Towards the Roots of the Negation Problem: A Multilingual NLI Dataset and Model Scaling Analysis

Tereza Vrabcová[♣] **Marek Kadlčík**[♣] **Petr Sojka**[♣] **Michal Štefánik**[♡♣†] **Michal Spiegel**[♣♢]

[♣]Faculty of Informatics, Masaryk University, Czech Republic [♡]Language Technology, University of Helsinki

[†] National Institute of Informatics, Research and Development Center for Large Language Models, Tokyo

[♢]Kempelen Institute of Intelligent Technologies

Correspondence: negation@fi.muni.cz

Abstract

Negations are key to determining sentence meaning, making them essential for logical reasoning. Despite their importance, negations pose a substantial challenge for large language models (LLMs) and remain underexplored.

We constructed and published two new textual entailment datasets NoFEVER-ML and NoSNLI-ML in four languages (English, Czech, German, and Ukrainian) with *paired* examples differing in negation. It allows investigation of the root causes of the negation problem and its exemplification: how popular LLM model properties and language impact their inability to handle negation correctly.

Contrary to previous work, we show that increasing the model size may improve the models' ability to handle negations. Furthermore, we find that both the models' reasoning accuracy and robustness to negation are language-dependent and that the length and explicitness of the premise have an impact on robustness. We observe higher accuracy in languages with relatively fixed word order like English, compared to those with greater flexibility like Czech and German. Our entailment datasets pave the way to further research for explanation and exemplification of the negation problem, minimization of LLM hallucinations, and improvement of LLM reasoning in multilingual settings.

1 Introduction

Understanding negation is a cornerstone of sentiment analysis (Pang et al., 2002), logical reasoning and nuanced communication. Still, it often remains a 'pink elephant in the room' for Large Language Models (LLMs). Despite their impressive advancements, effectively understanding and processing negation poses a fundamental, and often underestimated, challenge to their reliability and logical reasoning capabilities. Processing negation is critical for reliable performance, and leads to hallucinations as Varshney et al. (2025) showed on four tasks with negation: 'false premise completion', 'constrained fact generation', 'multiple choice question answering', and 'fact generation'. Vision-language models do not understand negation either (Alhamoud et al., 2025): when asked to provide a picture with no elephant in the room, they put an elephant there anyway. It appears that negation tokens (e.g., "not") have a limited effect on the representations learned distributionally (Singh et al., 2024).

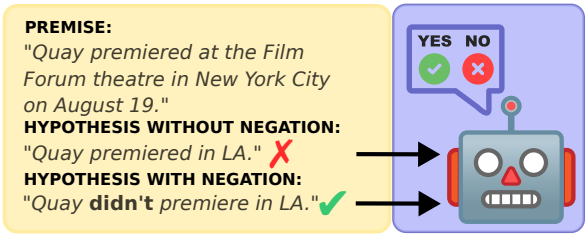

Figure 1: We extend existing entailment datasets FEVER and SNLI with negated hypotheses in four languages (ENG, UKR, CZE, GER) and measure how negations degrade the accuracy of model reasoning.

This difficulty, sometimes termed ***negation blindness*** or ***NO syndrome***, is well-documented. Studies show LLMs may struggle to distinguish between facts and their negations, misunderstand the semantic impact of negative particles, and fail to generalize negation handling robustly, even with instruction tuning.

Prior Work on Negation in LLMs

The challenge of negation for LLMs has been approached from several angles, as in the parable about the elephant met by blind monks (Ireland, 1997). We categorize prior work into the following areas:

1. Foundational Issues & Early Observations Early research highlighted fundamental difficulties for language models in capturing logical operations

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 25537–25551

such as negation within continuous vector space representations (Hermann et al., 2013). Subsequent work observed that LLMs can be equally prone to generating factual statements and their incorrect negations (Kassner and Schütze, 2020), and specific assessments were made on models like BERT regarding their sensitivity to the meaning conveyed by negation (Ettinger, 2020).

2. Datasets & Benchmarks for Negation Several benchmarks have been developed to probe negation understanding. For instance, the NeQA dataset was introduced to test question answering with negation, where models did not exhibit straightforward positive scaling (Zhang et al., 2023). Another large-scale benchmark was created to challenge LLMs with a wide array of negative sentences, finding that while fine-tuning helps, generalization in handling negation remains elusive (García-Ferrero et al., 2023). The NaN-NLI test suite specifically focuses on sub-clausal negation (Truong et al., 2023). Mondorf and Plank (2024) introduced a benchmark for suppositional reasoning that convincingly shows that "proficient models struggle with inferring the logical implications of potentially false statements". Srikanth and Rudinger (2025); Srikanth et al. (2024) measure the inferential consistency of a model, the degree to which a model makes consistently correct or incorrect predictions about the same fact under different contexts.

3. Analysis of LLM Failures & Characteristics of Negation Handling Researchers have investigated whether the semantic constraints of function words like negation are adequately learned and how context impacts their embeddings (Kalouli et al., 2022). It has been shown that LLMs tend to underestimate the significant impact of negation on sentence meaning (Anschütz et al., 2023). A broad analysis across various LLMs, including instruction-tuned models, confirmed their inability to capture the lexical semantics of negation and to reason effectively under negation, with model size not initially appearing as a clear mitigating factor (Truong et al., 2023) (this paper re-evaluates the model size aspect). Furthermore, negation tokens like "not" appear to have a limited effect on the distributional representations learned by models (Singh et al., 2024). The phenomenon often leads to hallucinations in tasks involving negation, such as false premise completion and constrained fact generation (Varshney et al., 2025). Other studies have also investigated the limits of LLMs on general compositional tasks, which implicitly include negation (Dziri et al., 2023).

4. Negation in Multimodal & Vision-Language Models (VLMs) The challenge of negation extends beyond text-only models. On NegBench, a comprehensive benchmark for VLMs covering retrieval and multiple-choice questions with negated captions across images, videos, and medical data, Alhamoud et al. (2025) find that models often collapse affirmative and negated statements into *similar* embeddings, which is a powerful technical insight. Efforts are being made to address this, such as proposing negation context-aware reinforcement learning feedback loops to ensure generated text and images correctly reflect negated queries (Nadeem et al., 2024). Zhang et al. (2025) find that even state-of-the-art VLMs struggle significantly with negation. They also uncover an interesting "U-shaped" scaling trend, where performance first drops with model size before improving again, which adds nuance to the scaling debate.

5. Mitigation Strategies & Improving Negation Handling Various strategies have been proposed to improve negation handling. These include augmenting training data with more negative examples (Helwe et al., 2022), paraphrasing input text into affirmative terms to remove explicit negation while preserving meaning (Rezaei and Blanco, 2024), and exploring fine-tuning processes, including instruction prompt adjustments, to explicitly account for negation (Anschütz et al., 2023; Truong et al., 2023). Rezaei and Blanco (2025) proposes a self-supervised method to improve negation handling by pre-training models on a novel task called Next Sentence Polarity Prediction (NSPP). They show that further pre-training on this task improves performance on downstream negation benchmarks. Castricato et al. (2024) study controllable generation of LLM through the lens of the "Pink elephant paradox" (Spiers, 2002; Wegner, 1994). They think that form of Reinforcement Learning from AI Feedback (RLAIF, that they term *Direct Principle Feedback*) assesses the problem.

