Separating form and meaning: Using self-consistency to quantify task understanding across multiple senses

Xenia Ohmer$^{\Omega}$ and **Elia Bruni**$^{\Omega *}$ and **Dieuwke Hupkes**$^{\infty *}$
$^{\Omega}$Osnabrück University $^{\infty}$ FAIR
{xenia.ohmer, elia.bruni}@uni-osnabrueck.de
dieuwkehupkes@meta.com

Abstract

At the staggering pace with which the capabilities of large language models (LLMs) are increasing, creating future-proof evaluation sets to assess their understanding becomes more and more challenging. In this paper, we propose a novel paradigm for evaluating LLMs which leverages the idea that correct world understanding should be consistent across different (Fregean) senses of the same meaning. Accordingly, we measure understanding not in terms of correctness but by evaluating consistency across multiple senses that are generated by the model itself. We showcase our approach by instantiating a test where the different senses are different languages, hence using multilingual self-consistency as a litmus test for the model's understanding and simultaneously addressing the important topic of multilinguality. Taking one of the latest versions of ChatGPT as our object of study, we evaluate multilingual consistency for two different tasks across three different languages. We show that its multilingual consistency is still lacking, and that its task and world understanding are thus not language-independent. As our approach does not require any static evaluation corpora in languages other than English, it can easily and cheaply be extended to different languages and tasks and could become an integral part of future benchmarking efforts.

1 Introduction

The staggering pace at which the capabilities of large language models (LLMs) have increased in the recent past comes with many questions related to what kind of progress we are making on the road towards true machine intelligence and human-level understanding. To assess such progress, practitioners often rely on benchmarks that measure natural language understanding (e.g. Williams et al., 2018; Nie et al., 2020), commonsense reasoning (e.g. Sap et al., 2019; Bisk et al., 2020), or probe for factual knowledge (e.g. Hendrycks et al., 2021), among other things. The extent to which such benchmarks can be used to assess whether LLMs "understand" language is widely debated (e.g. Mitchell and Krakauer, 2023; Raji et al., 2021). Often mentioned concerns in this context are that LLMs may learn specific lexical patterns rather than general principles (e.g. Ray Choudhury et al., 2022) and, relatedly, that benchmark scores may confuse competence in *form* with competence in *meaning* (e.g. Heineman, 2023). In support of these concerns, LLMs have been found to bypass certain tasks by relying on memorised information from the training data (McKenna et al., 2023). More recently, the enormous amount of data that models are trained on and the fact that this data is often not publically accessible have further increased the difficulty of assessing whether benchmarks really quantify what they are meant to quantify. A benchmark always makes assumptions about what a model has seen in its training phase, and, given the rapid changes on that front, it is difficult to design challenging benchmarks that remain informative past training rounds of new models. In addition, novel evaluation data may leak into the training data of newly trained models[1] – which even the most future-proofed benchmarks may not withstand.

In this paper, we propose a novel approach to evaluate models' task or world understanding that aims to create some separation between form and meaning in benchmarking and simultaneously mitigates the challenging evaluation-contamination loop. Our method is based on the idea that language is used to describe or act in the world (Wittgenstein, 1953) and that this world functions as an anchor for diverse linguistic forms. Having a genuine understanding of the world thus implies consistency among different linguistic expressions that pertain to the same entities within the world. To give an

[1] E.g. portions of the BIG-Bench data (Srivastava et al., 2022) were inadvertently added to the GPT4 training corpus (OpenAI, 2023, footnote 5).

Proceedings of the Third Workshop on Natural Language Generation, Evaluation, and Metrics (GEM), pages 258–276
December 6, 2023 ©2023 Association for Computational Linguistics

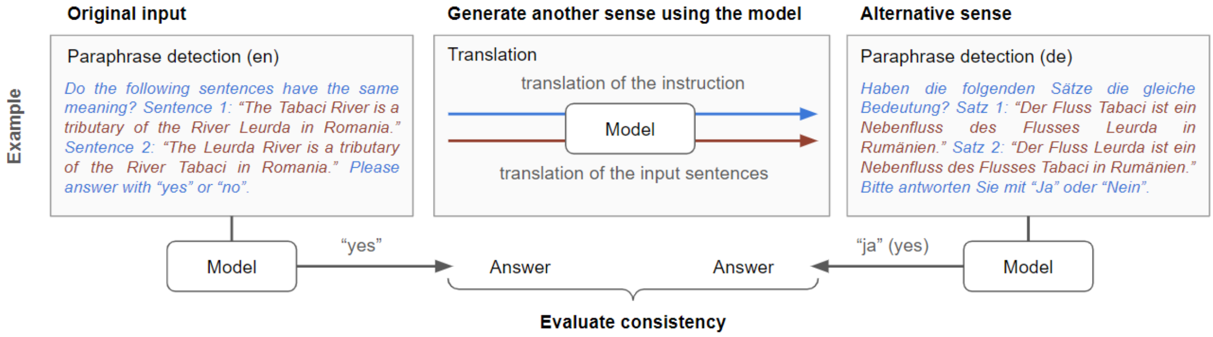
Original input
Generate another sense using the model
Alternative sense
Example
Paraphrase detection (en)
Do the following sentences have the same meaning? Sentence 1: "The Tabaci River is a tributary of the River Leurda in Romania." Sentence 2: "The Leurda River is a tributary of the River Tabaci in Romania." Please answer with "yes" or "no".
Translation
translation of the instruction
Model
translation of the input sentences
Paraphrase detection (de)
Haben die folgenden Sätze die gleiche Bedeutung? Satz 1: "Der Fluss Tabaci ist ein Nebenfluss des Flusses Leurda in Rumänien." Satz 2: "Der Fluss Leurda ist ein Nebenfluss des Flusses Tabaci in Rumänien." Bitte antworten Sie mit "Ja" oder "Nein".
Model
"yes"
Answer
Answer
"ja" (yes)
Model
Evaluate consistency

