOps: Equivalence ($\equiv$), Forward Entailment ($\sqsubseteq$), Backward Entailment ($\sqsupseteq$), Negation ($\neg$), and Alternation ($|\!\uparrow$). For example, for the negation NatOp, we can ask the question *"Is the phrase X a negation of Y?"*, where *X* and *Y* represent claim and evidence spans, respectively.

While LLMs tend to perform reasonably well

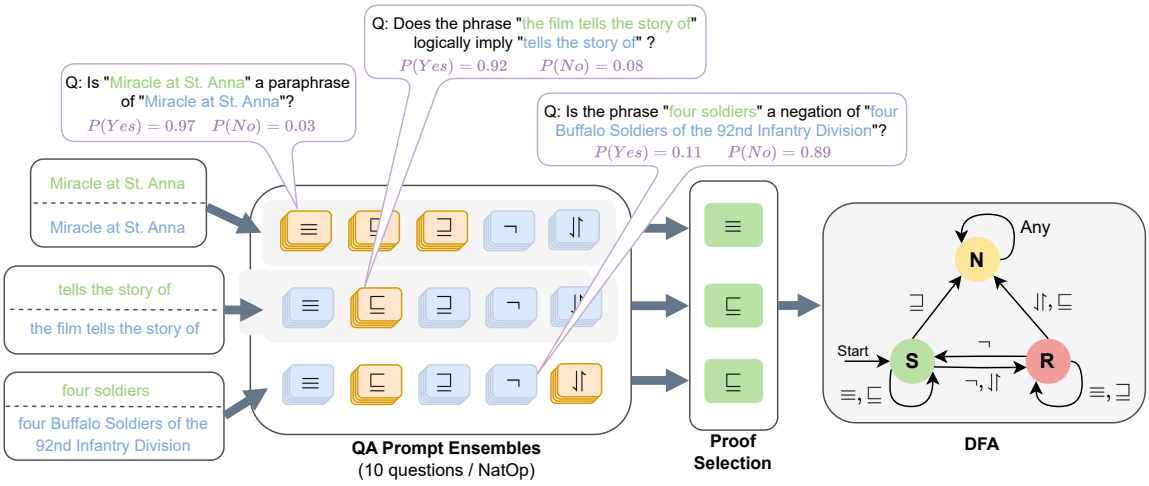

Figure 3: **Verdict generation from aligned claim-evidence pairs.** We use prompt ensembles to determine the relevancy of each NatOp to a particular claim-evidence pair, resulting in a set of NatOp candidates (orange blocks). A single NatOp (green blocks) is then selected among the candidates in the proof selection set, and the sequence of NatOps (one for each claim-evidence pair) is then passed to the DFA, which determines the final verdict.

on tasks such as semantic exclusion, they are also overly sensitive to small changes in their prompts, resulting in inconsistent outcomes (c.f. Section 2). In order to reduce the variability of outcomes, we use a large number of *Yes/No* questions to prompt the model, thereby obtaining several micro-judgments per NatOp which are then aggregated as a weighted average. Instead of manually hand-crafting these question templates, we prompt the *Llama2* model to generate them. This approach ensures the questions are more aligned with the model's distribution. In our experiments, we employ 10 templates for each NatOp, though it is easy to generate and use additional templates.

For a given claim-evidence pair a and operator o, we compute a NatOp score $s_{o,a}$ as a weighted average over all micro-judgments:

$$s_{o,a} = \sum_{i=1}^{N} w_i P(Yes|T_i, a) \tag{1}$$

where T is a collection of prompt templates, and w represents confidence weights for each template, with $\sum_{i=1}^{N} w_i = 1$.

We compute w by iterating over the entire dataset in a single pass and capturing the log-likelihood scores for each template. For each instance, we always capture only the Yes/No option, which has the higher log-likelihood score (i.e. the option that the model favors more).

Using Equation 1, we then compile a list of NatOps candidates C, considering only $s_{o,a} > \alpha$, where α can be seen as a confidence threshold for the model. Since we are not using any validation data to determine hyper-parameters, we set $\alpha = 0.5$ as we are considering two output classes.

