Lost in Inference: Rediscovering the Role of Natural Language Inference for Large Language Models

Lovish Madaan[∞◇] **David Esiobu**[∞] **Pontus Stenetorp**[◇♠]
Barbara Plank[♣△] **Dieuwke Hupkes**[∞]
[∞]Meta [◇]University College London
[♠]Research and Development Center for LLMs, National Institute of Informatics
[♣]MaiNLP, Center for Information and Language Processing, LMU Munich, Germany
[△]Munich Center for Machine Learning (MCML), Munich, Germany
{lovish,dieuwkehupkes}@meta.com

Abstract

In the recent past, a popular way of evaluating natural language understanding (NLU), was to consider a model's ability to perform natural language inference (NLI) tasks. In this paper, we investigate if NLI tasks, that are rarely used for LLM evaluation, can still be informative for evaluating LLMs. Focusing on five different NLI benchmarks across six models of different scales, we investigate if they are able to discriminate models of different size and quality and how their accuracies develop during training. Furthermore, we investigate the extent to which the softmax distributions of models align with human distributions in cases where statements are ambiguous or vague. Overall, our results paint a positive picture for the NLI tasks: we find that they are able to discriminate well between models at various stages of training, yet are not (all) saturated. Furthermore, we find that while the similarity of model distributions with human label distributions increases with scale, it is still much higher than the similarity between two populations of humans, making it a potentially interesting statistic to consider.

1 Introduction

Before the state-of-the-art (SoTA) in NLP was constituted almost exclusively by large language models (LLMs), a popular way of evaluating models' understanding of natural language was to consider their ability to perform *natural language inference* (NLI) tasks (most famously, Bowman et al., 2015; Williams et al., 2018). Motivated by the idea that concepts such as entailment and contradiction are central to many aspects of language meaning (Bowman et al., 2015), in NLI tasks, a model is asked to judge the relationship between the meaning of two sentences, typically chosing between entailment, contradiction, and no relationship. Included in the then widely-used natural language understanding benchmark GLUE (Wang et al., 2019b), the NLI

benchmark *Multi-Genre Natural Language Inference* (MNLI, Williams et al., 2018) was up until relatively recently one of the most popular benchmarks to evaluate language models, and is – with over 600 citations to date in 2024 – well-cited even in the recent past.

However, with the arrival of LLMs, MNLI and other datasets have lost their spot on the SoTA leaderboards. With the exception of Brown et al. (2020), who reported low scores for GPT-3 on the NLI benchmark *Adversarial NLI* (ANLI, Nie et al., 2020a), not a single LLM release paper considers an NLI benchmark in their evaluation suite.[1] In this paper, we investigate why. Are NLI benchmarks simply not suitable to evaluate modern-day LLMs? Are their examples too difficult or too easy? Are their scores not informative? Or do they, in fact, still provide a useful signal?

Focusing on five different NLI benchmarks across six different models, we first show that they provide signal and are able to discriminate models of different size and quality (§ 4.1), but that one or more few-shot examples are needed to obtain a somewhat reasonable accuracy for models of any size. We furthermore show that performance on the datasets develop steadily during training, albeit with some fluctuations, making the benchmarks suitable for monitoring training progress (§ 4.2). For some of the benchmarks, accuracies of the best models are reaching 80-90%, for ANLI (Nie et al., 2020a), however, the accuracy of even the best model does not exceed 70%. Furthermore, we show that high scores are not caused by data contamination (§ 4.3).

Next, we investigate the extent to which improvement on the higher-scored benchmarks is still possible (§ 4.4). In a manual analysis, we find that, for

---

[1]Some recent papers do report low scores for LLMs, for e.g. GPT3.5 on XNLI (Ohmer et al., 2024, 2023) and Llama 1 models on ANLI, HANS, and MNLI (McCoy et al., 2019; Weber et al., 2023).

the best performing model, examples with 'incorrect' predictions are rarely in fact incorrect; most concern questions on which humans may disagree as well. Motivated by this result, we further explore this topic by considering the ChaosNLI benchmark (Nie et al., 2020b), which contains 100 human annotations for over 4500 samples for three of the benchmarks we consider (§ 4.5). We find that accuracies (as computed on the majority label) are higher if the entropy of the human labels is low; when humans disagree, models are more likely to select one of the less preferred labels. Lastly, we consider the distributional differences between model outputs and human labels, as measured by the Jensen Shannon Divergence (JSD) between the human label distributions and the models' softmax distributions over the possible answers. We find that the JSDs are lower than the ones reported by Nie et al. (2020b) for the previous generation of models, and they are also better than chance distributions. However, they are still substantially higher than the JSD between two populations of humans[2], bearing interesting implications on the viablity of using an ensemble of LLM judges as a 'jury' (e.g. Verga et al., 2024). Interestingly, contrary to the findings of Nie et al. (2020b), we observe an effect of scale and model quality: JSD shows a clear decrease during training, and larger models have lower JSD than smaller models.

In sum, we find that NLI benchmarks are still useful for model development and improvement. Specifically, they are able to discriminate between models of different scale and quality, develop steadily during training, and are not completely saturated. Furthermore, as even the best models are still far away from human performance in this respect, we see promise in monitoring the development of the distributional differences between models and humans, both during and after training.

2 Related work

Before presenting our analysis, we first discuss NLI or *recognising textual entailment* (RTE) tasks (§ 2.1) and touch upon the related topic of subjectivity for NLP tasks (§ 2.2) before presenting our analysis.