Figure 1: Illustration of the basic mechanism of our paradigm: We use the model to generate other senses of the original input. The model's answers on the original input and the alternative sense are used to evaluate its consistency. In this example, the model is presented with the task of paraphrase detection in English (sentences taken from PAWS-X) and generates another sense by translating from English to German.

example, if you ask a colleague who is fluent in both French and English if a particular statement is true, you expect their answer to be invariant to the language (French or English) in which you ask this question. We leverage this intuition to investigate whether models have a consistent world model across different senses (in the case above: languages) and, consequently, a consistent understanding of the tasks that they are asked to execute. Loosely inspired by Frege (1892), we take different senses to be different modes of presentation or notational variants. Crucially, rather than generating different senses ourselves, we ask the model to create different versions of the same question. This ensures that potential inconsistencies are really due to model-internal inconsistencies rather than misinterpretations of the question. Additionally, the method is protected from data contamination: as the different senses are regenerated for every evaluation, they cannot leak into new training data. Lastly, it can cheaply and easily be applied to already available benchmarks and therefore reduces the burden on data generation.

Our approach can be applied to a number of different senses. Here, we showcase it focusing on the multilingual case described in the example above, by asking whether models are consistent across different languages (see Figure 1). In essence, we are thus using multilingual self-consistency as a litmus test for their understanding, simultaneously addressing the important topic of multilinguality. Taking one of the latest SOTA versions of ChatGPT[2] as our object of study, we evaluate multilingual understanding for two different tasks (paraphrase detection and natural language inference) across three different languages (Chinese, German and English). It turns out that the model is inconsistent across all languages and tasks, despite being able to perform the tasks reasonably well in English and generating high-quality translations. Taken together, our analyses provide strong evidence that the model's task understanding is modulated by the representational form of the task.

In sum, we make the following contributions:

i) We introduce multisense consistency as a novel, cheap, and data-contamination-proof evaluation paradigm for LLMs;

ii) We showcase this paradigm by implementing a specific version that utilises *multilinguality* to create different senses;

iii) Using this implementation, we evaluate ChatGPT to illustrate that multilingual consistency of SOTA LLMs is still lacking;

iv) With a range of ablation experiments (see Figure 2), we demonstrate that the observed inconsistencies in fact arise from a language-dependent task understanding (rather than an inability to translate or perform the task).

With our work, we hope to not only present an interesting set of empirical results on multilingual consistency but also propose a novel, easily applicable method to generate many more challenging evaluation tests. Our framework targets models that can follow instructions to generate alternative senses and are able to generate these senses based on these instructions. Thus, with the growing popularity and capabilities of chat-models and instruction-tuned models, such as GPT-4 (OpenAI,

[2] https://openai.com/blog/chatgpt

2023) or Llama-2 (Touvron et al., 2023), our framework is becoming increasingly relevant.[3]

2 Related work

Existing benchmarks for *evaluating language understanding* in LLMs form the foundation for our work. The main idea of our paradigm is to evaluate LLMs in terms of their consistency across different senses of these benchmarks and is therefore related to other work on *self-consistency in LLMs*. In creating multiple senses through translation, there is also a close connection between our execution of this paradigm and *multilingual evaluation*. Appendix A provides a GenBench eval card (Hupkes et al., 2023) that classifies our work in the context of generalisation research.

Evaluating language understanding. A wide range of benchmark tasks has been developed to evaluate specific aspects of natural language understanding in LLMs. To evaluate *general* language understanding across diverse tasks, genres, and datasets, several of these tasks have been combined into multi-task benchmarks, such as GLUE (Wang et al., 2018) or SuperGLUE (Wang et al., 2019a). These benchmarks focus on English and evaluate, among others, paraphrase identification (e.g. PAWS; Zhang et al., 2019), natural language inference (e.g. MNLI; Williams et al., 2018), and commonsense reasoning (e.g. COPA; Roemmele et al., 2011). In response to the rapid improvements of LLMs on these benchmarks other multi-task benchmarks have been developed. MMLU, for example, assesses world knowledge and problem-solving ability across a large number of subjects, covering STEM, humanities, social sciences, and more (Hendrycks et al., 2021). While our paradigm also makes an effort to find more appropriate evaluation methods, it not only assesses performance but also evaluates the model's ability to consistently solve a task across multiple languages, thereby providing insights into its ability to abstract from specific representational forms.