6. Cross-lingual & Multilingual Negation Studies The study of negation awareness has also been extended to multilingual contexts. For example, Hartmann et al. (2021) developed a multilingual benchmark to probe negation-awareness across languages including English, Bulgarian, German,

French, and Chinese, setting a precedent for cross-lingual investigations in this domain. Vrabcová and Sojka (2024) developed CsFEVER dataset and confirmed that negation causes problems in Czech as well.

In this paper, we focus on the effects of moderating variables on robustness to negation, posed by the following research questions:

RQ1: Can the *scale of model size* improve models' ability to handle negations?
In one of the findings of Truong et al. (2023), it is posited that the LLMs are unable to reason under negation, with the increased size of the LLM not being a mitigating factor. We review this claim by evaluating models from the same LLM family with varying model sizes.

RQ2: Is the models' capability to handle negations *language-dependent*?
As most studies concerning negation in LLMs focus on English datasets, we extend our scope to include multiple languages with differing levels of word-order flexibility.

The primary contributions of this work are:

- The development and release of two novel, multilingual textual entailment datasets specifically designed to evaluate negation. These datasets span English, Czech, German, and Ukrainian and feature paired examples differing only in negation.
- Empirical evidence demonstrates that larger model sizes can indeed improve the handling of negations, offering a counterpoint to some earlier research. An analysis reveals that while linguistic features of a language play a role, the specificity and length of the premise context have a more pronounced impact on an LLM's robustness to negation than the language itself.

The paper is structured as follows. Section 2 presents the preparation of our two new NLI datasets. Section 3 describes the methodology and the LLMs we use in our experiments. Results are discussed in Section 4. We conclude and outline future directions in Section 5.

2 Datasets

We evaluate the accuracy of the models in our experiments using two datasets that have been originally formatted for the Natural Language Inference (NLI) task, i.e., containing premise-hypothesis pairs and the label denoting their logical relationship (entailment, neutral, contradiction).

SNLI (Bowman et al., 2015) is a popular NLI dataset focusing on the evaluation of general and common sense knowledge of the LLMs. The premises are captions of images capturing everyday situations and, as such, are quite short, consisting of one to two sentences. The hypotheses have been generated manually by human annotators.

FEVER-NLI (Thorne et al., 2018) is a modification of the FEVER dataset, which originally focused on the task of fact extraction and verification. The dataset consists primarily of factual statements retrieved from Wikipedia.

Both of the original datasets are available under permissive Creative Commons licenses, The SNLI dataset is licensed under CC BY-SA 4.0 and the FEVER NLI dataset is licensed under CC BY-SA 3.0.

As we aim to study the effects of negation on the accuracy of the models, we modify the format of these datasets for the TE task and extend them to contain the negation of the original hypothesis. The original datasets contain a low percentage of hypotheses with common negation markers, such as *no, n't, not, nobody, neither,* or *nothing*, and differ in the verbosity of premises, as seen in the Table 1.

In Table 2, we include average lengths of premises and hypotheses of the newly prepared datasets for comparison, NoFEVER-ML and NoSNLI-ML.

	SNLI	FEVER-NLI
Rows in the dataset	20,000	10,000
Rows w/ negation	868	189
% of rows w/ negation	4.34	1.89
Premise word count	13.94	51.96
Hypothesis word count	7.50	8.80

Table 1: Comparison of the original English SNLI and FEVER-NLI datasets

2.1 Preparation of new datasets in English

In order to prepare new datasets containing negation, we first sanitize the datasets and then extend the datasets with hypotheses including negation.

	CES FEVER MANUAL	CES FEVER	CES SNLI	ENG FEVER	ENG SNLI	DEU FEVER	DEU SNLI	UKR FEVER	UKR SNLI
Premise (chars)	268.36	265.06	51.71	266.31	58.02	306.60	70.33	278.54	58.55
Hypothesis without negation (chars)	42.87	41.69	24.06	42.12	28.71	49.63	34.55	43.38	27.71
Hypothesis with negation (chars)	44.89	43.58	26.42	51.13	37.26	53.69	38.87	45.11	29.70
Premise (words)	49.36	48.85	10.74	54.12	14.05	55.43	13.58	47.90	10.55
Hypothesis without negation (words)	7.86	7.66	5.20	8.72	7.13	8.97	7.01	7.35	4.99
Hypothesis with negation (words)	7.86	7.69	5.32	10.95	9.23	9.82	7.93	8.34	5.99

Table 2: Comparison of average lengths across evaluated datasets, NoFEVER-ML and NoSNLI-ML.

Contrary to Gururangan et al. (2018), we do not distinguish between contradiction and logical negation of tokens expressing negation. We then manually replace incorrectly generated sentences and create 2 datasets.

NoFEVER-ENG dataset: This dataset contains 2,600 premise-hypothesis-negated_hypothesis triplets, with 1,513 premises entailing the hypothesis without negation, and 1,087 premises entailing the hypothesis with negation.

NoSNLI-ENG dataset: This dataset contains 6,261 triplets, with 3,245 premises entailing the hypothesis without negation, and 3,016 premises entailing the hypothesis with negation.

Sanitization and Filtration Firstly, we sanitize the datasets by removing all incomplete or invalid rows, i.e., removing rows with empty premises, empty hypotheses, or invalid labels. Secondly, we remove rows that have the *neutral* entailment label, as the effect of negation on their entailment value is not uniform.

Negation We employ the `negate` Python module for the first phase of our negation process. This module is a rule-based negation tool for the English language, utilizing the `spaCy` model `en_core_web_md` for POS tagging and dependency parsing in order to correctly locate the verb in the sentence. The verb is then negated by the inclusion of an auxiliary verb and a negation marker, with our preference in the module set to the contraction *n't*. The tool is able to negate in both directions, meaning that it can remove negation from the verb as well.

Rows with hypotheses that do not contain verbs are discarded during the generation process.

Manual verification As the tool we used for the hypothesis generation is rule-based, it comes with certain limitations and cases that are not covered by the rules and have to be fixed manually. For the second phase of our negation process, we have performed a full manual check to ensure the correctness of the newly created hypotheses.

***Some, any*, nor *yet* are properly supported.**
Example 1:
Input: There is someone jumping.
Negate output: There isn't someone jumping.
Manual change: There isn't anyone jumping.
Example 2:
Input: Keegan-Michael Key has yet to play anyone.
Negate output: Keegan-Michael Key doesn't have yet to play anyone.
Manual change: Keegan-Michale Key has played someone.

The rules of the module as aimed at the *Not-negation*, most likely due to its higher frequency in the English language, in comparison to the *No-negation*.
Example 3:
Input: There is no one running.
Negate output: There isn't no one running.
Manual change: There is someone running.

The typos in the original datasets have been fixed only in cases where the typo directly impacted the negation of the verb.