2.1 RTE tasks and their results

In RTE tasks – also referred to as 'natural language inference' (NLI) tasks, models are asked to judge

whether the meaning of one sentence can be inferred from the meaning of another. Because the concept of entailment (and contradiction) are considered central to many aspects of language meaning (e.g. Bowman et al., 2015), such tasks were for some period considered an important task to determine whether one model could understand language better than another (Poliak, 2020). Over the years, many different, increasingly more difficult NLI tasks have been proposed in the literature. Included in the popular benchmarks GLUE (Wang et al., 2019b) and SuperGLUE (Wang et al., 2019a), the benchmarks MNLI (Williams et al., 2018) and RTE (Dagan et al., 2005) and RTE were used to claim SoTA by many influential models such as BERT (Devlin et al., 2019) and T5 (Raffel et al., 2023). When performance on MNLI and RTE started to saturate, several adversarial NLI benchmarks were introduced, such as ANLI (Raffel et al., 2023) and HANS (McCoy et al., 2019), on which BERT-style models performed poorly compared previous datasets. For LLMs, NLI benchmarks are rarely used. Of all big LLM releases, only GPT-3 (Brown et al., 2020) reported an NLI score, and only on one partition of ANLI. They concluded that NLI is a difficult task for general purpose LLMs. Similar trends were observed by Ohmer et al. (2024) and Weber et al. (2023) for decoder-only LLMs on various NLI tasks, in part motivating the study presented here.

2.2 Subjectivity in NLP tasks

Another line of work relevant to ours considers the behaviour of models in cases where humans disagree on the correct label for a particular sample (for an overview, see Plank, 2022). The ground truth labels for NLP benchmarks are often decided according to the majority label by human annotators. This simplifies the data annotation process, while also making the evaluation easier. However, several previous studies have noted that human disagreements in annotations for NLP datasets reflect the lack of a single ground truth label, rather than noise in the annotation process (e.g. de Marneffe et al., 2012; Plank et al., 2014; Pavlick and Kwiatkowski, 2019; Nie et al., 2020b)and that we could benefit from embracing rather than resolving it (Plank, 2022; Aroyo and Welty, 2015).

For NLI, label disagreements were captured in more detail in the dataset ChaosNLI (Nie et al., 2020b), comprising of 100 annotations for each sample for a subset of three benchmarks – MNLI,

---

SNLI, and αNLI. As human disagreements are ubiquitous yet mostly ignored in LLM evaluation, in our study, we analyse not only the model accuracies, but also the relationship between these disagreements and models' probability distributions across the different labels. A study with a similar aim is presented by Chen et al. (2024). They explore whether the softmax probability distributions solicited by a few explanations from two LLMs (Mixtral and Llama 3) can approximate human judgement distributions (HJD) on the ChaosNLI and VariErr NLI (Weber-Genzel et al., 2024) benchmarks. Similarly, Baan et al. (2024) and Lee et al. (2023) compare human and model judgement distributions on ChaosNLI, finding that models fail to capture human distributions in LLMs and model confidences. Furthermore, Baan et al. (2024) argue how the softmax probability distribution can be interpreted as both an approximation to human label distribution and confidence estimation in language models.

3 Setup

Before reporting our results, we first describe the benchmarks and models we consider in our experiments and describe our evaluation procedure.

3.1 Benchmarks

In our experiments, we consider five NLI benchmarks as briefly described below. A more elaborate description can be found in Appendix A.

SNLI   The Stanford Natural Language Inference (SNLI) is one of the first large-scale NLI datasets for NLP evaluation, sourced via Amazon Mechanical Turk. We use the dev set of the corpus, which comprises 10K examples.

MNLI   The Multi-Genre NLI (MNLI) corpus (Williams et al., 2018) was introduced as an alternative to SNLI that captures more genres and challenging examples. For MNLI too we consider the 10K examples dev set of the corpus.

HANS   The adversarial dataset Heuristic Analysis for NLI Systems (HANS, McCoy et al., 2019) is generated to contain examples that cannot be solved through heuristics like lexical overlap, and contains 30K examples.

ANLI   The second adversarial dataset we consider is Adversarial NLI, or ANLI (Nie et al., 2020a). The dataset, created with the primary aim to make SoTA models fail, is sourced iteratively with a human-in-the-loop setup. In our experiments, we consider the dev set of iteration 3, the most challenging set of the benchmark.

αNLI αNLI or abductive NLI (Bhagavatula et al., 2020) has a different setup than the previous benchmarks, focusing specifically on *abductive reasoning*. Each sample consist of a pair of observations at two consecutive times, a plausible hypothesis that explains tho two observations, and an implausible hypothesis that does not (or to a lesser extent). The model has to select the most plausible hypothesis among the two choices. The dataset contains 1500 examples.

ChaosNLI   Lastly, we consider ChaosNLI (Nie et al., 2020b), which contains 100 additional (human) labels for 1500 examples for each of SNLI, MNLI, and αNLI.

3.2 Models

For each of these benchmarks, we compute and analyse scores for two different model families: Llama (Dubey et al., 2024) and Mistral (Jiang et al., 2023, 2024). Specifically, we use Meta-Llama 3.1 {8, 70, 405}B from the Llama series, and Mistral 7B / Mixtral 8x{7, 22}B from the Mistral series. We limit our analysis to pre-trained base models and leave the analysis on post-trained / instruct models to future work.

3.3 Evaluation details

As nowadays common for pre-trained models, we consider the choice-based rather than generative evaluation setup for all the tasks (Dubey et al., 2024; Fourrier et al., 2023). In this setup, the model is presented with the few shot examples (if present) along with the question and the available choices in a multiple choice set-up, and is then asked to predict the correct letter choice. Since there are only a limited number of choices depending on the task (two or three), we append the these choices to the prompt and compute the negative log likelihood (NLL) over the letter choice. We then choose the option which has the lowest NLL as the model's prediction. The prompt templates for all tasks are detailed in Table 5. We use simplistic prompt templates for each of the tasks without much prompt tuning, exploiting the finding presented by Dubey et al. (2024) that the Llama models are usually robust to prompt texts in the choice task evaluation setup.