Self-consistency in LLMs. Various studies have shown that inconsistencies are common in LLMs (and suggested methods for improving consistency, which is not our focus). These studies are mostly concerned with self-consistency in natural language inference (NLI) (e.g. Minervini and Riedel, 2018; Wang et al., 2019b; Li et al., 2019; Hosseini et al., 2021) and question answering (e.g. Kassner and Schütze, 2020; Alberti et al., 2019; Mitchell et al., 2022; Chen et al., 2021; Elazar et al., 2021; Kassner et al., 2021; Asai and Hajishirzi, 2020; Hosseini et al., 2021). For example, Kassner et al. (2021) created a dataset to measure a model's consistency by evaluating its responses to sentence pairs that are subject to certain constraints (e.g. if *X is a dog* is true, *X has a tail* must also be true). More similar to our work, Elazar et al. (2021) studied whether factual knowledge in masked language models is invariant to paraphrasing. To this end, they created PARAREL, a dataset containing cloze-style English paraphrases. In these two examples, consistency is either evaluated against a network of logical relationships between beliefs or by generating different forms of the same meaning through paraphrasing. BECEL (Jang et al., 2022) is a benchmark for evaluating these two types of consistency (logical and semantic) across various tasks. This benchmark has recently been used to evaluate ChatGPT, showing that it is more consistent for negations than other LLMs, but still likely to generate different responses to paraphrases of the same meaning (Jang and Lukasiewicz, 2023). Unlike previous work – except (Jang and Lukasiewicz, 2023) – we focus on true *self*-consistency: Different forms of the same meaning are generated by the model itself, rather than externally.

Multilingual evaluation. The development of cross- and multi-lingual LLMs has spurred interest in multilingual evaluation beyond translation. Several multilingual versions of benchmark tasks have been generated, such as PAWS-X (Yang et al., 2019), XCOPA (Ponti et al., 2020), and XNLI (Conneau et al., 2018) – usually through expert translations from the original task (for a more expansive overview, we refer to Hupkes et al., 2023, Appendix D). In addition, multilingual tasks have been combined to form multilingual multitask benchmarks, including XTREME (Hu et al., 2020), XTREME-R (Ruder et al., 2021), and XGLUE (Liang et al., 2020). All of these benchmarks reveal language-dependent differences in performance for current multilingual LLMs. Our approach is different in that we aim to evaluate self-consistency by detecting language-dependent changes in model responses, relying on the model's own translations instead of external translations.

[3] Our code is available at https://github.com/XeniaOhmer/multisense_consistency.

3 Methods

We now proceed with describing the model (§ 3.1) and the benchmark data (§ 3.2) we use for our experiments, as well as the procedure we use for extracting translations from the model (§ 3.3).

3.1 Model and hyperparameters

We showcase our paradigm using GPT-3.5-TURBO-0301. We use the default parameters but set the temperature to 0.25. We found a low temperature to yield model responses that more closely match the template answers for benchmarking, as well as model translations that better capture the meaning of the source sentences. In addition, we set the maximal number of generated tokens to 256 for benchmarking and 2048 for translation.

3.2 Benchmarking

Tasks and languages. We evaluate understanding using the multilingual benchmarks PAWS-X and XNLI (test splits). While our paradigm does not require parallel multilingual datasets, we use them here to evaluate translation quality, compare translations between two languages in both directions, and analyse differences that arise from using model-internal instead of model-external translations. PAWS-X is an adversarial paraphrase identification task, consisting of sentence pairs created by word-swapping, resulting in negative pairs that have clearly distinct meanings, but a high lexical overlap (see, for instance, the example in Figure 1). XNLI, on the other hand, is an NLI benchmark, containing sentence pairs where one sentence entails the other, contradicts it, or neither of the two (neutral). Importantly, on either task, the model's judgment should not be dependent on nuances in meaning that may be lost in translation. For our experiments, we focus on the German, English, and Chinese partitions of the respective benchmarks.

Instructions. We design task instructions in English to evaluate the model's zero-shot performance. Given that the benchmarks are binary/ternary classification problems, the instructions can be formulated such that the model's responses can be easily standardised and evaluated:

- PAWS-X: *Do the following sentences have the same meaning? Sentence 1: "[sentence_1]" Sentence 2: "[sentence_2]" Please answer with "yes" or "no".*
- XNLI: *Given the following sentence pair, which one of the following is true: (A) the first sentence entails the second sentence, (B) the first sentence contradicts the second sentence, or (C) neither of the two? Sentence 1: "[sentence_1]" Sentence 2: "[sentence_2]" Please answer with "A", "B", or "C".*

In addition, these instructions are translated into German and Chinese by native speakers (see Appendix B), in sum giving us ground truth input data and instructions in each language.