2.2 Translation of new datasets into Czech, German, and Ukrainian

To explore the impact of negation on the accuracy of the LLMs across languages as started by (Hartmann et al., 2021) for English, Bulgarian, German, French and Chinese, we translate our English (ENG) datasets into Czech (CES), German (DEU), and Ukrainian (UKR). We have chosen these languages due to their differing features, such as mor-

phology, orthography, alphabet, and projectivity, as specified in Table 3. Additionally, we highlight Slavic-language negation phenomenon, **negative concord**, in Czech and Ukrainian, where the negation of multiple words within the sentence is interpreted as a single semantic negation. While similar constructions can appear in English, it is generally classified as gramatically incorrect and is not a part of our English datasets. This phenomenon may challenge the LLMs, which could misinterpret the multiple negated words as separate negations.

Translator Selection There are many machine translation tools to choose from. To choose between Google Translate and DeepL, we have translated 100 rows from the modified English FEVER-NLI dataset into Czech, comparing the quality of translation between the translations and to a preexisting translation of the FEVER-NLI dataset that has been manually verified by a human annotator. From these 100 rows, DeepL provided higher quality for 44 rows, Google Translate for 30, with 26 rows having equal translation quality between them.

We carry out the Statistical Sign Test with the ties equally split and with the alpha level $\alpha = 0.05$. The null hypothesis is that both translators are of equal quality. Our result value is p-value$= 0.2$ and thus we cannot reject our null hypothesis.

As the difference between the translations is not statistically significant in the 100 examples, we have chosen to use DeepL to translate the datasets due to the higher percentage of winning translations.

Translation validation To validate the translations, we manually reviewed a random sample of 600 rows from each translated dataset. Our review focused on preserving the logical relationship and the correct application of negation. Across this substantial sample, we found no systematic errors, confirming the high quality of the machine translation for this task.

Dataset Availability Our datasets are publicly available in a Zenodo repository (Vrabcová et al., 2025a).

3 Methodology

To study the effect of negation on the reasoning reliability of large language models, we evaluate them on *textual entailment* (Putra et al., 2024). In this task, the model receives two claims, a *premise* and a *hypothesis*, and decides whether the hypothesis follows from the premise.

In our experiments, we inspect whether the presence of **lexical negation** (e.g. "not") in the hypothesis *causes* degradation in the accuracy of the language model reasoning. For this purpose, we prepare a set of paired evaluation sentences that differ only in a negation of the primary verb. We call the variants **hypotheses without negation** and **hypotheses with negation** based on whether they contain a negation. Necessarily, if one follows from the premise, the other contradicts it, and vice versa. If the models' reasoning is robust to negations, its accuracy should be the same between the two groups. We deliberately chose such a simple task in order to minimize confounding factors and to demonstrate that the difficulty with negation arises even in the most basic form of the task.

The evaluated models are LLama 3 (Llama Team, 2024b) with 3B, 8B, and 70B parameters (Llama Team, 2024a), Qwen 2.5 (Yang et al., 2024) with 1.5B, 7B, and 72B parameters (Qwen Team, 2024), and Mistral Nemo 12B, Small 22B, and Large 123B (Mistral AI, 2024).

All of the evaluated models are publicly available state-of-the-art models based on the `transformer` architecture (Vaswani et al., 2017), have open weights, a multi-lingual pre-training, and are fine-tuned to follow user instructions. In the evaluation, we prompt the model to judge if a hypothesis follows from a premise or not. We let the model finish a response "The answer is:" with either *True* or *False* using a greedy decoding. To ensure the response follows the predefined structure, we employ Outlines (Willard and Louf, 2023) and LMFE (Gat, 2024). More details of the prompt format can be found in Appendix B.

Our implementation for running the experiments is publicly available as a GitHub repository (Vrabcová et al., 2025b).

4 Results

We summarize our results in this section, including the results of the statistical tests to support these claims. We find that accuracy and negation sensitivity improve with increasing model size and the information specificity of dataset premises, while variations in languages are not statistically significant.

Unless otherwise stated, for the statistical tests, we are using an alpha level (significance threshold)

	ENG	CES	DEU	UKR
Language Family	West Germanic	West Slavic	West Germanic	East Slavic
Script	Latin	Latin	Latin	Cyrillic
Linguistic Typology	Analytic	Fusional	Fusional w/ Agglutinative Features	Fusional
Word Order Flexibility	Low	High	Moderate	High

Table 3: Comparison of linguistic features between languages English (ENG), Czech (CES), German (DEU), and Ukrainian (UKR).

$\alpha = 0.05$. More details of the specific values used during computation can be found in Appendix C.

Model size has a positive correlation with models' ability to handle negation. Figure 2 shows the accuracy change of different models for the hypotheses with and without negation. We found that all evaluated models perform better than a random guess and that the larger models *narrow* the gap between the accuracies, with two outliers on the SNLI dataset: a slight increase on the middle-sized Llama model (8B) and a marked drop on the middle-sized Mistral model (22B).

Using the results shown in Figure 4, we compute the correlation of the model size and the relative accuracy change for each evaluated dataset, as well as the overall correlation. Using the Shapiro-Wilk test, we find that the majority of datasets do not have a normal distribution over the different model sizes. Thus, we compute correlation using the Spearman correlation coefficient, as the Pearson correlation coefficient presumes a normal distribution of values.

The results indicate a strong positive correlation between model size and the relative accuracy change on the FEVER dataset and a weak-to-moderate positive correlation on the SNLI dataset. For an averaged relative accuracy change for all languages per model, the results indicate a strong positive correlation with the Spearman correlation coefficient of 0.867, supporting out RQ1.

We posit that the differences in correlation are due to different philosophies behind the datasets. The FEVER dataset was originally based on the fact verification dataset and consists of factually specific premises that are sufficient for the prediction of the textual entailment. The SNLI dataset, on the other hand, relies more heavily on implicit semantical patterns that the models learn during training, introducing an uncertainty we cannot easily predict.

Our findings are in contrast to the results of (Truong et al., 2023, Section 3, Finding 1), where it is posited that pre-trained language models perform comparably or worse than a random baseline, and model scaling has almost no effect. The difference in our findings can be caused by a particular choice of the evaluated models and their level of fine-tuning, or the increase in the quality of the models themselves since the prior study. Differences in methodology may also contribute, as our benchmark specifically focuses on simple examples of lexical negation. Another plausible explanation might be that the measured accuracy strongly benefits from fine-tuning of prompts.

Larger LMs are more sensitive to negation for TE task. Table 4 shows the negation sensitivity of each model across individual datasets. Our negation sensitivity metric represents the percentage of datasets in which the model assigned opposite entailment scores to pairs of hypotheses with and without negation. The best results were achieved only by the biggest models, supporting our RQ1.

Similarly to the relative accuracy change, we compute the correlation between model size and the model's sensitivity to negation. As per the computed p-values for the Shapiro-Wilk test, we can see that the majority of our negation sensitivity metric across datasets does not have a normal distribution either. Therefore, we are again using the Spearman correlation.

The results indicate a very strong positive correlation of negation sensitivity with the model size on the FEVER datasets, with a rather weak correlation on the SNLI datasets. For an average negation sensitivity for all languages per model, the results indicate a strong positive correlation, with a Spearman correlation coefficient of 0.866.