4 Results

We now present our results, focusing in particular on whether NLI benchmarks provide a discriminative signal for fully trained out models (§ 4.1), how their performance develops during training (§ 4.2), and what is the impact of evaluation data contamination (§ 4.3). Next, we investigate whether there is still room for improvement (§ 4.4) and how model judgements compare to human judgements for ambiguous or vague questions (§ 4.5).

4.1 Informativeness for fully trained models

To get an overall estimate of the difficulty of NLI and the extent to which it can discriminate models, we first consider the performance of fully pretrained models from the respective model series.

Accuracy across shots For each of the models, we compute results with a variable number of shots.[3] We report the results in Figure 1, along with a chance baseline and the results of a finetuned BERT model. For all models, we observe rather poor zero-shot performance for all tasks, except αNLI, confirming previously reported results by Ohmer et al. (2024), Weber et al. (2023), and Dutt et al. (2024). When more few-shot examples are added, performance starkly improves. Even with just one in-context example, the performance is significantly better than the zero shot accuracy. Adding more than three or four examples marginally improves performance and saturates around ten few-shot examples. Among the five benchmarks, the most challenging benchmark is ANLI. Although the larger models in the Llama and Mistral series far outperform the finetuned BERT baseline, they do not exceed 70% accuracy – to some extent confirming the difficulty of the benchmark reported by Brown et al. (2020).

Model discriminability For discriminability, virtually all benchmarks provide a clear gap between the smaller and larger models for both families of models. For example, for the Llama series, 405B performs the best followed by the 70B and then the 8B model. Though these three models are trained on the same amount of text tokens, performance clearly improves with scale. The exception to this pattern is αNLI, which appears to be near-saturated already at 70B with an accuracy of around 85%.

---

[3]For each number of few-shot examples, samples were chosen randomly from the dev or train set once and kept fixed throughout the experiments.

	8B		70B	
	mon_{Acc}	mon_{NLL}	mon_{Acc}	mon_{NLL}
αNLI	0.62	0.62	0.79	0.79
ANLI	-	-	0.67	0.47
HANS	0.32	0.46	0.57	0.63
MNLI	0.34	0.51	0.77	0.80
SNLI	0.05	0.38	0.64	0.65

Table 1: **Monotonicity values during the course of training.** We report monotonicity both for accuracy (mon_{Acc}) and negative log likelihood of the correct answer (mon_{NLL}).

We conclude from these results, that the benchmarks provide a useful signal to compare trained-out models, though it is unclear to what extent their performance has saturated, which we discuss in more detail in § 4.4.

4.2 Informativeness during training

Next, we investigate if NLI datasets provide a good signal during training. To this end, we pre-train Llama-3 architecture-based 8B and 70B models from scratch for 2T tokens. The details of the pre-training setup are provided in Appendix D.

Training curves In Figure 2, we show how four-shot performance develops during training. We see that for most benchmarks, the 8B and 70B model quickly start to diverge. The 70B model starts improving after 250B tokens for ANLI and αNLI; it crosses fine-tuned BERT performance after 500B tokens. On the other hand, the development of performance for the 8B model is slow, not exceeding chance accuracy for HANS, MNLI, and SNLI. From the final model performance of 8B depicted in Figure 1, we can conclude that the 8B model improves fairly late in pre-training. We did not have the budget for a full pre-training run, but longer training seems to help for NLI tasks, further supporting the claim that NLI models can provide a useful signal.

Utility for ablations A requirement for a benchmark to provide a useful signal during training is that it develops relatively monotonically during training. The plots in Figure 2 suggest that this is not the case for most of the benchmarks for the 8B model. Following Madaan et al. (2024), we quantify the benchmarks' monotonicity – defined as the rank correlation (Kendall Tau) between a monotonically increasing array and the array containing
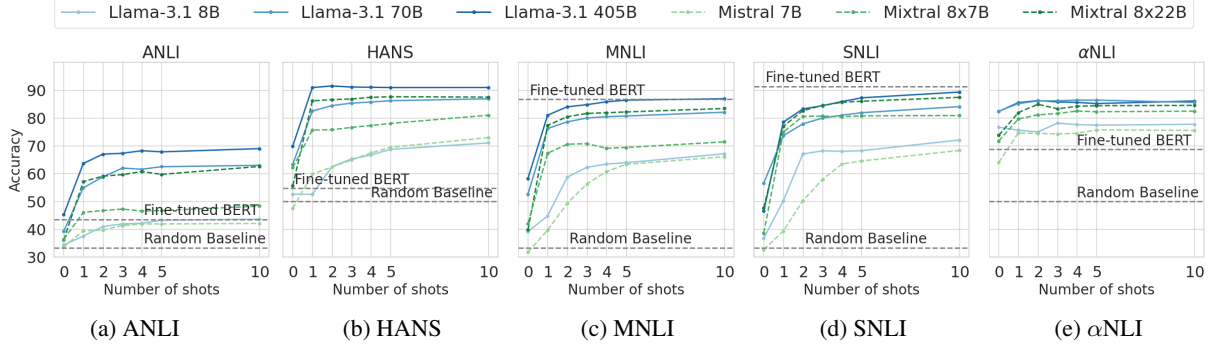
Figure 1: **Performance across shots.** We show the accuracies for six fully pre-trained models on the five NLI benchmarks. Dashed lines indicate random and finetuned-BERT baselines.
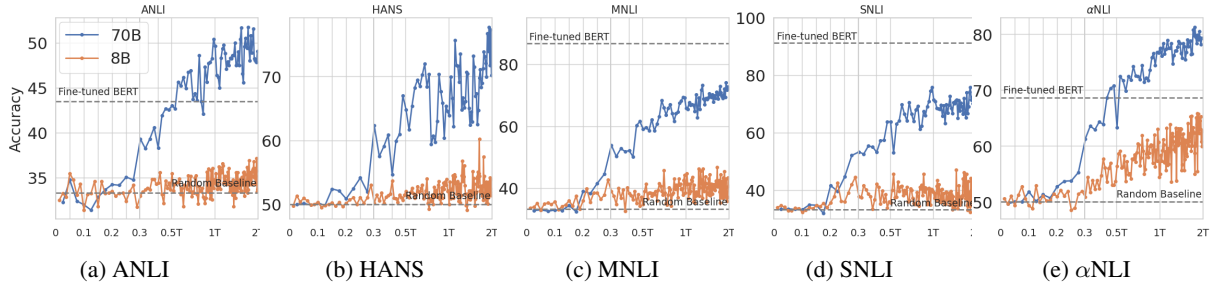