Evaluation. We process each input in a separate request. We only accept model responses matching the template answer (e.g. "yes") or containing it (e.g. "Yes, the sentences have the same meaning."), ignoring casing.[4] In the second case, we apply a semi-automatic standardisation procedure: a function maps the model's responses to one of the template answers, and these mappings are checked, and if necessary corrected, by hand. Using the standardised responses, we can calculate the model's accuracy on the task, as well as the model's consistency across different runs.

3.3 Model-internal translations

We experiment with translations from English to Chinese and German, and from Chinese and German to English. The original English, Chinese, and German tasks serve as baselines for our simulations. Our main goal is to evaluate the consistency between the model's responses on these baselines and the model's responses on a model-internally generated translation, always comparing the source language to the translation from source to target.

Translation procedure and notation. We generate model-internal (zero-shot) translations of both input data and instructions. The translation instructions (see Appendix C), written by native speakers, are always given in the source language. For the task instructions, the model translates the instruction prefix (e.g. *Do the following sentences have the same meaning?*), the word for sentence (e.g. *sentence*), and the instruction suffix (e.g. *Please answer with "yes" or "no".*) in separate requests, and we recompose these translations to an instruction in the target language (see Appendix B). For

[4] Only a negligible amount of responses do not fall into one of these two categories ($< 1\%$). These are mapped onto an additional label indicating invalidity.

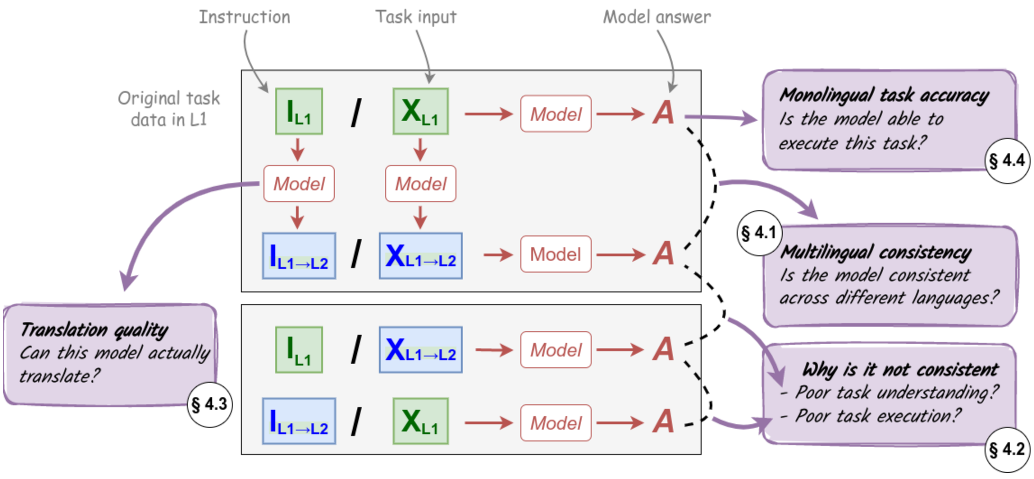
Instruction
Task input
Model answer
Original task data in L1
I_{L1} / X_{L1} → Model → A
Model
Model
$I_{L1\to L2}$ / $X_{L1\to L2}$ → Model → A
I_{L1} / $X_{L1\to L2}$ → Model → A
$I_{L1\to L2}$ / X_{L1} → Model → A
Monolingual task accuracy
Is the model able to execute this task?
§ 4.4
§ 4.1
Multilingual consistency
Is the model consistent across different languages?
Why is it not consistent
- Poor task understanding?
- Poor task execution?
§ 4.2
Translation quality
Can this model actually translate?
§ 4.3

Figure 2: An overview of our experiments and analyses.

the input data, the model translates each sentence per input sentence pair in a separate request.

In what follows, we will denote the instruction of a task T with I and the input to which it is applied with X. We annotate the language in which either of those is given with a subscript, which also indicates whether it is a model translation from another language. Thus, T_{en} refers to the scenario in which both the instruction and the input are given in English, using the original benchmark data, while $T_{en\to de}$ denotes the model's translation of instruction and input sentences from English to German. Following the same principle, $I_{en\to de}$ and $X_{en\to de}$ indicate instructions and input, respectively, that the model has translated from English to German.

Evaluation. The model's translations of the task instructions were reviewed by native speakers who found the translations to be appropriate, apart from slight deviations in the translations from Chinese to English: For PAWS-X the instructions mention a single sentence instead of a sentence pair (*Does the following sentence have the same meaning?*) and for XNLI the word *covers* is used for *entails*. To evaluate the quality of the model's translations of the actual input sentences, we employ BLEU scores (Papineni et al., 2002) calculated with SacreBLEU (Post, 2018), as well as ROUGE (Lin, 2004), and COMET-22 (Rei et al., 2022) scores (see § 4.3).

4 Results

In this section, we discuss the results of our experiments (see Figure 2), beginning with our primary experiment in which we assess how consistent the model's task understanding is across languages (§ 4.1). In subsequent experiments, we investigate the individual effects of translating the dataset or the instructions (§ 4.2), and ensure that inconsistencies do not arise from inaccurate translations (§ 4.3) or poor task performance (§ 4.4).