Larger models are less likely to classify pair hypotheses with the same textual entailment label, and the impact of negation is more effectively propagated through the model. One outlier in our results

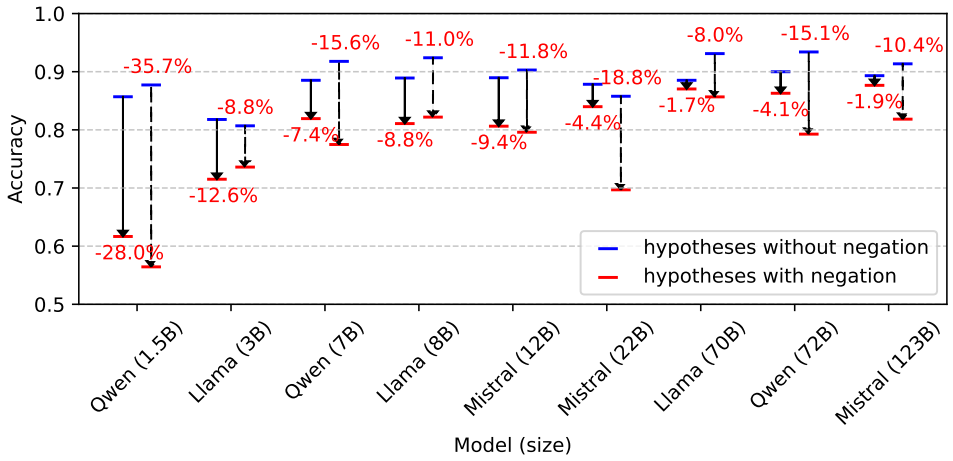

Figure 2: **Robustness to negations across model sizes:** Models' accuracy difference on textual entailment when facing a hypothesis (i) including and (ii) not including a negation. The results are an aggregate across four languages (CES, DEU, ENG, UKR) measured for FEVER (solid line; left) and SNLI (dashed line; right) datasets.

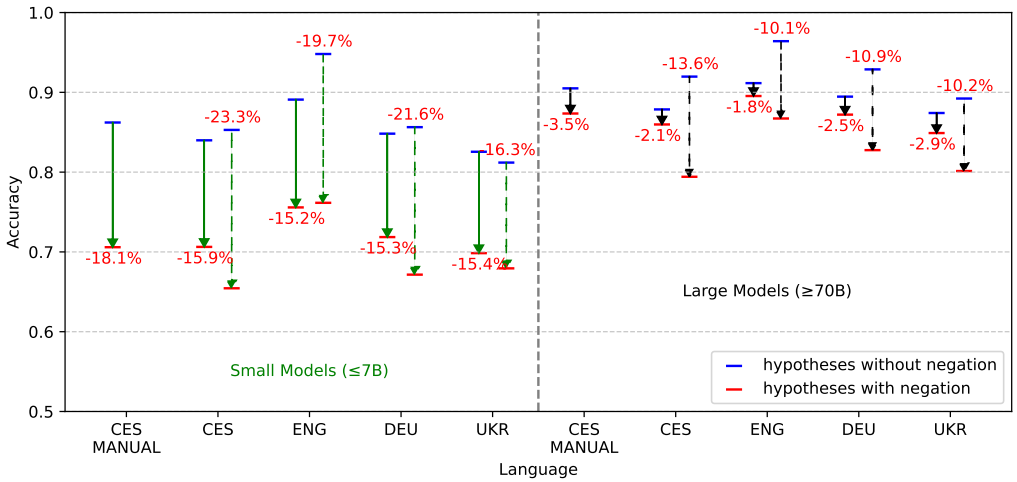

Figure 3: **Robustness to negations across languages:** Accuracy difference of *smaller* (≤7B; left) and *large* models (≥70B; right) on textual entailment when facing a hypothesis (i) including and (ii) not including a negation. Inputs are *identical* across *different* languages. The results are aggregated across three models in each size group, measured for two datasets: FEVER (solid line) and SNLI (dashed line).

is the Mistral (22B) model, which has been outperformed by the Mistral (12B) model on the CES, ENG, and UKR SNLI datasets, as well as on the DEU FEVER dataset.

Language plays a role in the models' accuracy. Figure 3 shows the accuracy change on different models with regard to the language of the dataset. As the language is a categorical variable, instead of correlation, we compute and compare the variance of the relative accuracy change across the four evaluated languages (CES, ENG, DEU, UKR), excluding the Czech manually translated dataset. We aggregate the values shown in Figure 4 per language and use the Friedman test as our analysis method. Our null hypothesis is that all languages have the same average of relative accuracy change, and the alternate hypothesis is that the averages are not the same.

Analysis has shown that for our degrees of freedom $d_{fn} = 3$, $d_{df} = 33$, the results are F-statistic = 13.000 and p-value = 0.0046.

As the p-value is less than our alpha level, we can reject our null hypothesis. Thus, we conclude that there is a significant difference in the relative accuracy change across different languages.

Delving further and applying the Wilcoxon Signed-Rank Test on the results of language pairs per specific dataset, we find that there is a statistically significant difference on the Czech SNLI dataset and all of the other languages, with the models performing worse on the Czech dataset. On the FEVER dataset, the models performed the worst

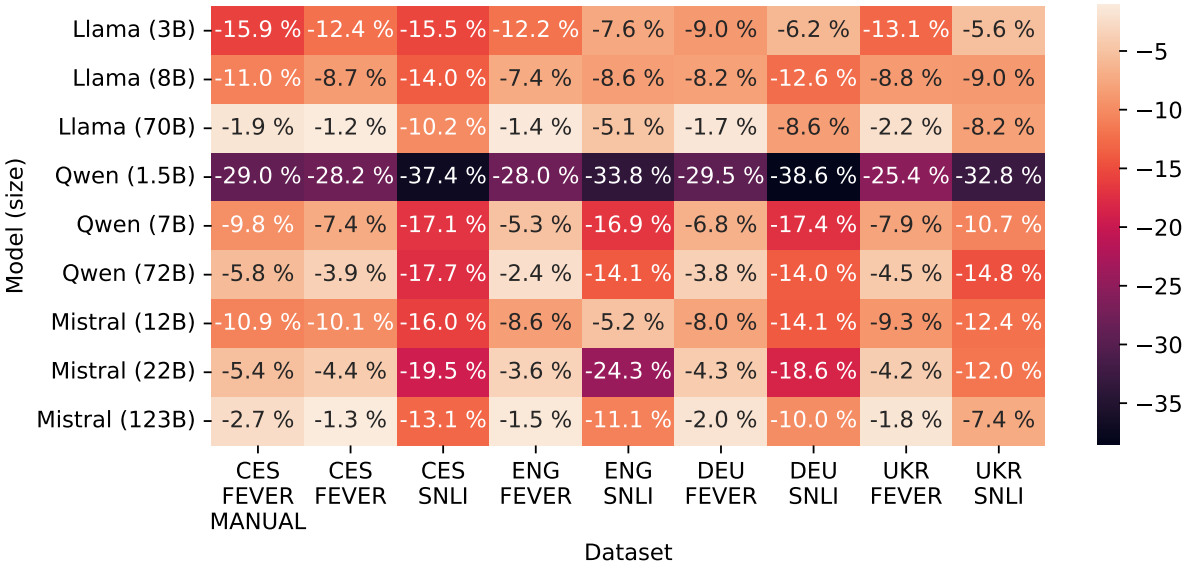

Figure 4: **Robustness to negations in detail:** Accuracy difference caused by negation for all 9 evaluated models across all 4 evaluated languages and our two new datasets. The bigger models consistently handle negation better.