Figure 2: **Performance during pre-training**. We show how accuracy for the five benchmarks develops during pre-training for two Llama-3 style models.

the benchmark scores during training – on both a discrete (accuracy) and continuous metric (NLL). In Table 1, we can see that, despite the benchmarks' moderate sizes, monotonicity is low for accuracy as well as NLL, suggesting that the benchmark may not be suitable for monitoring performance on closely-spaced checkpoints at this scale.

4.3   Contamination analysis

Now that we have concluded that NLI benchmarks provide a signal both for final models and during pre-training, we analyse the extent to which scores may be driven by evaluation data contamination. Using the methodology of Singh et al. (2024) and Dubey et al. (2024), we analyse contamination by assigning a score to each evaluation sample which represents the percentage of tokens in the sample that is part of an 8-gram occurring in the pre-training datamix. Following Singh et al. (2024), we select contamination thresholds using ConTAM, considering for each threshold the estimated performance gain (EPG), defined as the difference in performance on the full evaluation set and the 'uncontaminated' subset.[4] Since EPG is

model-dependent, we use the 8B and 70B models we trained from scratch to observe the effect of contamination.

In Figure 3, we plot the EPG as a function of the percentage of the evaluation data that is marked contaminated according to a given contamination threshold. We can see that there is virtually no score inflation as a consequence of contamination for the 8 and 70B model that we considered. For HANS, this is because little to no contaminated samples are found through 8-gram overlap. For MNLI, SNLI, and αNLI, instead, large parts of the dataset are marked contaminated at low thresholds, but there is no performance impact. We suspect that this is because the *premises* of the samples are sourced from publicly available datasets, and their presence per se is thus not indicitive of contamination. Likely for similar reasons, almost all of the ANLI samples contain at least one overlapping 8-gram, resulting in very high percentages of detected contamination at low thresholds. At the lowest threshold, also the EPG shoots up. However, the erratic behaviour as examples are gradually added to the clean set suggests that the EPG

---

[4]A threshold of 0 implies that all examples with non-zero contamination scores to be marked as contaminated, as we

increase the threshold, examples are moved to the 'uncontaminated' subset.

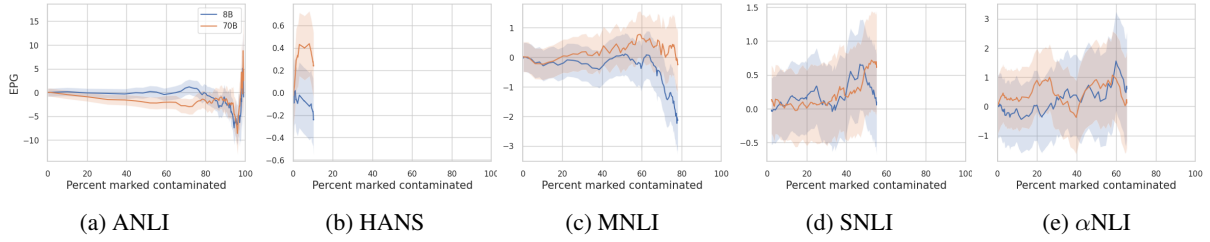(a) ANLI   (b) HANS   (c) MNLI   (d) SNLI   (e) αNLI

Figure 3: **Contamination results.** We show the EPG vs the percent of the evaluation dataset marked as contaminated according to different thresholds.

observed at those threshold is likely an artefact of the small size of the clean partition, and this result is likely not indicative of true performance gain. Thus, we believe that contamination does not play a participatory role in high performances for NLI benchmarks.

4.4 Dataset saturation

Having confirmed that the NLI benchmarks under scrutiny provide discrimination between LLMs of different sizes and are not affected much by contamination, we now turn to the question of saturation: the results show that the benchmarks would have been useful so far, but how about the future? As already pointed out before, the benchmark that still has the clearest room for improvement is adversarial benchmark ANLI, with performances not exceeding 70% even for the largest models. For the other benchmarks, performances are substantially higher, and it is unclear to what extent the benchmarks may suffer saturation.

Error types To address this question, we first conduct a manual error analysis on the examples that the largest models assign an incorrect label. Specifically, we focus on MNLI, one of the datasets with the highest scores and analyse 40 predictions of Llama 3.1 405B model. In Table 2, we show examples, along with distribution of the 100 human annotations for that sample from the previously mentioned dataset ChaosNLI. For all the examples we investigated, we found that there was at least some degree of interpretation or ambiguity in assigning a label to the premise-hypothesis pair. For example, in the first row of Table 2, whether the correct label is contradiction or negation would depend on a not-specified sentiment towards Cuba. In fact, 31 out of 100 human annotators would agree with the model's judgement, despite it being labeled as "incorrect". Also in the second row of Table 2, we see an example where human opinions diverge on what the correct label is. In this

case, the model's prediction matches the majority label found by Nie et al. (2020b), but the sample is marked incorrect because the MNLI label is different. It is worth pointing out, that the same is sometimes true for examples that are marked *correct*. Consider, for instance, the last row in Table 2, where the model prediction matches the gold label in the MNLI dataset, but not the majority label collected by ChaosNLI. In sum, it appears that for the best model, most of the 'mistakes' in MNLI are cases in which humans may not agree on the correct label.