4.1 Multilingual consistency

In our primary experiment, we assess the consistency of a model's understanding by comparing the model's responses in a monolingual setting – with the original input data and instruction language – with its responses when using a model-internal translation of those. Crucially, the task translations are produced by the model itself, rather than externally. Assuming that the model is a good model of translation (see § 4.3), its translations should be meaning-preserving. In that case, *if* the model has a meaning-based task understanding, its responses to both task versions should be consistent.

The results are reported in column T (*Task*) of Table 1. As we can see, there is not a single case where the scores are near-maximal, indicating that the task understanding of the model is not consistent across the evaluated languages. Regarding the language pairs, consistencies tend to be higher when translating between English and German compared to English and Chinese, with an exception for translations *to* English on XNLI (bottom rows). More details on the differences in predictions before and after translation can be found in Appendix D and a qualitative analysis of the translations from English to German in Appendix E. Besides, to provide an example where our paradigm is applied to a monolingual benchmark, we run the main experiment also for BoolQ (Clark et al., 2019), which yields similar results (see Appendix F).

	Src→Tgt	Consistency		
		T	*I*	*X*
PAWS-X	en→de	0.84	0.93	0.85
	en→zh	0.76	0.91	0.79
	de→en	0.86	0.93	0.86
	zh→en	0.70	0.87	0.75
XNLI	en→de	0.74	0.81	0.76
	en→zh	0.67	0.77	0.71
	de→en	0.63	0.69	0.81
	zh→en	0.67	0.79	0.72

Table 1: Consistency between baseline (T_{src}) and model-internal translation from source to target language ($T_{src \rightarrow tgt}$). Shown are the consistencies for translating input data and instruction (column *T*), instruction only (*I*), or input data only (*X*).

4.2 Interpretation and execution consistency

When the model is inconsistent across languages, we need to determine whether it is due to an inadequate understanding of what it is asked to do in the target language or an inability to perform what it is asked to do in that language. We differentiate these effects by assessing the model's consistency when translating only the instruction, while retaining the original input sentences (e.g. comparing T_{en} and $I_{en \rightarrow de}/X_{en}$) and its consistency when translating only the input sentences while preserving the original instruction (e.g. comparing T_{en} and $I_{en}/X_{en \rightarrow de}$). We show the results in Table 1.

Neither consistencies for translating only the instructions (column *I*) nor those for translating only the input sentences (column *X*) are at their maximum, indicating that the model is inconsistent in both interpretation and execution. Inconsistencies are consistently higher for PAWS-X than XNLI, probably because PAWS-X is a binary and XNLI is a ternary classification problem. However, even translating a simple instruction, such as the one for PAWS-X, leads to inconsistencies for all translations. Consistencies seem to decrease more when translating the input sentences compared to the instructions (except for German to English on XNLI) and even more when translating both (column *T*). Thus, inconsistencies in complete translations seem to be driven by differences in both task interpretation and execution, although differences in execution are more pronounced.

	Src→Tgt	BLEU
PAWS-X	en→de	56.5
	en→zh	49.2
	de→en	60.0
	zh→en	37.6
XNLI	en→de	41.4
	en→zh	43.5
	de→en	45.8
	zh→en	28.0

Table 2: BLEU scores for the model-internal translation of the input data.

4.3 Consistency and translation quality

The metric we propose in some way conflates monolingual task understanding and translation quality: inconsistencies can be driven by misalignment in task understanding, but also by poor translation quality. While both are important, and the metric therefore favours models that do well across the board, it is worth further investigating which of the two drives the observed inconsistencies.

We start by considering the hypothesis that the model's consistency is suboptimal simply because it is not a good model for translation. If the translation quality is poor, inconsistencies may arise from differences in meaning between original and translated inputs, rather than a language-dependent task understanding. To evaluate the model's translation quality specifically on the benchmark data, we examine the translations of the input data for all languages and directions using BLEU scores (see Table 2) and other commonly adapted metrics for translation quality (see Appendix G).

All metrics indicate that the model's translations are of high quality across tasks and languages, with the sole exception of translations from Chinese to English. The scores are generally higher for PAWS-X than XNLI, which might be due to the more challenging and diverse text sources used in generating XNLI. The high scores thus suggest that, for most of the considered source-target language combinations, inconsistencies cannot be ascribed to changes in meaning induced by the translation.

To further substantiate this claim, we compute the Pearson correlations between the BLEU score of the translation and the (binary) consistencies between the model's original responses and its responses on the translated benchmark data (see Table 3, top row for each task). We focus on the simulations with model-internal translations of the input sentences, keeping the instruction in the source language (e.g. $I_{en}/X_{en\rightarrow de}$). For these simulations, we can obtain a translation quality score per data point, which is not confounded with the translation quality of the instruction. The BLEU score for a given data point is calculated by averaging the scores of the two sentences from the sentence pair. All correlations are positive, yet, the absolute values are very low (≤ 0.09). These findings suggest that the observed inconsistencies are largely independent of the translation quality, at least in light of the generally high translation quality observed for this specific model. Additional evidence is presented in Appendix H, revealing significant inconsistencies even when exclusively using the best translations.