Natural logic entailment relationships are by set-theory definition mutually exclusive (MacCartney and Manning, 2008), and so a should be associated with exactly one NatOp. However, since we don't train the model on this task and questions over alignments can often be non-trivial, it can happen that $|C| > 1$. Therefore, we must resolve this conflict and select a single NatOp from C. Due to the design of the DFA, a single incorrectly chosen NatOp can lead to the *Not Enough Evidence* state, from which there are no outgoing transitions to other states. Aly et al. (2023) deals with this problem by considering multiple alignments at different granularities and choosing the proofs with the highest verdict score. Instead, our method simply uses a NatOp priority approach and prioritizes operators in the following order: $[\equiv, \neg, \sqsubseteq, \sqsupseteq, \downarrow\uparrow]$. Thus, we choose the operator from C that occurs at the leftmost position in the priority list. We defined the NatOp order by considering the difficulty of each NatOp task. For instance, in a scenario where the candidate list C consists of equivalence ($\equiv$) and alternation ($\downarrow\uparrow$), we postulate that identifying equivalence (i.e., assessing textual similarity) is a simpler task compared to identifying alternation (i.e., recognizing non-exhaustive exclusion). We decided on this order prior to running our experiments and did not optimize this order.

	Climate-FEVER		PubHealth		SciFact	
	Macro-F1	Micro-F1	Macro-F1	Micro-F1	Macro-F1	Micro-F1
Transfer						
ProoFVer	40.70	43.35	36.21	**51.65**	45.57	49.16
QA-NatVer	44.74	47.43	28.35	36.05	52.02	56.67
Pan et al. (2023)	40.60	-	36.46	-	50.71	-
Generalization						
ProoFVer	26.63	34.75	14.91	16.89	25.58	34.67
QA-NatVer	22.20	36.86	12.96	14.73	23.56	40.67
Zero-NatVer	44.71	46.78	36.19	42.64	57.47	60.33
Zero-NatVer-R	**45.78**	**49.38**	**36.44**	46.03	**61.07**	**64.00**

Table 1: **Final results on different fact-checking datasets in zero-shot generalization and zero-shot transfer setups.** The numbers in **bold** denote the best scores for a given dataset across all models.

4 Experimental Methodology

4.1 Zero-shot Setups

To better assess the zero-shot capabilities of our approach, we differentiate between two types of zero-shot setups– **zero-shot generalization** and **zero-shot transfer**. We define zero-shot generalization as a model's ability to handle entirely new tasks or domains it has not encountered during training. Conversely, zero-shot transfer refers to training a model on a specific task or dataset and subsequently applying it to a different but related task or dataset without further training. For example, consider a model trained on a broad spectrum of general data, such as BART, T5, or Llama2. Applying this model to an entirely new task like FV using natural logic, which was not part of its training, exemplifies zero-shot generalization according to our definition. In contrast, if the same model is fine-tuned on a FV dataset like FEVER and applied to perform FV on a different dataset, this would be an instance of zero-shot transfer.

4.2 Datasets

Previous works on NLI-based FV models mainly examined the performance on artificial claims from FEVER-like datasets (Krishna et al., 2022; Aly et al., 2023; Chen et al., 2023). However, these datasets tend to cover more general topics, and artificial claims are often rather simple in their structure. For a more comprehensive assessment of zero-shot capabilities, we evaluate our method on natural claims from datasets Climate-FEVER (Diggelmann et al., 2020), PubHealth (Kotonya and Toni, 2020), and Scifact (Wadden et al., 2020) (See Appendix A for mode details).

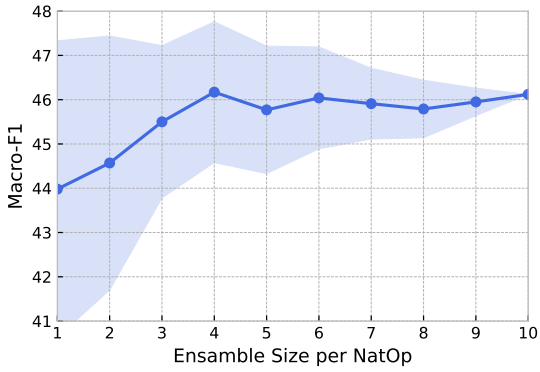

Figure 4: Macro-F1 scores across all datasets for various ensemble sizes. The light blue area represents the standard deviation from 10 independent measurements.