Model (size)	CES FEVER MANUAL	CES FEVER	CES SNLI	ENG FEVER	ENG SNLI	DEU FEVER	DEU SNLI	UKR FEVER	UKR SNLI
Llama (3B)	68.04 %	68.04 %	64.88 %	73.54 %	81.23 %	49.08 %	61.85 %	65.19 %	68.36 %
Llama (8B)	80.5 %	80.5 %	75.87 %	88.62 %	86.93 %	78.07 %	84.46 %	80.38 %	**79.08 %**
Llama (70B)	87.62 %	87.62 %	**80.95 %**	**94.81 %**	**92.29 %**	**84.22 %**	**89.69 %**	84.77 %	78.98 %
Qwen (1.5B)	55.92 %	55.92 %	40.44 %	64.77 %	62.37 %	48.08 %	58.58 %	54.31 %	37.85 %
Qwen (7B)	81.85 %	81.85 %	71.03 %	89.54 %	77.35 %	70.95 %	83.85 %	79.46 %	72.85 %
Qwen (72B)	87.65 %	87.65 %	71.83 %	92.65 %	80.93 %	75.56 %	88.38 %	**85.38 %**	70.34 %
Mistral (12B)	80.65 %	80.65 %	69.4 %	87.88 %	87.06 %	74.8 %	83.81 %	77.04 %	66.19 %
Mistral (22B)	85.23 %	85.23 %	62.11 %	91.46 %	66.79 %	63.79 %	87.08 %	77.81 %	62.08 %
Mistral (123B)	**87.81 %**	**87.81 %**	72.1 %	91.31 %	80.59 %	77.3 %	89.08 %	84.23 %	74.09 %

Table 4: Evaluation of the negation sensitivity of each model across individual datasets, with the best results bolded. Our negation sensitivity metric represents the percentage of datasets in which the model assigned opposite entailment scores to pairs of hypotheses with and without negation. The best results (in bold) were achieved only by the biggest models supporting our RQ1.

	ENG	DEU	UKR
CES	0.030	0.0273	0.0390
ENG		0.4961	0.4961
DEU			0.0195

	ENG	DEU	UKR
ENG		0.6523	0.6523
DEU			0.0977
UKR			0.2031

Table 5: p-values of Wilcoxon Signed-Test for language pairs on the SNLI (left) and the FEVER (right) datasets

on the Czech dataset as well, but the difference is statistically significant only in comparison to the English dataset.

Our results partially conform to our presuppositions in RQ2 that the models will perform worse in languages with higher word-order flexibility. However, part of the observed differences may also be explained by language-specific phenomena, such as **negative concord** in Czech, where a single negation is expressed by multiple negated words, or by biases in the models, as English and German are likely more prevalent in their training data.

5 Conclusions and future work

This work presents novel datasets enabling the evaluation of language models' robustness in handling negations in four languages.We have extended existing FEVER (Thorne et al., 2018) and SNLI (Bowman et al., 2015) datasets to include paired hypotheses that are logically opposite to each other by negating the primary verb in the

original hypothesis. We have evaluated the models' accuracy separately on **hypotheses with and without negation** and computed the accuracy drop that occurred due to the addition of negation in the hypothesis. New datasets NoFEVER-ML and NoSNLI-ML are freely available (Vrabcová et al., 2025a), together with the code used in the negation project (Vrabcová et al., 2025b).

We find that model size plays a key role in handling negation correctly, countering the previous work of Truong et al. (2023). The results on differing language datasets only partially match the initial expectations, with models performing worse on the Czech configuration. This may be due to the relative simplicity of hypotheses, which do not lend to the inclusion of more complex negation structures, such as the negation token being farther away from the verb, as may be the case in German.

In regard to the differing accuracy from one dataset to another, we hypothesize that the accuracy is also heavily dependent on the specificity of information in the premise; the more common-sense SNLI dataset performed worse than the more fact-explicit FEVER dataset. Our finding that negation handling is less reliable on common-sense tasks (SNLI) suggests a potential source for *hallucinations*, where models may default to learned semantic patterns over strict logical constraints. Improving robustness to negation is, therefore, a direct step towards more factually grounded and reliable LLMs.

Our work provides new resources for addressing the deficiency of language models in handling negations. It shows that while the 'pink elephant' of negation is still in the room, larger models are becoming better at seeing what isn't there, particularly when given a clear description of the room's other contents.

In future, studying and understanding which part of model architecture is responsible for the problems with negation is our ultimate goal. We will use the lens of LLM-microscope (Razzhigaev et al., 2025) to look at the tokens that play decisive role, depending on subsets of our newly created datasets.

Acknowledgments

Michal Štefánik's work has been supported by the Research Council of Finland through project No. 353164 "Green NLP – controlling the carbon footprint in sustainable language technology". This project has received funding from the European Union's Horizon Europe research and innovation programme under Grant agreement No. 101070350 and from UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee (grant number 10052546). The contents of this publication are the sole responsibility of its authors and do not necessarily reflect the opinion of the European Union.

Limitations

The details of model preparation (training data, hyperparameters) limit the full understanding of the results we report, as we use pretrained models. Additionally, this paper does not delve into the effects of post-training and fine-tuning, and possible exposure to classes of negation examples as contradictory usage of negation.

Our study focuses on *lexical negation* (e.g., "not") and does not cover more complex forms of negation, such as *negation scope* or *double negation*, which can further challenge model reasoning in natural language inference. Similarly, it does not cover *negative polarity items* or *downward entailment* from a superset to a subset (e.g., from "not animal" to "not dog"), which are also known to pose difficulties for language models. We deliberately focus on lexical negation to isolate its effect and show that the problem with negation persists even in this simplified setting. For a similar reason, we focus exclusively on the *textual entailment* task.

The results may be biased by the semi-automated translations of datasets into their German, Czech, and Ukrainian versions, as shown in Table 6. Although the translation quality of the model we have used is generally high, it is still an LLM that is not immune to hallucinations on rarely seen data. We have, for example, in the Ukrainian SNLI dataset, observed the outlier case shown in Table 6. This behavior does not occur when the sentence does not include negation. We suppose that the combination of the inclusion of negation, the tense in the original English sentence, and the inclusion of an initialism (*pb and j*) created confusion in the model. However, as the training data of the model isn't public, we cannot confirm our hypothesis.

However, we did a manual check to ensure the quality of a subset of the translations and did not encounter any errors related to the application of negation in context.

Additionally, our study is limited to two Germanic (English, German) and two Slavic (Czech,

Ukrainian) languages, which constrains the generalizability of our findings. Evaluating additional languages from other language families could reveal further challenges and insights regarding negation in LLMs.