4.4 Consistency and performance

While we have shown that the model's inconsistency does not stem from poor translation quality, it could still stem from an inability to perform the task, leading to somewhat "random" responses on different task versions. To investigate this hypothesis we look at the model's accuracies.

Task accuracies. In Table 4 (column T_{src}), we report the monolingual task accuracies for the model on both tasks, for all languages. Accuracies are generally higher for PAWS-X (with only two class labels) than XNLI (with three class labels). In particular, accuracies for German on XNLI are very low. Appendix I presents the accuracies for various combinations of input data and instruction languages, which indicate that the model struggles with the German instruction (rather than input) for XNLI. Furthermore, the accuracies for English are higher than for other languages. While this may not be surprising given the predominantly English training data, it does raise an intriguing point: if a model can perform a particular task in English, and it can correctly translate the task into a different language, why is it not able to perform the task at a similar level in that other language?

	ρ (BLEU, consistency)			
	en→de	en→zh	de→en	zh→en
PAWS-X	0.02	0.07	0.06	0.03
XNLI	0.03	0.02	0.05	0.09

Table 3: Pearson correlation between BLEU scores and model consistency between original and translated inputs ($I_{source}/X_{source\rightarrow target}$).

To further investigate this point, we now consider the accuracies of the model on the task using the model's own translation, which we report in Table 4 (column $T_{src\rightarrow tgt}$). Accuracies for translating either instructions or input sentences only can be found in Appendix J. The results confirm our earlier observation that the model does not maintain consistent meaning representations across languages: even though translations are generated by the model itself and thus should be meaning-preserving according to the model, they lead to differences in performance (compared to the baselines in column T_{src}).

These differences in performance also have practical consequences. While translating from English to German or Chinese leads to a decrease in accuracy, translating from German or Chinese to English leads to an *increase* in accuracy for both PAWS-X and XNLI. Such improvements can also be observed when translating to English from other languages, like French and Spanish, and are largely due to translating the instruction (see Appendix J). It seems that the model's language-dependent task understanding – especially interpretation – can be exploited to increase performance on "lower"-resource languages by instructing the model to first translate the incoming prompt to English and then to perform the task.

Consistent correct vs incorrect. We further investigate if there is a difference in consistency between examples for which the model provides a correct answer and those for which it provides an incorrect answer. This comparison is interesting because correct and incorrect consistent examples provide different levels of evidence for the consistency of a model. Being consistently *incorrect* across two

			Accuracy	
	Src	**Tgt**	T_{src}	$T_{src\rightarrow tgt}$
PAWS-X	en	de	0.77	0.76
	en	zh		0.66
	de	en	0.71	0.73
	zh	en	0.60	0.68
XNLI	en	de	0.71	0.60
	en	zh		0.60
	de	en	0.48	0.65
	zh	en	0.56	0.61

Table 4: Accuracies on PAWS-X and XNLI for the original task T_{src}, and model-internal translations $T_{src\rightarrow tgt}$ from source (src) to target (tgt) language.

		T_{en}	$T_{en\rightarrow de}$	$T_{en\rightarrow zh}$	$T_{de\rightarrow en}$	$T_{zh\rightarrow en}$
PAWS-X	consistency all	0.99	0.84	0.76	0.86	0.70
	consistency correct	0.99	0.89	0.78	0.92	0.82
	consistency incorrect	0.98	0.67	0.71	0.72	0.52
XNLI	consistency all	0.98	0.74	0.67	0.63	0.67
	consistency correct	0.99	0.77	0.71	0.83	0.80
	consistency incorrect	0.96	0.66	0.57	0.45	0.50

Table 5: Detailed consistencies for the core experiment as well as for a baseline of two different runs with T_{en}. Listed are the consistency across all responses (consistency all), as well as the consistency across responses that were correct (consistency correct) and responses that were incorrect (consistency incorrect) on the source task.

examples points to an error in the model's understanding but provides stronger evidence for the consistency of its underlying representations than examples that are consistently correct. Whereas the latter are correct in both languages and could, in theory, have been inferred independently from the data for those respective languages, it is more unlikely that a model makes an identical but unrelated incorrect inference in two different languages.

First, we establish a baseline, by computing the consistency between two runs with the same T_{en} inputs (Table 5, first column, row 1 for each task, respectively). The overall consistencies for this baseline are very high: 99% for PAWS-X and 98% for XNLI. Accordingly, when asked the same question multiple times, the model usually gives the same response. In the second and third row (per task, respectively), we further break down consistency and compute what percentage of the correct and incorrect examples were consistent. As we can see, the baseline case has a high consistency for incorrect responses (98% and 96%), implying that the model's errors are systematic and not due to random guessing.