Majority accuracy Next, we study this phenomenon more quantitatively, again utilising the ChaosNLI dataset which contains 100 human annotations for over 1500 samples each for MNLI, SNLI, and αNLI. First, we consider how model accuracies change when we replace the original labels with the majority label of the ChaosNLI dataset. This alters 32%, 25%, and 11% of the labels of MNLI, SNLI, and αNLI, respectively. In Table 3, we can see that the results differ per model and benchmark. The largest effect is observed for MNLI, where for some models, there's an increase of more than 10% in performance and the average accuracy across models is more than five points higher on the 'corrected' datasets. For the other two benchmarks, the results are more mixed, with little to no difference on average. Only for the largest Llama-3.1 model (405B), the majority accuracy is systematically higher than the original accuracy, suggesting it may have honed in more on the majority label. Interestingly, the MNLI and SNLI subsets of ChaosNLI appear substantially more difficult than the average dataset; even the Llama-3.1 405B model stays below 70% for both these subsets, suggesting that there is room for improvement.

Accuracy versus entropy Next, we consider the distribution of accuracy with entropy for the three

Example	Label distribution	Gold	Prediction
Premise: of course you could annex Cuba but they wouldn't like that a bit **Hypothesis**: Annexing Cuba is a great idea.	e: 0, n: 31, **c: 69**	c	n
Premise: I thought working on Liddy's campaign would be better than working on Bob's. **Hypothesis**: I thought I would like working on Liddy's campaign the best.	**e: 66**, n: 32, c: 2	n	e
Premise: Sorry but that's how it is. **Hypothesis**: This is how things are and there are no apologies about it.	**e: 48**, c: 40, n: 12	c	c

Table 2: **Examples from ChaosNLI with different label distributions.** We show different samples from ChaosNLI with different annotator distributions. The first column shows the example; the second the distribution of human labels across the 100 collected annotations (the majority label is marked bold), the third column indicates the label that the example had in the original dataset and the fourth column the Llama 3.1 405B prediction.

Model		α**NLI**		**MNLI**		**SNLI**	
		Og.	*Maj.*	*Og.*	*Maj.*	*Og.*	*Maj.*
Llama-3.1	8B	77.55	**78.00**	49.47	**50.97**	55.28	55.48
Llama-3.1	70B	86.36	**87.21**	57.66	**67.54**	**60.44**	58.52
Llama-3.1	405B	85.51	**86.10**	64.04	**69.67**	64.60	**67.31**
Mistral	7B	74.41	**75.78**	49.97	**53.22**	**49.47**	48.15
Mixtral	8x7B	**82.44**	81.59	**54.03**	51.53	63.14	**64.27**
Mixtral	8x22B	**84.14**	83.68	60.23	**67.04**	64.86	**67.83**
Average		76.01	76.21	50.86	55.93	54.90	54.35

Table 3: **Majority accuracy.** For each model and dataset, we show the accuracy on the ChaosNLI subsets of αNLI, MNLI, and SNLI. The original accuracy (Og.) represents the accuracy as computed according to the original labels of the respective datasets; the majority accuracy (Maj.) expresses accuracy according to the majority label.

benchmarks in Figure 4a. We only show the smallest (8B) and the largest (405B) model in the Llama series across the three benchmarks. We observe that for the largest model, there is a clear decreasing trend in accuracy with increasing entropy for all three benchmarks, whereas for the 8B model, the trend is not as prominent, and some entropy bins have similar accuracies. In Figure 4a, we can furthermore see that all models perform 'better' on samples where the entropy of the labels is low. For the larger models, this effect is larger on samples where humans have high disagreement (and the majority label is thus in a way more representative of the average human judgements), their accuracies are often near maximal, and they drop as human judgements become more dispersed. For the entropy vs accuracy distributions for other models, we refer the reader to Appendix B.

Implications for saturation What the implication of these results is for the question of saturation of these benchmarks is, like the samples of the benchmark themselves, open to interpretation. On the one hand, it appears that for several cases, the model predictions do not align with the human ma-

jority label, suggesting that the performance may still improve as models continue training. On the other, the results put into question whether strong alignment with the majority label is in fact what we should strive for: if humans have low agreement on their judgements, should the desired behaviour of a model be to side with the majority? In the next section, we approach this question by considering the distribution of model outputs, rather than only the highest probability label.

4.5 Model versus human distributions

The ChaosNLI dataset does not only help us estimate the adequacy of the original label sets of the respective datasets, but also reveals model behaviour in scenarios without a single correct answer. In a time where LLMs are deployed for many users with potentially different preferences, this question has become very practically relevant. To do so, we consider how the probability distribution of the models over the three possible labels *Entailment*, *Neutral*, and *Contradiction*) compares with the label distributions observed in the human annotations. Following Nie et al. (2020b), we consider Jensen-Shannon Divergence (JSD Menéndez

