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

Anonymous ACL submission

Abstract

The recent development of fact verification 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 improve stability. We evaluate our system on artificial and real-world fact verification data. In a zero-shot setup where models were not trained on any data annotated with natural logic, our method surpasses the best baselines by an average of 7.52 accuracy points. We also demonstrate multilingual capabilities in other languages, such as Danish, where we outperform our baselines by 8.72 accuracy points.

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. While many improvements have been made in this field due to the recent rapid growth in NLP (Mubashara et al., 2023; Guo et al., 2022; Nakov et al., 2021), FV systems often employ pipelines with black-box components that hide the underlying reasoning.

One line of research 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). However, 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) provide faithful justifications by expressing semantic relations between claim/evidence pairs. Modeling these logical relations and their aggregation explicitly with natural logic (such as double-negation) has also resulted in more accurate and robust fact-checking systems.

However, a severe limitation of natural logic-based FV systems is the necessity for large amounts of training data annotated with entire natural logic proofs. For example, ProoFVer (Krishna et al., 2022) was trained on 145K instances artificially obtained from structured knowledge bases such as PPDB (Ganitkevitch et al., 2013) and Wikidata (Vrandečić and Krötzsch, 2014). While recent work (Aly et al., 2023) attempts to alleviate this issue by proposing a few-shot learning method trained on as few as 32 instances, human annotation of even a small number of proofs can be impractical and expensive, as it requires substantial linguistic knowledge and familiarity with natural logic. Moreover, few-shot systems might require additional training data in order to generalize effectively to new domains, further increasing the costs.

To this end, we propose **Zero-NatVer**, a zero-shot fact verification approach for constructing natural logic proofs that leverages prompting and question-answering with instruction-tuned large language models (LLMs). Unlike some previous works that combine several fine-tuned models, our method uses a single language model for all stages of the pipeline and does not require any adjustments when transferring to different domains or languages. Furthermore, we investigate evidence rephrasing to address the lack of clear alignment between claim and evidence, a common problem of fact verification with natural logic. For example, as illustrated in Figure 1, evidence rephrasing can improve alignments by reformulating text into a more detailed form.

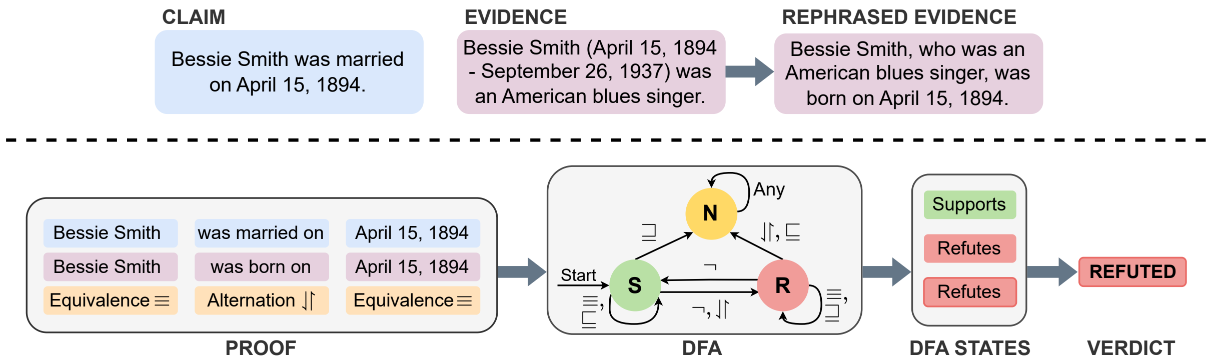

Figure 1: **Proof generation with natural logic using rephrased evidence.** Evidence is first rephrased to facilitate alignment with the claim (e.g., it introduces the word ’born’). Claim and evidence texts are then processed to generate a proof sequence, which consists of *(claim, evidence, NatOp)* triples. Lastly, NatOps are used as transitions in the DFA, and the final state (i.e. Refutes) determines the verdict.