Moving to the model-internal translations, we observe a general decrease in consistency that affects both correct and incorrect responses. However, the consistency for incorrect examples is notably lower than for correct examples. Given that the model's errors are systematic, this discrepancy suggests that at least some of the consistently correct examples have been inferred independently in both languages. In conclusion, the comparatively low consistencies for incorrect examples provide corroborating evidence for a sense-dependent task understanding.

5 Conclusion

In this paper, we presented a novel paradigm for evaluating language models, which leverages consistency across different linguistic senses. Our method can be used to assess generalisation ability beyond specific forms. It offers affordability and applicability to different evaluation tasks, while also mitigating the risk of evaluating on data that the model has already encountered during training. As such, multisense evaluation is not an *alternative* to current benchmarks but a *complement*. Reporting consistency next to standard evaluation metrics will make model evaluation more meaningful in providing an estimate of how well the model understands a given task beyond its specific form. Therefore, we encourage other researchers to treat multisense consistency as an essential part of benchmarking.

To showcase the effectiveness of our paradigm, we conducted a *multilingual* multisense consistency evaluation of ChatGPT (gpt-3.5-turbo), a SOTA LLM. The results of this evaluation unveiled significant inconsistencies across different senses generated by the model itself through translation, suggesting a lack of genuine, cross-sense task understanding. To ensure the validity of this interpretation, we ruled out alternative explanations such as model-subjective or objective changes in meaning caused by the translation as well as inadequate performance on the original task. Collectively, these findings show that ChatGPT exhibits a language- and therefore sense-dependent task understanding, which might also affect other leading LLMs.

Our paradigm can be cheaply and easily expanded to include more languages, tasks, models, and notions of "sense". Our choice to generate multiple senses through translation is well-suited for evaluating current and future models, given the growing trend towards multilingual models with increasingly proficient translation abilities. Nevertheless, numerous other multisense evaluations are conceivable. For instance, instead of using model-internal translations, one could employ model-internal paraphrases. Multiple senses could also be generated in different domains, such as arithmetic (different formulas yielding the same result) or code (different functions producing the same input-output mapping). Last but not least, calculating consistency for various tasks may help disentangle "unfounded" language-specific differences (forming the focus of our analysis) from differences related to cultural bias.

In conclusion, multisense consistency can be applied as long as the model under investigation can create different senses of a given task and has some understanding of the task in its original sense. It offers the possibility of evaluating the task understanding detached from a specific task realisation, and we hope it will contribute to making standard benchmark evaluations more meaningful.

Limitations

While our method can certainly be extended to other tasks and models, some of these extensions may prove more challenging than others. In particular, evaluating consistency between model responses that are more variable than the ones in our experiments is less straightforward, and requires an appropriate definition of consistency. More variable responses may arise when working with LLMs that have not been adapted to deal with instructions. For example, we instruct ChatGPT to choose a response from a set of predefined responses (*Please answer with "yes" or "no"*, *Please answer with "A", "B", or "C"*) and it largely follows these instructions. A standard LLM may deviate from these answer templates, leading to complications in calculating the consistency. In addition, more variable responses may arise when dealing with tasks that do not correspond to a classification problem. Even testing factual knowledge with a question-answering task may lead to variations in responses. For example, model responses like *5*, *5 times*, or *five*, may be consistent but are different. Further, the model can generate responses that are only partially overlapping, e.g. *disastrous financial situation* versus *bad financial situation*, which might require a graded definition of consistency. Thus, moving forward, it is important to develop appropriate definitions of consistency as well as corresponding automatic evaluation procedures. Given that judging whether two answers have the same meaning is much easier than providing these answers, the consistency evaluation might even be outsourced to the model under investigation.

Ethics statement

We proposed a novel method for evaluating the self-consistency of LLMs by using the models themselves to generate alternative forms or senses of the same task. If a model is self-consistent according to this evaluation, its task understanding goes beyond matching patterns that are present in specific forms. Importantly, though, the model can still be subject to the many problems that currently pertain to pretrained LLMs such as hallucinations or biases. Thus, when using the model to generate different forms or to make predictions for a certain task, its output may contain wrong information, as well as biased and offensive content. These problematic outputs may or may not lead to inconsistencies, and as discussed in the conclusion, future work could try to employ multisense consistency as a tool to detect them. As of now, however, multisense consistency is a means to evaluate a model's robustness, not a means to determine whether the content of its answers is desirable.

Acknowledgements

We would like to thank Felipe Cerdas, Xin Huang, and Vera Lamprecht for their help with the translations. In addition, we would like to thank Marco Baroni and Ryan Nefdt for giving us feedback on an earlier version the draft.

References

Chris Alberti, Daniel Andor, Emily Pitler, Jacob Devlin, and Michael Collins. 2019. Synthetic QA corpora generation with roundtrip consistency. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6168–6173, Florence, Italy. Association for Computational Linguistics.

Akari Asai and Hannaneh Hajishirzi. 2020. Logic-guided data augmentation and regularization for consistent question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5642–5650, Online. Association for Computational Linguistics.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2020. PIQA: Reasoning about physical commonsense in natural language. In *Thirty-Fourth AAAI Conference on Artificial Intelligence*.