References

Kumail Alhamoud, Shaden Alshammari, Yonglong Tian, Guohao Li, Philip H.S. Torr, Yoon Kim, and Marzyeh Ghassemi. 2025. Vision-Language Models Do Not Understand Negation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 29612–29622.

Miriam Anschütz, Diego Miguel Lozano, and Georg Groh. 2023. This is not correct! Negation-aware Evaluation of Language Generation Systems. In *Proceedings of the 16th International Natural Language Generation Conference*, pages 163–175, Prague, Czechia. ACL.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. ACL.

Louis Castricato, Nathan Lile, Suraj Anand, Hailey Schoelkopf, Siddharth Verma, and Stella Biderman. 2024. Suppressing Pink Elephants with Direct Principle Feedback. *Preprint*, arXiv:2402.07896.

Nouha Dziri, Ximing Lu, Melanie Sclar, Xiang Lorraine Li, Liwei Jiang, Bill Yuchen Lin, Sean Welleck, Peter West, Chandra Bhagavatula, Ronan Le Bras, Jena D. Hwang, Soumya Sanyal, Xiang Ren, Allyson Ettinger, Zaid Harchaoui, and Yejin Choi. 2023. Faith and Fate: Limits of Transformers on Compositionality. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Allyson Ettinger. 2020. What BERT Is Not: Lessons from a New Suite of Psycholinguistic Diagnostics for Language Models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

Iker García-Ferrero, Begoña Altuna, Javier Alvez, Itziar Gonzalez-Dios, and German Rigau. 2023. This is not a Dataset: A Large Negation Benchmark to Challenge Large Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8596–8615, Singapore. ACL.

Noam Gat. 2024. LM-Format-Enforcer: A text formatter for language models. `https://github.com/noamgat/lm-format-enforcer`. Accessed: 2025-09-19.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation Artifacts in Natural Language Inference Data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. ACL.

Mareike Hartmann, Miryam de Lhoneux, Daniel Hershcovich, Yova Kementchedjhieva, Lukas Nielsen, Chen Qiu, and Anders Søgaard. 2021. A Multilingual Benchmark for Probing Negation-Awareness with Minimal Pairs. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 244–257, Online. ACL.

Chadi Helwe, Simon Coumes, Chloé Clavel, and Fabian Suchanek. 2022. TINA: Textual Inference with Negation Augmentation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4086–4099, Abu Dhabi, United Arab Emirates. ACL.

Karl Moritz Hermann, Edward Grefenstette, and Phil Blunsom. 2013. "Not not bad" is not "bad": A distributional account of negation. In *Proceedings of the Workshop on Continuous Vector Space Models and their Compositionality*, pages 74–82, Sofia, Bulgaria. ACL.

John D. Ireland. 1997. *The Udana and the Itivuttaka: Two Classics from the Pali Canon*. Buddhist Publication Society, Kandy, Sri Lanka. The parable appears in Udana 6.4.

Aikaterini-Lida Kalouli, Rita Sevastjanova, Christin Beck, and Maribel Romero. 2022. Negation, Coordination, and Quantifiers in Contextualized Language Models. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3074–3085, Gyeongju, Republic of Korea. ACL.

Nora Kassner and Hinrich Schütze. 2020. Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly. In *Proceedings of the 58th Annual Meeting of the ACL*, pages 7811–7818, Online. ACL.

Llama Team. 2024a. Llama AI Models on Hugging Face. Hugging Face Model Repository. Accessed [2025-09-19]. Models: 3B, 8B, 70B.

Llama Team. 2024b. The Llama 3 Herd of Models. *ArXiv*, abs/2407.21783.

Mistral AI. 2024. Mistral AI Models on Hugging Face. Hugging Face Model Repository. Accessed [2025-09-19]. Models: 12B Nemo, 22B Small, 123B Large.

Philipp Mondorf and Barbara Plank. 2024. Liar, Liar, Logical Mire: A Benchmark for Suppositional Reasoning in Large Language Models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7114–7137, Miami, Florida, USA. ACL.

	Original/Translated Sentence	Sentence in English
ENG	A man isn't eating pb and j	A man isn't eating pb and j
CES	Muž nejí pb a j.	A man isn't eating pb and j.
DEU	Ein Mann isst kein Pb und J	A man isn't eating Pb and J
UKR	Чоловік не їсть пломбір з соєвими бобами	A man isn't eating ice cream with soy beans

Table 6: Translation of an outlier example sentence using DeepL

Mohammad Nadeem, Shahab Saquib Sohail, Erik Cambria, Björn W. Schuller, and Amir Hussain. 2024. Negation Blindness in Large Language Models: Unveiling the NO Syndrome in Image Generation. *Preprint*, arXiv:2409.00105.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? Sentiment Classification using Machine Learning Techniques. In *Proc. of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, pages 79–86, Philadelphia, Pennsylvania, USA. ACL.

I Made Suwija Putra, Daniel Siahaan, and Ahmad Saikhu. 2024. Recognizing textual entailment: A review of resources, approaches, applications, and challenges. *ICT Express*, 10(1):132–155.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Anton Razzhigaev, Matvey Mikhalchuk, Temurbek Rahmatullaev, Elizaveta Goncharova, Polina Druzhinina, Ivan Oseledets, and Andrey Kuznetsov. 2025. LLM-Microscope: Uncovering the Hidden Role of Punctuation in Context Memory of Transformers. In *Findings of the ACL: NAACL 2025*, pages 7757–7764, Albuquerque, New Mexico. ACL.

MohammadHossein Rezaei and Eduardo Blanco. 2024. Paraphrasing in Affirmative Terms Improves Negation Understanding. In *Proc. of the 62nd Annual Meeting of the ACL (Volume 2: Short Papers)*, pages 602–615, Bangkok, Thailand. ACL.

MohammadHossein Rezaei and Eduardo Blanco. 2025. Making Language Models Robust Against Negation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the ACL: Human Language Technologies (Volume 1: Long Papers)*, pages 8123–8142, Albuquerque, New Mexico. ACL.

Jaisidh Singh, Ishaan Shrivastava, Mayank Vatsa, Richa Singh, and Aparna Bharati. 2024. Learn "No" to Say "Yes" Better: Improving Vision-Language Models via Negations. *Preprint*, arXiv:2403.20312.

Jude A. Spiers. 2002. The Pink Elephant Paradox (or, Avoiding the Misattribution of Data). *International Journal of Qualitative Methods*, 1(4):36–44.

Neha Srikanth, Marine Carpuat, and Rachel Rudinger. 2024. How Often Are Errors in Natural Language Reasoning Due to Paraphrastic Variability? *Transactions of the Association for Computational Linguistics*, 12:1143–1162.

Neha Srikanth and Rachel Rudinger. 2025. NLI under the Microscope: What Atomic Hypothesis Decomposition Reveals. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2574–2589, Albuquerque, New Mexico. ACL.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a Large-scale Dataset for Fact Extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. ACL.

Thinh Hung Truong, Timothy Baldwin, Karin Verspoor, and Trevor Cohn. 2023. Language models are not naysayers: An analysis of language models on negation benchmarks. In *Proc. of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, pages 101–114, Toronto, Canada. ACL.