We evaluate our method on real-world and artificial FV datasets, including Climate-FEVER ((Diggelmann et al., 2020)), PubHealth ((Kotonya and Toni, 2020)), SciFact ((Wadden et al., 2020)), and Hover ((Jiang et al., 2020)). In a zero-shot setup, where models have not been trained using any data labeled with natural logic, our approach outperforms the top baseline models by an average accuracy improvement of 10.09 points. By rephrasing relevant evidence, we further improve these results by additional 1.96 points. Our method also surpasses fully supervised and few-shot trained models on natural datasets, obtaining the average improvement of 2.22 accuracy points without evidence rephrasing and 5.29 points with evidence rephrasing. Lastly, we evaluate our system on multilingual datasets, including Danish (DanFever (Nørregaard and Derczynski, 2021)) and Chinese (CHEF (Hu et al., 2022)), demonstrating that it can generalize to other languages.

Our study also evaluates the performance benefits of using natural logic in FV. To this end, we conduct experiments, comparing Zero-NatVer with LLMs of similar sizes that are prompted directly for determining the final verdict. Zero-NatVer outperforms these methods by 3.67 accuracy points.

2 Related Work

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). Other neuro-symbolic reasoning systems for FV use 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).

Previous work on zero-shot FV is limited and largely relies 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 and 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 Zero-NatVer

Given a claim c and 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 reach a verdict. Zero-NatVer obtains the verdict in four steps, executed by an instruction-tuned LLM. 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, Zero-NatVer first rephrases evidence E into R so that relevant information is easier to align with the claim c while staying semantically equivalent to E (Sec. 3.1). In the second step, Zero-NatVer segments c into several chunks and aligns each such chunk with relevant information from R (Sec. 3.2). This process results in a sequence of l claim-evidence alignment pairs $A = a_1, a_2, ..., a_l$. Next, Zero-NatVer determines the relation of each pair in terms of natural logic and generates a sequence of natural logic operators $O = o_1, o_2, ..., o_l$, which correspond to alignment pairs in A (Sec. 3.3). Finally, O is used to traverse a deterministic finite state automaton (DFA), which determines the claim's veracity. 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)" (Figure 1) 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.

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 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 alter the meaning of the evidence. Even though we instruct the LLM accordingly, it 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, we keep track of several most likely partial sequences and constrain 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, in Figure 1, 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 (Krishna et al., 2022). This preprocessing traditionally consists of two separate 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 predicting relevant NatOps.

Zero-NatVer performs both steps as a joint task, using the same prompt (details in Listing 2) and context window. As shown in Figure 2, the decoding starts with generating claim chunks as follows:

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) Segment the claim text into chunks:
* Miracle at St. Anna
* tells the story of
* four soldiers

Step 2) Align each claim chunk with relevant evidence:
* "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: **Decoding for chunking and alignment.** The blue text refers to generated claim chunks, and the purple text refers to generated evidence alignments. The remaining text was forced during the decoding.

1. The claim text is pre-processed as a queue of tokens Q_C.
2. The decoding is prefixed with the phrase *"Step 1) Segment the claim text into chunks:"* to encourage the generation of claim chunks.
3. The model is constrained to sample one of two outputs - the next token from Q_C or a new-line character.
4. Repeats step 3 until Q_C is empty (all claim tokens are consumed).

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, Zero-NatVer then starts generating alignments:

1. The previously generated chunks are parsed and stored in queue Q_A.
2. The decoding is prefixed with the phrase *"Step 2) Align each claim chunk with relevant evidence:"* to encourage alignment generation.
3. The model is prefixed with a chunk from Q_A.
4. Aligned evidence text is sampled with constrained decoding.
5. Repeats steps 3-4 until Q_A is empty.

As shown in Figure 2, the outcome of this process is a bulleted list of claim-evidence segments. While the decoding of claim chunks is constrained by design and does not allow for hallucinations, the alignment generation in step 4 relies on general sampling and needs to be constrained. In order to prevent hallucinations and guarantee reliability, we post-process the alignments and remove any text that does not form sequences of tokens in E or R. This approach ensures the aligned text comprises only sub-strings from E or R.

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 in a zero-shot setting.