9235
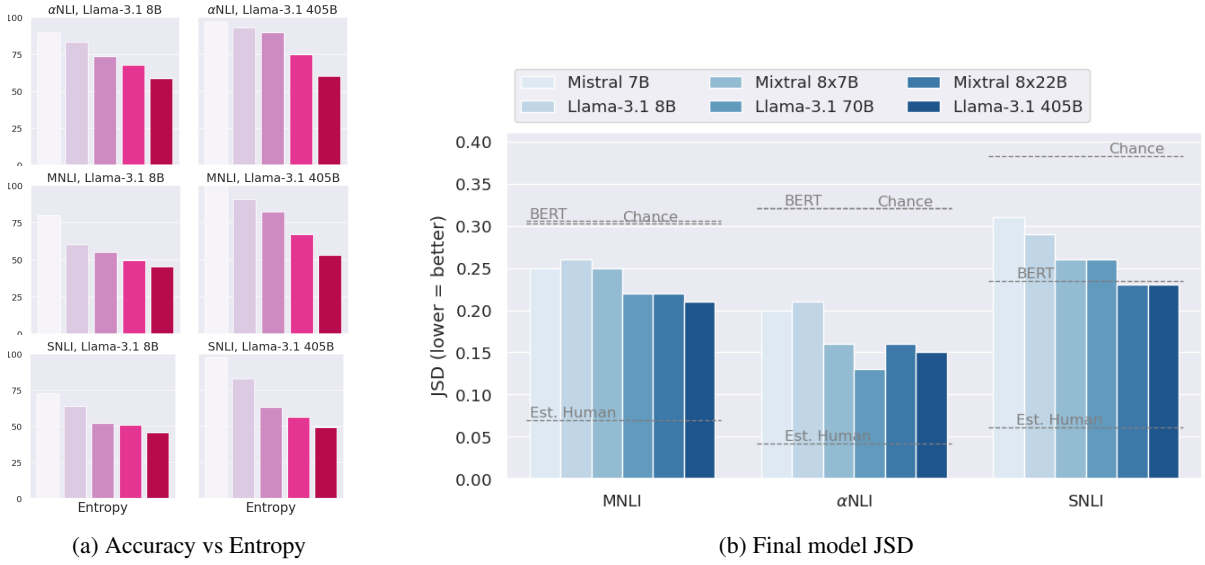
(a) Accuracy vs Entropy

(b) Final model JSD

Figure 4: **Accuracy vs entropy and final model JSDs** a) Accuracy vs entropy for Llama8B and Llama 405B. We show how the accuracy of Llama 8B and Llama 405B changes as the entropy of the human label distributions increases. Accuracy-vs-entropy plots for all other models can be found in Figure 6. b) Final model JSDs for each of the benchmarks in ChaosNLI. JSDs are substantially lower than chance and BERT JSDs, but substantially higher than JSDs between humans.
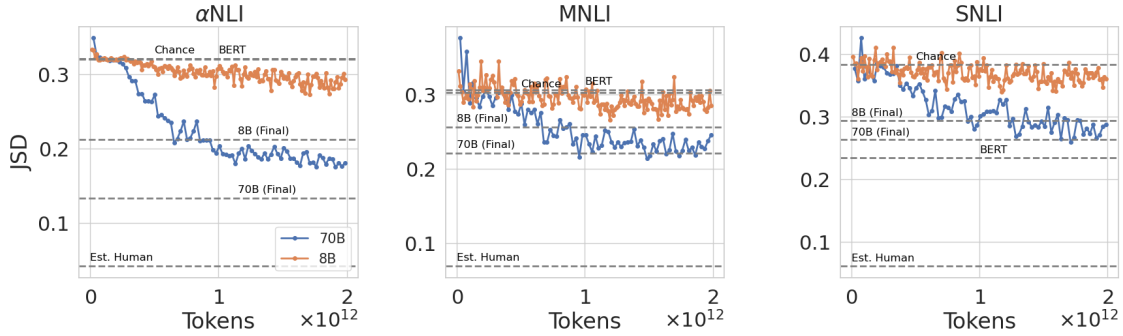

Figure 5: **Development of JSD during training.** We show how the JSD of our trained-from-scratch 8B and 70B model develops during training.

et al., 1997) to measure the distance between the two distributions. Contrary to KL divergence, JSD is symmetric and bound between 0 and 1, making it more interpretable for our specific case.

In Figure 4b, we can see that, for MNLI, all models have probability distributions more similar to the human distributions than chance, but substantially different than humans have among each other.[5] This pattern is constant across datasets. Interestingly, larger models have lower divergence, contradicting the finding of Nie et al. (2020b) that better or larger models do not have more similar distributions. Yet, the effect of scale is further confirmed in Figure 5: though even trained-out models

are far away from a low JSD, the JSD of the model softmax distributions and the human label distributions decreases steadily during training. This is an exciting finding, since it suggests that measuring similarity with human distributions could be an interesting venue to explore during training.

5 Conclusion

In this work, we revisit natural language inference (NLI) benchmarks and investigate if they may still play a role in LLM evaluation, both during and after pre-training. We consider five different NLI benchmarks – αNLI, ANLI, HANS, and MNLI – and evaluate them across six different models of two model families: Llama 3.1 8B, 70B, and 405B, and Mistral 7B, 8x7B, and 8x22B. Furthermore,

---

[5]The human estimate was computed by Nie et al. (2020b), on an independent sample of human labels on the same data.

we consider how the benchmark behave during the training of two Llama-style 8B and 70B models. We find that, with the exception of αNLI, all benchmarks are able to discriminate between models of different qualities, and in particular ANLI is challenging even for the largest models. Furthermore, we find next to no effect of contamination for the benchmarks. Considering the benchmark ChaosNLI (Nie et al., 2020b), containing 100 human annotations for over 4500 samples of three of the benchmarks we consider, we also find that the differences between human label distributions and model label distributions – as measured with Jensen Shannon Divergence (JSD) – has decreased for the new generation of models. However, they are still substantially higher than the distributional difference between two populations of humans. Interestingly, contrary to the findings of Nie et al. (2020b), we observe a clear effect of scale. JSD shows a steady decrease during training, and larger models have lower JSD than smaller models, making it a potentially interesting quality to consider for model development.