Neeraj Varshney, Satyam Raj, Venkatesh Mishra, Agneet Chatterjee, Amir Saeidi, Ritika Sarkar, and Chitta Baral. 2025. Investigating and Addressing Hallucinations of LLMs in Tasks Involving Negation. In *Proceedings of the 5th Workshop on Trustworthy NLP (TrustNLP 2025)*, pages 580–598, Albuquerque, New Mexico. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. In *Proc. of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pages 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Tereza Vrabcová, Marek Kadlčík, Petr Sojka, Michal Štefánik, and Michal Spiegel. 2025a. Datasets for: Towards the roots of the negation problem: A multilingual nli dataset and model scaling analysis.

Tereza Vrabcová, Marek Kadlčík, Petr Sojka, Michal Štefánik, and Michal Spiegel. 2025b. Towards the roots of the negation problem: A multilingual

nli dataset and model scaling analysis. https://github.com/MIR-MU/negation.

Tereza Vrabcová and Petr Sojka. 2024. Negation Disrupts Compositionality in Language Models: The Czech Usecase. In *Proceedings of the 18th Workshop on Recent Advances in Slavonic Natural Language Processing, RASLAN 2024, Kouty nad Desnou, Czech Republic, December 6–8, 2024*, pages 17–24, Brno. Tribun EU.

Daniel M Wegner. 1994. Ironic processes of mental control. *Psychological review*, 101(1):34.

Brandon T Willard and Rémi Louf. 2023. Efficient Guided Generation for Large Language Models. *arXiv preprint arXiv:2307.09702*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. 2024. Qwen2 Technical Report. *Preprint*, arXiv:2407.10671.

Yuhui Zhang, Yuchang Su, Yiming Liu, and Serena Yeung-Levy. 2025. NegVQA: Can vision language models understand negation? In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 3707–3716, Vienna, Austria. ACL.

Yuhui Zhang, Michihiro Yasunaga, Zhengping Zhou, Jeff Z. HaoChen, James Zou, Percy Liang, and Serena Yeung. 2023. Beyond Positive Scaling: How Negation Impacts Scaling Trends of Language Models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7479–7498, Toronto, Canada. ACL.

A Full details of accuracy across evaluated models

We include figures 5 and 6 as the full tables of accuracies of models for hypotheses with and without negation on the task of textual entailment. In both tables, we can see that almost all models achieve the highest accuracy on the English datasets, on both hypotheses, with and without negation.

However, the accuracy of the models from the Qwen family shown in Figure 4 gave us pause, particularly the fact that overperformance is present only for one specific dataset. The English SNLI dataset with hypotheses without negation is very similar to the original SNLI dataset, as less than 5 percent of hypotheses in the original test dataset contain negation. As the original SNLI is a very popular benchmark for the evaluation of NLI and textual entailment tasks, and the Qwen overperformance did not translate across languages on the same dataset, nor did it translate into an increase of accuracy on hypotheses with negation, we have to wonder whether the training data of the Qwen model family inadvertently included the SNLI evaluation data.

B Prompting template

All the models we evaluated support a chat-like interface with a system prompt, a user prompt, and a model response.

As a system prompt, we used the following template:

> `You are a fact checker for queries in the` **`<English / Czech / German / Ukrainian>`** `language. You will be given a premise, which you know is factually correct, and a hypothesis. You will return the truth value of the hypothesis, based on the premise. Return True if the hypothesis is correct and False if the hypothesis is incorrect.`

The user prompt follows a simple format:

> `Premise:` **`<premise>`**
> `Hypothesis:` **`<hypothesis>`**

The model decides by finishing a prepared response:

> `The answer is:` **`<True / False>`**

C Statistical tests

In tables 7 and 8 we provide the values used during statistical computations explained in Section 4.

D Sentence complexity

As mentioned in Section 4, we posit that the larger models can more accurately capture nuanced language patterns, enabling them to correctly handle negations in more complex sentences that elude the smaller models. From each model family, we take

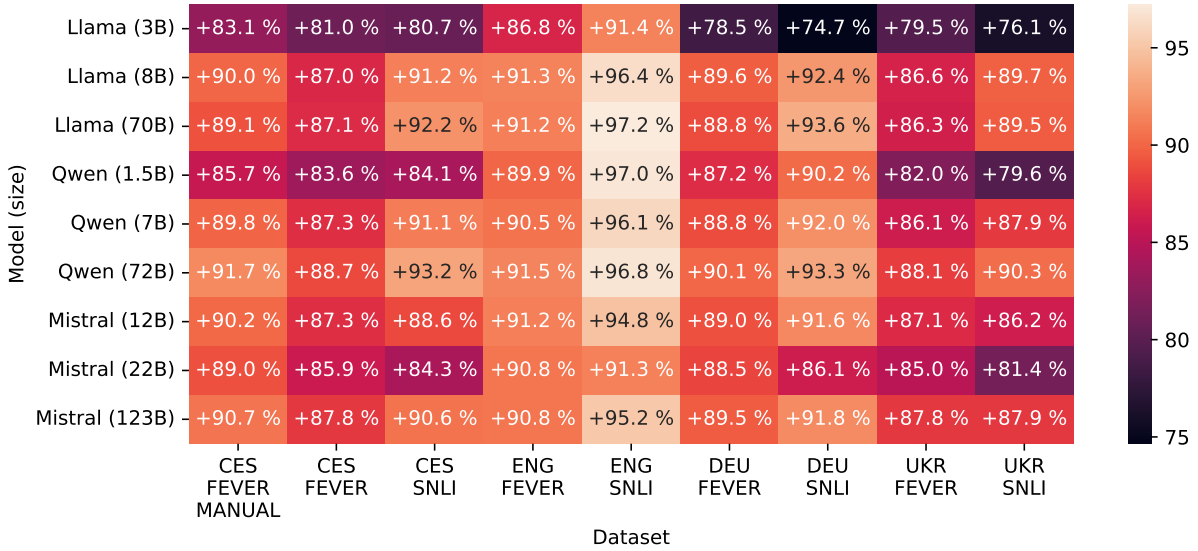

Figure 5: Absolute accuracy for all evaluated models across all evaluated languages and datasets for textual entailment of hypotheses without negation.