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 inferred 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 ($\downharpoonleft\upharpoonright$). 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.

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-judgements per NatOp, which are then aggregated as a weighted average. Instead of manually hand-crafting these question templates, we prompt the LLM 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 alignment pair a and operator o, we compute a NatOp score $s_{o,a}$ as a weighted average over all micro-judgments:

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

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_i 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).

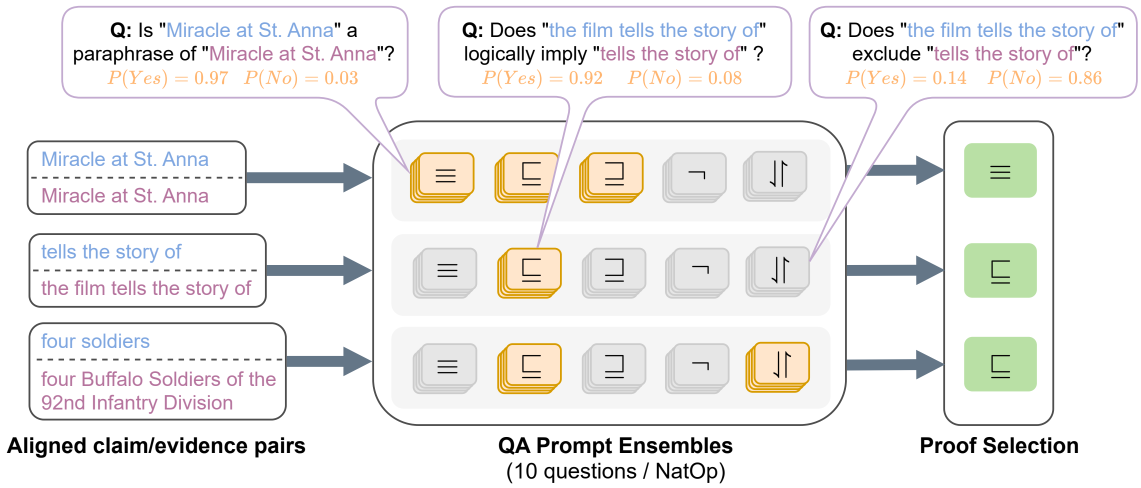

Figure 3: **Proof generation process of Zero-NatVer.** We use prompt ensembles to determine a set of NatOp candidates (orange blocks) for each claim-evidence pair. A single NatOp (green blocks) is then selected for each pair, using NatOp priority.

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.

Due to the ambiguity of natural language and the complexity of alignments, it frequently occurs that $|C| > 1$. However, we want to minimize the chance of incorrectly choosing NatOps that leads to the *Not Enough Evidence* state, from which there are no outgoing transitions to other states. Thus, we use a NatOp priority approach and select from the operators in C in the following order: $[\equiv, \neg, \sqsubseteq, \sqsupseteq, \downharpoonleft\upharpoonright]$.We defined the NatOp order by considering the difficulty of each task. For instance, in a scenario where the candidate list C consists of equivalence ($\equiv$) and alternation ($\downharpoonleft\upharpoonright$), 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 before our experiments and did not optimize this order.

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 (e.g., BART, T5, or Llama2) that did not include proofs with natural logic. Applying this model to FV with natural logic then exemplifies zero-shot generalization according to our definition. In contrast, if the same model is fine-tuned on a dataset annotated with natural logic proofs and then applied to perform FV with natural logic 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 mostly general topics, and artificial claims are often rather simple in their structure. For a more comprehensive assessment of zero-shot capabilities, we also 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 more details).