6   Limitations

One of the limitation of our work is that all our analysis is limited to pre-trained models. Post-training with supervised fine-tuning and RLHF might change the behaviour of models on NLI benchmarks. But, we do not believe this to be a major limitation as Dubey et al. (2024) show that it's possible to improve the performance of models during the post-training phase, rather than lose it because of instruction fine-tuning and safety. Therefore, we believe that NLI benchmarks would be informative for post-trained models as well. Another limitation is that we only consider the choice-based setup for evaluation. Using a generative setup might change model behaviour and correspondingly, the signal provided by NLI benchmarks. We leave this to future work.

References

Lora Aroyo and Chris Welty. 2015. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Mag.*, 36:15–24.

Joris Baan, Wilker Aziz, Barbara Plank, and Raquel Fernandez. 2022. Stop measuring calibration when humans disagree. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1892–1915, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Joris Baan, Raquel Fernández, Barbara Plank, and Wilker Aziz. 2024. Interpreting predictive probabilities: Model confidence or human label variation? In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 268–277, St. Julian's, Malta. Association for Computational Linguistics.

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Wen tau Yih, and Yejin Choi. 2020. Abductive commonsense reasoning. In *International Conference on Learning Representations*.

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. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Beiduo Chen, Xinpeng Wang, Siyao Peng, Robert Litschko, Anna Korhonen, and Barbara Plank. 2024. "seeing the big through the small": Can llms approximate human judgment distributions on nli from a few explanations? *Preprint*, arXiv:2406.17600.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The pascal recognising textual entailment challenge. In *Machine learning challenges workshop*, pages 177–190. Springer.

Marie-Catherine de Marneffe, Christopher D. Manning, and Christopher Potts. 2012. Did it happen? the pragmatic complexity of veridicality assessment. *Computational Linguistics*, 38(2):301–333.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Ritam Dutt, Sagnik Ray Choudhury, Varun Venkat Rao, Carolyn Rose, and V.G.Vinod Vydiswaran. 2024. Investigating the generalizability of pretrained language models across multiple dimensions: A case study of NLI and MRC. In *Proceedings of the 2nd GenBench Workshop on Generalisation (Benchmarking) in NLP*, pages 165–182, Miami, Florida, USA. Association for Computational Linguistics.

Clémentine Fourrier, Nathan Habib, Julien Launay, and Thomas Wolf. 2023. What's going on with the open llm leaderboard?

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Noah Lee, Na Min An, and James Thorne. 2023. Can large language models capture dissenting human voices? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4569–4585, Singapore. Association for Computational Linguistics.

Lovish Madaan, Aaditya K. Singh, Rylan Schaeffer, Andrew Poulton, Sanmi Koyejo, Pontus Stenetorp, Sharan Narang, and Dieuwke Hupkes. 2024. Quantifying variance in evaluation benchmarks. *Preprint*, arXiv:2406.10229.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

María Luisa Menéndez, JA Pardo, L Pardo, and MC Pardo. 1997. The jensen-shannon divergence. *Journal of the Franklin Institute*, 334(2):307–318.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020a. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online. Association for Computational Linguistics.

Yixin Nie, Xiang Zhou, and Mohit Bansal. 2020b. What can we learn from collective human opinions on natural language inference data? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9131–9143, Online. Association for Computational Linguistics.

Xenia Ohmer, Elia Bruni, and Dieuwke Hupkes. 2023. Separating form and meaning: Using self-consistency to quantify task understanding across multiple senses. In *Proceedings of the Third Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 258–276, Singapore. Association for Computational Linguistics.

Xenia Ohmer, Elia Bruni, and Dieuwke Hupkes. 2024. From Form(s) to Meaning: Probing the Semantic Depths of Language Models Using Multisense Consistency. *Computational Linguistics*, pages 1–51.

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014. Learning part-of-speech taggers with inter-annotator agreement loss. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 742–751, Gothenburg, Sweden. Association for Computational Linguistics.

Adam Poliak. 2020. A survey on recognizing textual entailment as an NLP evaluation. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, pages 92–109, Online. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2023. Exploring the limits of transfer learning with a unified text-to-text transformer. *Preprint*, arXiv:1910.10683.

Aaditya K. Singh, Muhammed Yusuf Kocyigit, Andrew Poulton, David Esiobu, Maria Lomeli, Gergely Szilvasy, and Dieuwke Hupkes. 2024. Evaluation data contamination in llms: how do we measure it and (when) does it matter? *Preprint*, arXiv:2411.03923.

Pat Verga, Sebastian Hofstätter, Sophia Althammer, Yixuan Su, Aleksandra Piktus, Arkady Arkhangorodsky, Minjie Xu, Naomi White, and Patrick S. H. Lewis. 2024. Replacing judges with juries: Evaluating LLM generations with a panel of diverse models. *CoRR*, abs/2404.18796.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019a. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019b.

Benchmark	# Samples	# labels
ANLI	1200	3
HANS	30000	2
MNLI	9815	3
SNLI	9842	3
αNLI	1532	2

Table 4: Dataset details

GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Lucas Weber, Elia Bruni, and Dieuwke Hupkes. 2023. Mind the instructions: a holistic evaluation of consistency and interactions in prompt-based learning. In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 294–313, Singapore. Association for Computational Linguistics.

Leon Weber-Genzel, Siyao Peng, Marie-Catherine De Marneffe, and Barbara Plank. 2024. VariErr NLI: Separating annotation error from human label variation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2256–2269, Bangkok, Thailand. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. 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 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

A Benchmarks

Below, we provide a more elaborate description of (construction of) the benchmarks we consider in our work. A summary of their statistics can be found in Table 4.