5 Results

Our proposed approach represents the zero-shot generalization setup. We conducted our main experiments with the Llama2-70B model, one of the largest open-source LLMs to date. Crucially, we did not fine-tune the model on any specific dataset, and we did not tune any hyperparameters. The only exposure to fact-checking datasets was when we were designing our prompts. In this context, we used Symmetric-Fever (Schuster et al., 2019), which differs from other evaluated datasets by using only 2 classes (i.e., SUPPORTS/REFUTES). We selected a small subset of 100 claims and tested that our prompts generated the desired responses in the desired format. For hyperparameters, we have adopted the recommendations of Perez et al. (2021) and did not rely on hyperparameters from prior works. More details can be found in Appendix C.

Baselines Our natural-logic-based baselines consist of ProoFVer (Krishna et al., 2022) and QA-NatVer (Aly et al., 2023). For zero-shot transfer

setups, we use ProofVer trained on 100K+ FEVER instances and QA-NatVer trained on 32 instances of NLI annotated FV data. For zero-shot generalization setups, we run both models with their corresponding pre-trained LLMs without fine-tuning on NLI data. We also include results reported by Pan et al. (2023b) as an additional baseline for zero-shot transfer experiments. More details about our baselines can be found in Appendix B.

Main Results We report the main results in Table 1, where we compare Zero-NatVer to other baselines in two zero-shot setups. Although rephrasing of evidence, as detailed in Section 3.1, is an important component of our approach, we consider it an optional step. This is due to its potential to alter evidence, which may subsequently impact the clarity of explanations. Consequently, we always report our results without evidence rephrasing (Zero-NatVer) and with rephrasing (Zero-NatVer-R) separately. This separation allows for assessing the trade-off between explainability and performance. Additionally, we report both macro-averaged F1 (Macro-F1) and micro-averaged F1 (Micro-F1) metrics because classes in all datasets are imbalanced.

Comparing our method with other zero-shot generalization baselines, our system achieves 46.12 Macro-F1 and 49.92 Micro-F1 points on average without evidence rephrasing. On average, our method outperforms ProofVer by 23.75 Macro-F1 and 21.15 Micro-F1 points, and it outperforms QA-NatVer by 26.55 Macro-F1 and 19.16 Micro-F1 points. Our results further improve if we rephrase evidence in the first state of the pipeline. Zero-NatVer-R achieves additional 1.6 Macro-F1 and 3.2 Micro-F1 points on average across all datasets.

Our method also surpasses all baselines in zero-shot transfer setups, although the gap between models is much smaller. On average, we outperform ProofVer by 5.3 Macro-F1 and 1.86 Micro-F1 points, QA-NatVer by 4.42 Macro-F1 and 3.2 Micro-F1 points, and we outperform the experimental results of Pan et al. (2023b) by 3.53 Macro-F1 and 3.2 Micro-F1 points. We were not able to obtain the Micro-F1 metrics of their results.

Ensemble size To assess the impact of the prompt ensemble size (Section 3.3), we run an experiment measuring performance across all datasets for various ensemble sizes. For each measured ensemble size S, we randomly sample S prompts for each NatOp from our prompt bank. We repeat this process 10 times and report means and standard deviations for each ensemble size in Figure 4.

The results show that the size of prompt ensembles has a large impact on the variability of outcomes. When using only one question per NatOp and sampling different prompts, we obtain Macro-F1 scores with a standard deviation of 3.59 points. In comparison, an ensemble of only 4 prompts substantially reduces this variation by more than half.

Model size Figure 3 compares our method with different sizes of Llama2 models, showing a substantial growth in performance as the model scales up. Additionally, we evaluated our method on the proprietary model ChatGPT-3.5, which is allegedly larger in size than our Llama2 models. The low scores for ChatGPT-3.5 can be caused by API limitations, which prevented us from using constrained decoding and weighted prompting (see appendix D for prompting details).

Ablation Study We perform three different ablation studies on SciFact, as reported in Table 2. First, we examine the performance of Zero-NatVer and Zero-NatVer-R without weighting ensemble prompts, observing a small drop of 0.72 and 0.75 Macro-F1 points, respectively. Second, we ablate

	Macro-F1	Micro-F1
Zero-NatVer	**57.47**	**60.33**
w/o weighted prompts	56.52	59.33
w/o prompt ensemble	49.56	53.67
w/o constrained decoding	55.45	58.00

Table 2: Ablation study on SciFact.

	Macro-F1	Micro-F1
Llama2-7B	20.57	41.67
Llama2-13B	30.96	42.16
Llama2-70B	57.47	60.33
GPT-3.5-Turbo	49.21	53.00

Table 3: SciFact results for different Llama-2 model sizes and ChatGPT.