5 Results

To effectively assess the impact of evidence rephrasing, we consistently report our results in two separate formats: without evidence rephrasing (denoted as **Zero-NatVer**) and with rephrasing

	Model	Climate-FEVER		PubHealth		SciFact		Hover	
		F1	Acc	F1	Acc	F1	Acc	F1	Acc
ProoFVer	BART	26.63	34.75	38.15	39.27	25.58	34.67	47.13	49.76
QA-NatVer	Flan-T5	22.20	36.86	44.42	48.73	23.56	40.67	35.65	50.85
QA-NatVer	Llama2-70B	36.13	47.28	57.05	63.12	37.78	46.67	55.45	55.47
Zero-NatVer	Llama2-70B	44.71	46.78	65.45	65.45	57.47	60.33	59.12	59.13
Zero-NatVer-R	Llama2-70B	**45.78**	**49.38**	**66.91**	**68.39**	**61.07**	**64.00**	**60.83**	**60.85**
Full Supervision	-	75.7	-	85.88	86.93	71.1	-	-	81.2

Table 1: **Zero-shot generalization results.** Macro-F1 and accuracy scores for systems that were **not** specifically trained on FV datasets. Where possible, we also report available SOTA results with fully-supervised models trained on in-domain data as a reference.

	Model	Training data (size)	Climate-FEVER		PubHealth		SciFact		Hover	
			F1	Acc	F1	Acc	F1	Acc	F1	Acc
Pan et al. (2013)	BERT	FEVER/VitC (800)	40.60	-	60.06	-	50.71	-	-	-
ProoFVer	BART	FEVER (145K)	40.70	43.35	57.78	61.22	45.57	49.16	57.08	57.89
QA-NatVer	Flan-T5	FEVER (64)	44.74	47.43	61.8	61.8	52.02	56.67	**70.27**	**70.5**
Zero-NatVer	Llama2-70B	None	44.71	46.78	65.45	65.45	57.47	60.33	59.12	59.13
Zero-NatVer-R	Llama2-70B	None	**45.78**	**49.38**	**66.91**	**68.39**	**61.07**	**64.00**	60.83	60.85
Full Supervision	-	-	75.7	-	85.88	86.93	71.1	-	-	81.2

Table 2: **Zero-shot transfer results.** Macro-F1 and accuracy scores for systems trained on fact-checking datasets. For each system, we report the type and size of FV training data. Where possible, we also report available SOTA results with fully-supervised models trained on in-domain data as a reference. Results from Pan et al. (2013) do not include accuracy scores and experiments on Hover.

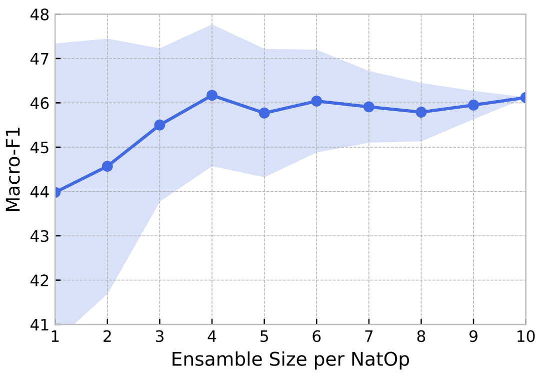

Figure 4: Macro-F1 scores across all datasets for various ensemble sizes. The light blue area represents the standard deviation from 10 independent measurements.

(denoted as **Zero-NatVer-R**)."

We conducted our main experiments with the Llama2-70B model (Touvron et al., 2023), 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. For this purpose, we used a separate dataset, Symmetric-Fever (Schuster et al., 2019). We selected a small subset of 100 claims and tested that our prompts generated 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 (details in Appendix C).

Baselines Our natural-logic-based baselines consist of ProoFVer (Krishna et al., 2022), which is based on the BART model (Lewis et al., 2019), and QA-NatVer (Aly et al., 2023), which uses Flan-T5 (Chung et al., 2022). In zero-shot generalization setups, we run both models with their corresponding pre-trained LLMs without fine-tuning on NLI data. In order to provide a directly comparable baseline, we also implemented support for Llama2 in QA-NatVer. ProoFVer currently supports only models from the Fairseq toolkit[1], which does not include models of similar sizes to Llama2. For zero-shot transfer setups, we use ProoFVer with BART trained on 145K FEVER instances and QA-NatVer with Flan-T5 trained on 64 instances of NLI annotated data. We were unable to fine-tune QA-NatVer with Llama2-70B model due to computational constraints. We 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 for zero-shot generalization in Table 1. Zero-NatVer

[1]https://github.com/facebookresearch/fairseq

	Macro-F1	Accuracy
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	Accuracy
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
separate chunking/alignment	52.3	54.33

Table 4: Ablation study on SciFact.