A.1 SNLI

Introduced by Bowman et al. (2015), the Stanford Natural Language Inference (SNLI) dataset was one of the first large-scale NLI dataset for NLP evaluation. The dataset comprises of 570K human-authored English sentence pairs, sourced by asking Amazon Mechanical Turk workers to supply hypotheses for the three labels available in the dataset – entailment, neutral and contradiction – given a premise comprised by an image caption drawn from a pre-existing corpus. For 57K of the resulting samples were then labeled by four additional annotators. In this work, we consider the 10K development set of the corpus. Like the original paper, we exclude samples with no gold label because there was no label that the annotators agreed on.

A.2 MNLI

The Multi-Genre NLI (MNLI) corpus (Williams et al., 2018) was introduced as an alternative to SNLI that captures more genres and more challenging examples, representing both written and spoken speech in a range of different styles, degrees, formalities, and topics. The data collection procedure of the corpus is similar to the SNLI procedure both in terms of sourcing and validation. Unlike SNLI, the MNLI premise sentences are derived from nine different sources, aiming to represent the full range of American English rather than a single image captioning corpus. As for SNLI, we consider the validation corpus of the dataset and exclude samples that have no gold label.

A.3 HANS

Deviating from previous datasets, the adversarial dataset Heuristic Analysis for NLI Systems (HANS, McCoy et al., 2019) is not crowd-sourced, but synthetically generated using templates. Specifically, the templates are designed to generate examples that can not be solved through heuristics such as lexical, subsequence, or constituent overlap. At the time of proposal, none of the SOTA models were able to solve such examples.

A.4 ANLI

The second adversarial dataset we consider is Adversarial NLI (or ANLI, Nie et al., 2020a). The dataset, created with the primary aim to make SOTA models fail, is sourced iteratively in a human-in-the-loop setup. Given a premise and a target label, annotators are asked to propose hypotheses that may fool models. The produced samples are then tested on a model, and the examples that do indeed receive an incorrect label are re-validated by one or more human validators. The dataset consists of three sets of increasingly challenging examples, where in each round more powerful models are considered that are trained on examples from the previous round. The third round furthermore contains a set of more diverse premises. For our experiments, we are using the dev set of round 3, the most challenging set of the benchmark.

A.5 αNLI

Differing in setup from the previously described benchmarks, αNLI or abductive NLI (Bhagavatula et al., 2020) focuses on *abductive reasoning –* which they describe as the inference of the most plausible explanation for an incomplete observation. The samples in αNLI consist of a pair of observations at two consecutive times, and a plausible hypothesis that explains tho two observations, and an implausible hypothesis that does not (or to a lesser extent). The task is to select the most plausible hypothesis. To construct the data Bhagavatula et al. (2020) first draw observation pairs from a stories dataset and then ask crowd-sources to generate plausible and implausible hypotheses. For each observation pair, multiple plausible and implausible hypotheses are crowd-sourced, and adversarial filtering is applied to retain one challenging pair of hypotheses. We use the development set of the corpus for our experiments.

B   Entropy vs accuracy plots

In addition to Figure 4a highlighting the results on the Llama-3.1 8B and 405B models, we also show the accuracy vs entropy plots of all other models we use for our analysis: Llama-3.1 70B, Mistral 7B, Mixtral 8x7B, and Mixtral 8x22B in Figure 6. The Mistral family of models also show a similar trend where larger models have lower accuracies for higher entropy buckets.

C   Prompt templates

The prompt templates used for each task are presented in Table 5.

D   Experimental Details

For the 8B and 70B models we pre-train from scratch, we use our custom pre-training datamix, a mix of data available from publicly available sources including web data, code, and reasoning datasets. For pre-training hyperparameters, we use similar settings as reported in Dubey et al. (2024). We use a batch size of 4M tokens and pre-train the models for 500,000 steps, resulting in a total of 2T token training. We use 512 GPUs for a single pre-training run of both models.

For running the evaluations, we use 16 GPUs for each model comprising all five NLI benchmarks and different shot settings in a single job. A single evaluation job takes on average takes around 55 mins for the five benchmarks.

Figure 6: **Accuracy versus entropy.** We show how the accuracy of Llama-3.1 70B and the Mistral series of models changes as the entropy of the human label distributions increases.

Benchmark	Prompt Template
MNLI, SNLI, ANLI	``` {% for x in few_shot -%} Premise: {{ x["premise"] }} Hypothesis: {{ x["hypothesis"] }} A. Entailment B. Neutral C. Contradiction Answer: {{ x["answer"] }} {% endfor -%} Premise: {{ premise }} Hypothesis: {{ hypothesis }} A. Entailment B. Neutral C. Contradiction Answer: {{ choice_text }} ```
AbductiveNLI	``` {% for x in few_shot -%} Observation 1: {{ x["obs1"] }} Observation 2: {{ x["obs2"] }} A. {{ x["choices"]["A"] }} B. {{ x["choices"]["B"] }} Answer: {{ x["answer"] }} {% endfor -%} Observation 1: {{ obs1 }} Observation 2: {{ obs2 }} A. {{ choices["A"] }} B. {{ choices["B"] }} Answer: {{ choice_text }} ```
HansNLI	``` {% for x in few_shot -%} Premise: {{ x["premise"] }}. Hypothesis: {{ x["hypothesis"] }}. A. Entailment B. Non-Entailment Answer: {{ x["answer"] }} {% endfor -%} Premise: {{ premise }}. Hypothesis: {{ hypothesis }}. A. Entailment B. Non-Entailment Answer: {{ choice_text }} ```

Table 5: Prompt Templates for each task