	Macro-F1	Micro-F1
Llama2-70B w/o NatLog	57.61	60.33
GPT-3.5-Turbo w/o NatLog	54.63	59.33
Zero-NatVer	57.47	60.33
Zero-NatVer-R	**61.07**	**64.00**

Table 4: Comparison of our method with other non-NatLog systems on SciFact.

	Dan-FEVER		CHEF	
	Macro-F1	Micro-F1	Macro-F1	Micro-F1
Transfer				
ProoFVer (mBART)	36.12	55.22	20.18	37.72
QA-NatVer (mT0)	**63.64**	**68.41**	-	-
Generalization				
ProoFVer (mBART)	29.80	41.97	20.16	38.57
QA-NatVer (mT0)	35.68	37.05	-	-
Zero-NatVer	43.28	57.47	51.10	**58.75**
Zero-NatVer-R	44.93	57.53	**51.34**	58.46

Table 5: F1 metrics for a Danish (Dan-FEVER) and Chinese (CHEF) fact verification dataset.

our method by omitting prompt ensembles and using a single randomly sampled prompt instead. We observe a substantial drop in performance of 7.91 and 9.56 Macro-F1 points, which agrees with our previous findings regarding ensemble sizes. Third, we ablate Zero-NatVer by using unconstrained generation in decoding, observing performance drop of 2.02 and 1.33 Macro-F1 points.

Non-NatLog Systems Although natural logic provides improved explainability, it is also important to assess its impact on performance compared to similar models that are not grounded in natural logic. To this end, we conducted an experiment with Llama2 and ChatGPT-3.5 models, prompting them to determine the verdict directly (see Appendix D for details). Our experimental results are reported in Figure 4. While Zero-NatVer's performance is on par with other systems that do not use natural logic, Zero-NatVer-R substantially outperforms those systems by 3.46 Macro-F1 points, demonstrating the importance of evidence rephrasing. Notably, it also outperforms ChatGPT-3.5

Multilingual Capabilities For a more comprehensive evaluation, we also assess the multi-lingual capabilities of Zero-NatVer on two non-English fact-checking datasets– DanFEVER (Danish) and CHEF (Chinese). We do not make any adjustments to our approach except for the chunking strategy with CHEF, where we consider new chunks after every token instead of every space character. To evaluate our baselines, we use models based on multi-lingual backbones. Thus, we use mBART (Liu et al., 2020) for ProoFVer, and we use mT0 (Muennighoff et al., 2022) for QA-NatVer.

Our results are reported in Table 5. When compared to ProoFVer in both setups, Zero-NatVer outperforms the best baseline model by 19.0 Macro-F1 and 11.6 Micro-F1 points. Although ProoFver was trained on a large NLI-annotated dataset, our results show that it is not able to generalize to other languages as well as Zero-NatVer. Conversely, QA-NatVer outperforms our system by 16.5 Macro-F1 and 10.3 Micro-F1 points on Dan-FEVER. However, it is important to note that QA-NatVer uses a multi-lingual backbone model mT0, while the proportion of Chinese and Danish in Llama2 pretraining data was only 0.13% and 0.02% Danish, respectively (Touvron et al., 2023). Additionally, we were not able to obtain results for QA-NatVer on CHEF because QA-NatVer's chunking system currently does not support Chinese. This highlights the simplicity of our method, which uses a single multi-lingual model for most stages of its pipeline. Lastly, our results show no benefit of evidence rephrasing on non-English datasets, suggesting future research in this area.

6 Conclusion

We have presented Zero-NatVer, a zero-shot method for fact verification based on natural logic. Our method leverages the generalization capabilities of instruction-tuned LLMs and generates faithful justifications for proofs without relying on training data annotated with natural logic. In our study, we have compared Zero-NatVer with other fact verification (FV) systems grounded in natural logic, evaluating their performance in two zero-shot scenarios: generalization and transfer. Our method surpasses all baselines in both setups. Moreover, we explored the impact of evidence rephrasing, which further improves our model's performance across all datasets. We hope that the methods and analyses presented here enable further progress toward improving the efficiency and explainability of fact verification systems.