	Macro-F1	Accuracy
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 5: Comparison of our method with other non-NatLog systems on SciFact.

achieves 57.92 accuracy points on average, surpassing all baselines. Our system outperforms ProoFVer and QA-NatVer with Flan-T5 backbone by 18.31 and 13.65 accuracy points, respectively. Notably, it also outperforms QA-NatVer with Llama2-70B backbone by 4.79 accuracy points. Evidence rephrasing (Zero-NatVer-R) further improves our results by additional 2.73 accuracy points. The main results for zero-shot transfer are reported in Table 2. When considering only datasets that contain natural claims, Zero-NatVer outperforms ProoFVer and QA-NatVer with Flan-T5 backbone by 6.28 and 2.22 accuracy points on average, respectively. Zero-NatVer-R further improves results by additional 3.07 accuracy points. However, QA-NatVer outperforms Zero-NatVer-R by 9.65 accuracy points on the artificial claims from Hover. While QA-NatVer's results demonstrate generalization capabilities beyond the training domain, the high scores can be attributed to the fact that both FEVER (i.e., QA-NatVer's training data) and Hover consist of artificial claims compiled from Wikipedia.

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. 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 Table 3 compares our method with different sizes of Llama2 models, showing a substantial improvement in performance as the model scales up. Additionally, we evaluated our method using the proprietary model ChatGPT-3.5 (OpenAI, 2023), 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 four ablation studies on SciFact, as reported in Table 4. First, we examine the performance of Zero-NatVer and Zero-NatVer-R without weighting ensemble prompts, observing a small drop of 1 accuracy point. Second, we ablate our method by omitting prompt ensembles and using a single randomly sampled prompt instead. We observe a substantial drop in performance of 6.66 accuracy points, which agrees with our previous findings regarding ensemble sizes. Third, we ablate Zero-NatVer by using unconstrained generation in decoding, observing an accuracy drop of 2.33 points. Last, we ablate our method by processing chunking and alignments as two separate consecutive steps, resulting in 6.0 points drop in accuracy.

Non-NatLog Systems We also compared our method with similar models that are not grounded in natural logic and conducted experiments with Llama2 and ChatGPT-3.5 models, prompting them to determine the verdict directly (see Appendix D for prompting details). Our experimental results reported in Table 5 demonstrate that Zero-NatVer-R substantially outperforms Llama2-70B and ChatGPT-3.5 by 3.67 and 4.67 accuracy points, respectively. These results demonstrate that natural logic provides improved performance in addition to the benefits of explainability.

	Model	Dan-FEVER		CHEF	
		Macro-F1	Acc	Macro-F1	Acc
ProoFVer	mBART	29.80	41.97	20.16	38.57
QA-NatVer	mT0	35.68	37.05	-	-
QA-NatVer	Llama2-70B	34.17	48.81	-	-
Zero-NatVer	Llama2-70B	43.28	57.47	51.10	**58.75**
Zero-NatVer-R	Llama2-70B	**44.93**	**57.53**	**51.34**	58.46
Full-Supervision	-	90.2	-	67.62	-

Table 6: **Zero-shot generalization results for multi-lingual datasets.** Macro-F1 and accuracy scores for systems that were **not** specifically trained on FV datasets. Where possible, we also report available SOTA results with fully-supervised models trained on in-domain data as a reference.

	Model	Training data (size)	Dan-FEVER		CHEF	
			Macro-F1	Acc	Macro-F1	Acc
ProoFVer	mBART	FEVER (145K)	36.12	55.22	20.18	37.72
QA-NatVer	mT0	FEVER (64)	**63.64**	**68.41**	-	-
Zero-NatVer	Llama2-70B	None	43.28	57.47	51.10	**58.75**
Zero-NatVer-R	Llama2-70B	None	44.93	57.53	**51.34**	58.46
Full-Supervision	-	-	90.2	-	67.62	-

Table 7: **Zero-shot transfer results for multi-lingual datasets.** Macro-F1 and accuracy scores for systems trained on fact-checking datasets. Where possible, we also report available SOTA results with fully-supervised models trained on in-domain data as a reference.