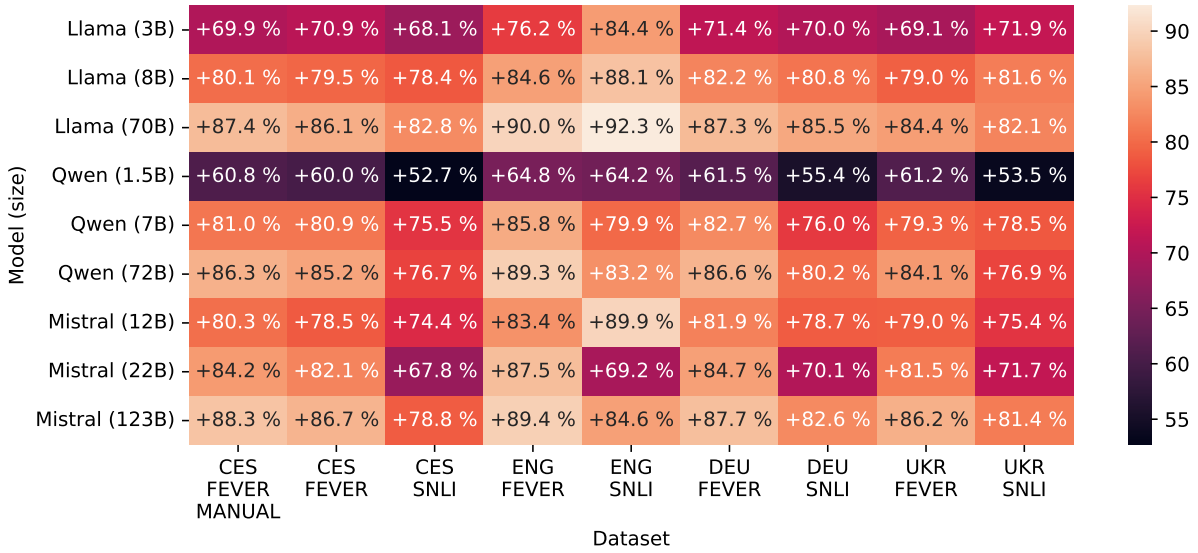

Figure 6: Absolute accuracy for all evaluated models across all evaluated languages and datasets for textual entailment of hypotheses with negation.

	CES FEVER MANUAL	CES FEVER	CES SNLI	ENG FEVER	ENG SNLI	DEU FEVER	DEU SNLI	UKR SNLI	UKR FEVER
Shapiro-Wilk test p-value	0.279	0.004	0.008	0.134	0.001	0.016	0.046	0.003	0.016
Spearman correlation coefficient	0.867	0.883	0.400	0.883	0.250	0.900	0.300	0.883	0.150

Table 7: Values used during the computation of correlation between the model size and the relative accuracy change across all evaluated datasets.

the smallest and the largest models and compare the average sentence complexity of the hypotheses that the model is able to correctly classify. However, we hypothesize that many hypotheses correctly classified by the smaller model are also correctly classified by the larger model. Therefore, for the larger model, we compute the average sentence complexity of the additional correctly classified hypotheses, i.e., the subset for the larger model consists of hypotheses that the larger model correctly classified, which are not present in the smaller model's subset.

	CES FEVER MANUAL	CES FEVER	CES SNLI	ENG FEVER	ENG SNLI	DEU FEVER	DEU SNLI	UKR SNLI	UKR FEVER
Shapiro-Wilk p-value	0.013	0.013	0.052	0.008	0.405	0.101	0.002	0.027	0.022
Spearman correlation w/ model size	0.950	0.950	0.583	0.817	0.267	0.683	0.900	0.817	0.400

Table 8: Values used during the computation of correlation between the model size and the negation sensitivity across all evaluated datasets.

	CES FEVER MANUAL	CES FEVER	CES SNLI	ENG FEVER	ENG SNLI	DEU FEVER	DEU SNLI	UKR FEVER	UKR SNLI
Llama Family	0.027	0.215	**0.064**	0.270	**0.196**	0.219	**0.056**	**0.155**	**0.049**
Qwen Family	**0.082**	0.173	0.006	0.215	0.174	0.024	0.030	0.148	0.041
Mistral Family	−0.016	**0.184**	0.037	**0.271**	0.137	**0.245**	0.001	0.121	−0.017

Table 9: Increase of average depth of **hypotheses without negation** that were correctly predicted only by the largest model of the LM family and the correctly predicted hypotheses by the smallest model of the LM family.

	CES FEVER MANUAL	CES FEVER	CES SNLI	ENG FEVER	ENG SNLI	DEU FEVER	DEU SNLI	UKR FEVER	UKR SNLI
Llama Family	0.019	**0.076**	**0.090**	0.089	**0.145**	0.041	**0.030**	0.150	**0.067**
Qwen Family	**0.038**	−0.09	0.027	−0.016	0.065	−0.053	−0.084	−0.006	−0.080
Mistral Family	−0.016	0.031	0.064	**0.125**	0.101	**0.103**	−0.012	**0.151**	−0.017

Table 10: Increase of average depth of **hypotheses with negation** that were correctly predicted only by the largest model of the LM family and the correctly predicted hypotheses by the smallest model of the LM family.

	CES FEVER MANUAL	CES FEVER	CES SNLI	ENG FEVER	ENG SNLI	DEU FEVER	DEU SNLI	UKR FEVER	UKR SNLI
Llama Family	**2.52 %**	−0.12 %	−4.31 %	−5.77 %	−15.30 %	−4.73 %	−16.01 %	−1.20 %	−6.83 %
Qwen Family	−0.47 %	4.19 %	**2.55 %**	**7.88 %**	−5.00 %	0.90 %	3.58 %	**4.66 %**	**1.44 %**
Mistral Family	0.89 %	**4.49 %**	0.42 %	6.94 %	**−1.63 %**	**1.02 %**	**6.26 %**	2.49 %	−1.92 %

Table 11: Increase of average sentence dissimilarity of premises and the evaluated **hypotheses without negation** that were correctly predicted only by the largest model of the LM family, and the correctly predicted hypotheses by the smallest model of the LM family.

	CES FEVER MANUAL	CES FEVER	CES SNLI	ENG FEVER	ENG SNLI	DEU FEVER	DEU SNLI	UKR FEVER	UKR SNLI
Llama Family	**2.00 %**	3.97 %	−3.02 %	2.36 %	0.62 %	5.52 %	21.12 %	5.56 %	−1.52 %
Qwen Family	0.24 %	**12.49 %**	**19.41 %**	**12.50 %**	**21.14 %**	**10.56 %**	**25.16 %**	**9.44 %**	**15.28 %**
Mistral Family	1.59 %	4.06 %	11.77%	5.22 %	6.11 %	5.30 %	11.96 %	4.94 %	10.36 %

Table 12: Increase of average sentence dissimilarity of premises and the evaluated **hypotheses with negation** that were correctly predicted only by the largest model of the LM family, and the correctly predicted hypotheses by the smallest model of the LM family.

We evaluate the sentence complexity using the following metrics:

Hypothesis depth. Depth of the hypothesis dependency tree. Sentence complexity increases with the depth of the dependency tree.

Tables 9 and 10 show the increase of average hypothesis depth between the smallest and the largest model per the language model family. Due to the relative simplicity of the hypotheses (short length as described in Table 2, average of one verb per hypothesis), the increase of depth is not substantial, being less than one level across all datasets for all language model families.

Dissimilarity of hypothesis to the premise. Number of content words present only in the hypothesis and not the premise, divided by the total number of content words in the hypothesis. Sentence complexity increases with the increase of dissimilarity. We compute the similarity on the lemmatized forms of the words, with the most common stop words per language filtered out. High dissimilarity may be caused by the usage of synonyms, hypernyms, hyponyms, or other linguistic features that require a higher understanding of syntactic patterns in the hypothesis.

Tables 11 and 12 show the increase in the average dissimilarity of the premise and its hypothesis between the smallest and the largest model per the language model family. Here, we can see a marked increase in dissimilarity for hypotheses with negation, especially within the Qwen family models, showcasing the largest model's increased ability to correctly reason and determine entailment on texts that use different formulations to express the same meaning.