Limitations

Evidence Rephrasing While rephrasing improved our results across all datasets, it represents a trade-off between performance and explainability. Despite the use of constrained beam-search decoding, it can still generate sentences that are not logically consistent with the original evidence text, leading to an incorrect verdict. Therefore, users should have access to both texts in order to make their own judgments about the reliability of rephrasing.

Natural Logic Natural logic is useful for explainability but is less expressive than semantic parsing methods such as lambda calculus (Zettlemoyer and Collins, 2005). This paper doesn't address natural logic's limitations. Furthermore, our method generates proofs, which are meant to be processed by the DFA from left to right. Nevertheless, natural logic-based inference is not constrained to such execution.

Ethics Statement

Intended Use and Misuse Potential. Our models can potentially captivate a wider audience and significantly reduce the workload for human fact-checkers. Nevertheless, it is crucial to acknowledge the possibility of their exploitation by malicious actors. As such, we strongly advise researchers to approach them with caution.

Accuracy and Infallibility. Our approach improves the clarity of FV models, enabling individuals to make better-informed decisions about trusting these models and their assessments. However, it is crucial for users to remain critical while interpreting the results of these systems and not mistake explainability for accuracy. We clarify that our evaluations do not determine the factual accuracy of a statement in the real world; instead, we use sources like Wikipedia as the basis for evidence. Wikipedia is a great collaborative resource, yet it has mistakes and noise of its own, similar to any encyclopedia or knowledge source. Therefore, we advise against using our verification system to make definitive judgments about the veracity of the assessed claims, meaning it should not be relied upon as an infallible source of truth.

References

Rami Aly, Marek Strong, and Andreas Vlachos. 2023. Qa-natver: Question answering for natural logic-based fact verification. *arXiv preprint arXiv:2310.14198*.

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Guided open vocabulary image captioning with constrained beam search. *arXiv preprint arXiv:1612.00576*.

Gabor Angeli and Christopher D Manning. 2014. Naturalli: Natural logic inference for common sense reasoning. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 534–545.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020. Generating fact checking explanations. *arXiv preprint arXiv:2004.05773*.

Greg Brockman, Peter Welinder, Mira Murati, and OpenAI. 2020. Openai: Openai api. https://openai.com/blog/openai-api.

Jiangjie Chen, Qiaoben Bao, Changzhi Sun, Xinbo Zhang, Jiaze Chen, Hao Zhou, Yanghua Xiao, and Lei Li. 2022. Loren: Logic-regularized reasoning for interpretable fact verification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):10482–10491.

Jiangjie Chen, Rui Xu, Wenxuan Zeng, Changzhi Sun, Lei Li, and Yanghua Xiao. 2023. Converge to the truth: Factual error correction via iterative constrained editing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 12616–12625.

Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. 2020. Autoregressive entity retrieval. *arXiv preprint arXiv:2010.00904*.

Thomas Diggelmann, Jordan Boyd-Graber, Jannis Bulian, Massimiliano Ciaramita, and Markus Leippold. 2020. Climate-fever: A dataset for verification of real-world climate claims. *arXiv preprint arXiv:2012.00614*.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. 2022. Gptq: Accurate post-training quantization for generative pre-trained transformers. *arXiv preprint arXiv:2210.17323*.

Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2013. Ppdb: The paraphrase database. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 758–764.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*.

Xuming Hu, Zhijiang Guo, Guanyu Wu, Aiwei Liu, Lijie Wen, and Philip S Yu. 2022. Chef: A pilot chinese dataset for evidence-based fact-checking. *arXiv preprint arXiv:2206.11863*.

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness? *arXiv preprint arXiv:2004.03685*.

Mingjian Jiang, Yangjun Ruan, Sicong Huang, Saifei Liao, Silviu Pitis, Roger Baker Grosse, and Jimmy Ba. 2023. Calibrating language models via augmented prompt ensembles.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Neema Kotonya and Francesca Toni. 2020. Explainable automated fact-checking for public health claims. *arXiv preprint arXiv:2010.09926*.

Amrith Krishna, Sebastian Riedel, and Andreas Vlachos. 2022. Proofver: Natural logic theorem proving for fact verification. *Transactions of the Association for Computational Linguistics*, 10:1013–1030.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Xingxuan Li, Ruochen Zhao, Yew Ken Chia, Bosheng Ding, Lidong Bing, Shafiq Joty, and Soujanya Poria. 2023. Chain of knowledge: A framework for grounding large language models with structured knowledge bases. *arXiv preprint arXiv:2305.13269*.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pretraining for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Bill MacCartney and Christopher D Manning. 2008. Modeling semantic containment and exclusion in natural language inference. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 521–528.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, et al. 2022. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*.