Multilingual Capabilities We also assess the multi-lingual capabilities of Zero-NatVer on two fact-checking datasets in languages other than English– DanFEVER (Danish) and CHEF (Chinese). 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) and Llama-70B for QA-NatVer. Table 6 reports our results for zero-shot generalization. On DanFEVER, Zero-NatVer-R outperforms both ProoFVer and QA-NatVer by 15.56 and 8.72 accuracy points, respectively. This gap is substantially larger on CHEF, where the difference is 21.03 points. We could not run QA-NatVer on CHEF because QA-NatVer relies on an additional model for chunking that currently does not support Chinese. This limitation highlights the simplicity of our method, which uses a single multi-lingual model for all stages of the pipeline and does not require any adjustments when transferring to different domains or languages. Table 7 then reports results for zero-shot transfer, comparing Zero-NatVer with two multilingual baseline models trained on data with natural logic. While our system's accuracy is worse than QA-NatVer by 10.88 points, it is important to note that QA-NatVer uses a multi-lingual backbone model mT0 with a balanced distribution of languages. In comparison, the proportion of Chinese and Danish in Llama2 pre-training data was only 0.13% and 0.02% Danish, respectively (Touvron et al., 2023). ProoFVer was unable to generalize to CHEF in this setup, and Zero-NatVer outperforms ProoFVer by 21.03 accuracy points.

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. We have evaluated Zero-NatVer in two zero-shot setups, outperforming our baselines on most datasets. The ablation study shows the importance of individual design choices, and our experiments with non-NatLog systems demonstrate that natural logic improves the performance of our system. Moreover, we explored the impact of evidence rephrasing, which further improves Zero-NatVer'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.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

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.

Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10:178–206.

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.

Yichen Jiang, Shikha Bordia, Zheng Zhong, Charles Dognin, Maneesh Singh, and Mohit Bansal. 2020. Hover: A dataset for many-hop fact extraction and claim verification. *arXiv preprint arXiv:2011.03088*.

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*.

Bill Yuchen Lin, Kangmin Tan, Chris Miller, Beiwen Tian, and Xiang Ren. 2022. Unsupervised cross-task generalization via retrieval augmentation. *Advances in Neural Information Processing Systems*, 35:22003–22017.

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.

Akhtar Mubashara, Schlichtkrull Michael, Guo Zhijiang, Cocarascu Oana, Simperl Elena, and Vlachos Andreas. 2023. Multimodal automated fact-checking: A survey. *arXiv preprint arXiv:2305.13507*.

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*.

Preslav Nakov, David Corney, Maram Hasanain, Firoj Alam, Tamer Elsayed, Alberto Barrón-Cedeño, Paolo Papotti, Shaden Shaar, and Giovanni Da San Martino. 2021. Automated fact-checking for assisting human fact-checkers. *arXiv preprint arXiv:2103.07769*.

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*.

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). The dataset contains four labels *True*, *False*, *Unproven*, and *Mixture*. However, the classes are heavily unbalanced and the labels *Unproven* and *Mixture* cover less than 10% of the data in total. Therefore, we use test set

```
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.

claims with only *True* and *False* labels, resulting in 987 claims 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.

Hover (Jiang et al., 2020) is an open-domain, multi-hop FV dataset, containing artificial claims built from the Wikipedia corpus. Its claims are labeled as either *SUPPORTED* and *NOT-SUPPORTED*. We use the development set, which consists of 4000 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).

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 64 training instances, which were further manually annotated with natural logic proofs.

QA-NatVer currently supports BART0 (Lin et al., 2022), Flan-T5 (Chung et al., 2022) and mT0 (Muennighoff et al., 2022) backbones. However, we also implemented support for Llama2 in QA-NatVer, and reported results for zero-shot generalization with the Llama2-70B model.

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.