Jeppe Nørregaard and Leon Derczynski. 2021. Danfever: claim verification dataset for danish. In *Proceedings of the 23rd Nordic conference on computational linguistics (NoDaLiDa)*, pages 422–428.

R OpenAI. 2023. Gpt-4 technical report. arxiv 2303.08774. *View in Article*, 2.

Liangming Pan, Wenhu Chen, Wenhan Xiong, Min-Yen Kan, and William Yang Wang. 2021. Zero-shot fact verification by claim generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 476–483, Online. Association for Computational Linguistics.

Liangming Pan, Xiaobao Wu, Xinyuan Lu, Anh Tuan Luu, William Yang Wang, Min-Yen Kan, and Preslav Nakov. 2023a. Fact-checking complex claims with program-guided reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6981–7004, Toronto, Canada. Association for Computational Linguistics.

Liangming Pan, Yunxiang Zhang, and Min-Yen Kan. 2023b. Investigating zero-and few-shot generalization in fact verification. *arXiv preprint arXiv:2309.09444*.

Ethan Perez, Douwe Kiela, and Kyunghyun Cho. 2021. True few-shot learning with language models. *Advances in neural information processing systems*, 34:11054–11070.

Kashyap Popat, Subhabrata Mukherjee, Andrew Yates, and Gerhard Weikum. 2018. Declare: Debunking fake news and false claims using evidence-aware deep learning. *arXiv preprint arXiv:1809.06416*.

Stephen E Robertson and Steve Walker. 1994. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *SIGIR'94: Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, organised by Dublin City University*, pages 232–241. Springer.

Victor Sanchez. 1991. *Studies on natural logic and categorial grammar*. Ph.D. thesis, University of Amsterdam.

Tal Schuster, Darsh J Shah, Yun Jie Serene Yeo, Daniel Filizzola, Enrico Santus, and Regina Barzilay. 2019. Towards debiasing fact verification models. *arXiv preprint arXiv:1908.05267*.

Kai Shu, Limeng Cui, Suhang Wang, Dongwon Lee, and Huan Liu. 2019. defend: Explainable fake news detection. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 395–405.

Joe Stacey, Pasquale Minervini, Haim Dubossarsky, Oana-Maria Camburu, and Marek Rei. 2023. Logical reasoning for natural language inference using generated facts as atoms. *arXiv preprint arXiv:2305.13214*.

Joe Stacey, Pasquale Minervini, Haim Dubossarsky, and Marek Rei. 2022. Logical reasoning with span-level predictions for interpretable and robust NLI models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3809–3823, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Joseph E Uscinski and Ryden W Butler. 2013. The epistemology of fact checking. *Critical Review*, 25(2):162–180.

Johan Van Benthem. 1986. *Natural Logic*, pages 109–119. Springer Netherlands, Dordrecht.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. *arXiv preprint arXiv:2004.14974*.

Dustin Wright, David Wadden, Kyle Lo, Bailey Kuehl, Arman Cohan, Isabelle Augenstein, and Lucy Lu Wang. 2022. Generating scientific claims for zero-shot scientific fact checking. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2448–2460, Dublin, Ireland. Association for Computational Linguistics.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations*.

Luke S. Zettlemoyer and Michael Collins. 2005. Learning to map sentences to logical form: Structured classification with probabilistic categorial grammars. In *Proceedings of the Twenty-First Conference on Uncertainty in Artificial Intelligence*, UAI'05, page 658–666, Arlington, Virginia, USA. AUAI Press.

A Dataset Processing

To effectively assess the zero-shot capabilities of FV systems, it is important to evaluate the performance on real-life claims and consider domains requiring various domain expertise. We evaluated all models on datasets covering natural claims and domains such as climate change, biomedical subjects, government healthcare policies, and scientific literature. We chose datasets that mainly focus on three-way classification, i.e., using three labels *Suppports*, *Refutes*, or *Not Enough Information*:

Climate-FEVER (Diggelmann et al., 2020) dataset comprises 1535 real-life climate change claims, each annotated with five evidence sentences retrieved from Wikipedia. Each evidence sentence was labeled by five human annotators as supporting, refuting, or inconclusive regarding the claim's veracity, resulting in 5 votes for each evidence sentence. These votes were then aggregated to micro-verdicts for each retrieved evidence sentence, and micro-verdicts were further aggregated to a single macro-label for the claim. In our data processing, we combined all evidence sentences into a single paragraph and paired them with the macro-label assessment. Besides the standard three labels, some claims in the datasets are labeled as *DISPUTED* if they are paired with both supporting and refuting micro-verdicts. Since our work focuses on three-label class prediction, we removed those 154 claims from the dataset.

PubHealth (Kotonya and Toni, 2020) is a dataset with natural claims in the public health domain. These claims are accompanied by evidence that requires subject matter expertise, along with expert explanations (judgments). For our experiments, we use the claim from the development set (1032 in total) paired with expert explanations as evidence.

SciFact (Wadden et al., 2020) is a dataset of expert-written scientific claims paired with evidence that was extracted from academic papers. We collect the claims with supporting and refuting rationale and construct claim-evidence pairs with *SUPPORT* and *REFUTE* labels. Claims lacking a specific rationale are categorized as *NEI*, and we pair them with the entire abstract text. We evaluate our pipeline on a test set that consists of 300 claims.

DanFEVER (Nørregaard and Derczynski, 2021) is a Danish dataset of counterfactual claims constructed from Danish Wikipedia. It consists of 6407 instances and provides gold evidence for *Supported* and *Refuted* claims. To obtain evidence for *NEI* claims, we use the BM25 retriever (Robertson and Walker, 1994).

```
You are given two sentences – Original Sentence and Syntax Reference Sentence.
Your task is to rephrase the first sentence (Original Sentence) so that it becomes syntactically closer to the structure of the second sentence (Syntax Reference Sentence), while ensuring that all the information from the original sentence remains logically consistent.

Original Sentence: {E}

Syntax Reference Sentence: {C}

Rephrase only the parts relevant to Syntax Reference Sentence. Don't change the logical meaning of Original Sentence.
```

Listing 1: Prompt template for the rephrasing task. Placeholders *{E}* and *{C}* get replaced by corresponding texts.

```
You are given two texts – a claim and evidence. Your task is to split the given claim into smaller verifiable segments and align each segment with the corresponding relevant information from the evidence text.

Proceed in two steps.
Step 1: Divide the provided claim text into smaller, independently verifiable segments.
Step 2: For each segmented chunk of the claim, identify and align it with the corresponding relevant information in the evidence text.

Segment and align the following claim and evidence texts:

CLAIM: {C}
EVIDENCE: {E}
```

Listing 2: Prompt template for the chunking and alignment task. Placeholders *{E}* and *{C}* get replaced by corresponding texts.

CHEF (Hu et al., 2022) is a Chinese dataset of real-world claims. We use their development set, which consists of 703 claims.

B Baselines

ProoFVer (Krishna et al., 2022) is a seq2seq FV model that generates natural logic proofs as sequences of *(claim, evidence, NatOp)* triples. ProoFVer is based on GENRE (De Cao et al., 2020), an end-to-end entity linking model that was obtained by fine-tuning the BART language model (Lewis et al., 2019). ProoFVer was trained on a large collection of 145,449 claims from FEVER that were heuristically annotated with natural logic proofs.

QA-NatVer (Aly et al., 2023) is also based on natural logic but uses a question-answering framework to determine proofs. As a few-shot method, QA-NatVer was trained only on a small subset of FEVER data. It uses 68 training instances, which were further manually annotated with natural logic proofs.

Pan et al. Pan et al. (2023b) recently published an extensive analysis of zero-shot FV over 11 FV datasets. In their work, they experimented with different combinations of datasets for training and testing. While Pan et al. (2023b) consider their experiments as zero-shot generalization tasks, in our work, we consider them as zero-shot transfer because they train their models on other FV datasets. Their results show useful zero-shot baselines over most of our datasets, providing a comparison with FV models that are not based on natural logic.

C Models

Llama2 We ran 7B, 13B, and 70B parameter models locally and used the GPTQ (Frantar et al., 2022) version of these models with 4-bit quantization to lower the computational requirements and speed up the inference.

Hyperparameters When decoding with Llama-2 models, we did not tune any hyperparameters and used the values described in Touvron et al. (2023). Specifically, in the question-answering task for NatOPs, we set temperature

to 1.0 and use nucleus sampling (Holtzman et al., 2019) with top-p set to 0.9. For all other tasks, we change temperature to 0.1.

Experimental Setup All experiments using Llama2 as the instruction-finetuned LLM were run on a machine with a single Quadro RTX 8000 with 49GB memory and 64GB RAM memory.

D Prompting

Listings 1 and 2 show prompt templates for the evidence-rephrasing task, and the chunking and alignment task, respectively. These prompt templates were used for all experiments with Llama2 and ChatGPT models.

NatOp assignment Listing 3 shows the prompt templates used in the question-answering task for NatOps. Given a claim-evidence pair, we generated 10 distinct questions for each NatOp in separate prompts, replacing *X* with the claim text and *Y* with the evidence text. Additionally, we added the phrase *"Answer Yes or No."* at the end of each prompt to encourage the *Yes/No* output format. Lastly, we used the default system prompt *"You are a helpful assistant."* for all prompts.

ChatGPT We used OpenAI's API (Brockman et al., 2020) to query *gpt-3.5-turbo-1106* and used the same hyperparamteres as with Llama2 models. Due to the API limitations, we were unable to use constrained decoding for rephrasing, chunking, and alignment. Moreover, we could not use weighted prompt ensembles due to the inability to access the model's log-likelihood scores. Otherwise, we could replicate all the steps of our method with ChatGPT.

Equivalence
Is X a paraphrase of Y?
Are X and Y semantically equivalent in meaning?
Is the meaning of X effectively the same as Y?
Do X and Y function as synonyms or paraphrases of each other?
Does X serve as a paraphrase or an alternative expression for Y?
Are X and Y synonymous or nearly synonymous in meaning?
Do X and Y mean the same, without using external knowledge or assumptions?
Are X and Y semantically identical when considered independently of external knowledge?
Considering just X and Y, do these expressions have the same meaning?
Comparing X with Y, are they semantically equivalent based solely on their respective content?

Entailment
Given the premise Y does the hypothesis X hold?
Does the expression Y entail X?
Does the phrase Y logically imply X?
Is it true that if Y then X?
Is X a valid inference from Y?
Can X be inferred from the statement Y?
Given just the statements Y and X, does the first statement logically and necessarily imply the second without any external information?
Is it true that the statement Y logically entails X based solely on the information within the statements?
Does Y imply X when only the information within these statements is considered?
Is it accurate to say that Y categorically entails X, without external interpretations?

Negation
Is the phrase X a negation of Y?
Do X and Y represent mutually exclusive states, where the presence of one negates the possibility of the other?
Is the relationship between X and Y binary, such that if X is true, Y must necessarily be false, and vice versa?
Do X and Y negate each other completely?
Are X and Y in a dichotomous relationship, where the existence of one implies the non−existence of the other?
Is there a mutually exclusive relationship between X and Y, indicating that only one can be true at any given time?
In the context of X and Y, does the affirmation of one mean the automatic negation of the other?
Do X and Y form a binary opposition, where one categorically negates the other?
Are X and Y opposites in such a way that they cannot be true simultaneously?
Is the relationship between X and Y characterized by a strict either/or dichotomy?

Alternation
Does X exclude Y?
Do X and Y represent distinct alternatives, but not the only possibilities in their category?
Are X and Y exclusively different without negating the existence of additional states or options?
Do X and Y denote exclusive but not exhaustive options within a larger set of possibilities?
In comparing X and Y, are they distinct yet not limiting the possibility of other variations or alternatives?
Are X and Y distinct entities or states that exclude each other without forming a complete, exhaustive set?
Are X and Y different entities or states, but not in a way that negates the possibility of other, different entities or states?
Are X and Y distinct entities or states that exclude each other without forming a complete, exhaustive set?
In comparing X and Y, are they exclusive in nature but not necessarily covering all possible alternatives?
Do X and Y define separate, non−intersecting options, while not encompassing all possible scenarios?

Listing 3: Template questions for determining NatOps.

Is the claim {C} supported or refuted by the evidence {E}?
Alternatively, reply that there is insufficient evidence to support or refute the claim.

Choices:
(A): Supported
(B): Refuted
(C): Not Enough Information

Answer in the following format:
Answer=A|B|C

Listing 4: Prompt template for FV experiments in a direct multiple-choice setup. Placeholders *{E}* and *{C}* get replaced by corresponding texts.