Jifan Chen, Eunsol Choi, and Greg Durrett. 2021. Can NLI models verify QA systems' predictions? In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3841–3854, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2924–2936, Minneapolis, Minnesota. Association for Computational Linguistics.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. 2021. Measuring and Improving Consistency in Pretrained Language Models. *Transactions of the Association for Computational Linguistics*, 9:1012–1031.

Gottlob Frege. 1892. Über Sinn und Bedeutung ["On sense and reference"]. *Zeitschrift für Philosophie und philosophische Kritik*, 100(1):25–50.

David Heineman. 2023. Rethinking reasoning evaluation with theories of intelligence.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations (ICLR)*.

Arian Hosseini, Siva Reddy, Dzmitry Bahdanau, R Devon Hjelm, Alessandro Sordoni, and Aaron Courville. 2021. Understanding by understanding not: Modeling negation in language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1301–1312, Online. Association for Computational Linguistics.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4411–4421. PMLR.

Dieuwke Hupkes, Mario Giulianelli, Verna Dankers, Mikel Artetxe, Yanai Elazar, Tiago Pimentel, Christos Christodoulopoulos, Karim Lasri, Naomi Saphra, Arabella Sinclair, Dennis Ulmer, Florian Schottmann, Khuyagbaatar Batsuren, Kaiser Sun, Koustuv Sinha, Leila Khalatbari, Maria Ryskina, Rita Frieske, Ryan Cotterell, and Zhijing Jin. 2023. State-of-the-art generalisation research in NLP: A taxonomy and review. *arXiv*, arXiv:2210.03050.

Myeongjun Jang, Deuk Sin Kwon, and Thomas Lukasiewicz. 2022. BECEL: Benchmark for consistency evaluation of language models. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3680–3696, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Myeongjun Jang and Thomas Lukasiewicz. 2023. Consistency analysis of ChatGPT. *arXiv*, arXiv:2303.06273.

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 Association for Computational Linguistics*, pages 7811–7818, Online. Association for Computational Linguistics.

Nora Kassner, Oyvind Tafjord, Hinrich Schütze, and Peter Clark. 2021. BeliefBank: Adding memory to a pre-trained language model for a systematic notion of belief. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8849–8861, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tao Li, Vivek Gupta, Maitrey Mehta, and Vivek Srikumar. 2019. A logic-driven framework for consistency of neural models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3924–3935, Hong Kong, China. Association for Computational Linguistics.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, Xiaodong Fan, Ruofei Zhang, Rahul Agrawal, Edward Cui, Sining Wei, Taroon Bharti, Ying Qiao, Jiun-Hung Chen, Winnie Wu, Shuguang Liu, Fan Yang, Daniel Campos, Rangan Majumder, and Ming Zhou. 2020. XGLUE: A new benchmark datasetfor cross-lingual pre-training, understanding and generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6008–6018, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Nick McKenna, Tianyi Li, Liang Cheng, Mohammad Javad Hosseini, Mark Johnson, and Mark Steedman. 2023. Sources of hallucination by large language models on inference tasks. *arXiv*, arXiv:2303.06273.

Pasquale Minervini and Sebastian Riedel. 2018. Adversarially regularising neural NLI models to integrate logical background knowledge. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 65–74, Brussels, Belgium. Association for Computational Linguistics.

Eric Mitchell, Joseph Noh, Siyan Li, Will Armstrong, Ananth Agarwal, Patrick Liu, Chelsea Finn, and Christopher Manning. 2022. Enhancing self-consistency and performance of pre-trained language models through natural language inference. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1754–1768, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Melanie Mitchell and David C. Krakauer. 2023. The debate over understanding in AI's large language models. *Proceedings of the National Academy of Sciences*, 120(13):e2215907120.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. 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.

OpenAI. 2023. GPT-4 technical report. *arXiv*, arXiv:2303.08774.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. XCOPA: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376, Online. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Inioluwa Deborah Raji, Emily Denton, Emily M. Bender, Alex Hanna, and Amandalynne Paullada. 2021. AI and the everything in the whole wide world benchmark. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*.

Sagnik Ray Choudhury, Anna Rogers, and Isabelle Augenstein. 2022. Machine reading, fast and slow: When do models "understand" language? In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 78–93, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S. Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *2011 AAAI Spring Symposium Series*.

Sebastian Ruder, Noah Constant, Jan Botha, Aditya Siddhant, Orhan Firat, Jinlan Fu, Pengfei Liu, Junjie Hu, Dan Garrette, Graham Neubig, and Melvin Johnson. 2021. XTREME-R: Towards more challenging and nuanced multilingual evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10215–10245, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Social IQa: Commonsense reasoning about social interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4463–4473, Hong Kong, China. Association for Computational Linguistics.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *CoRR*, abs/2206.04615.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv*, arXiv:2307.09288.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019a. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems (NeurIPS)*, Red Hook, NY, USA. Curran Associates Inc.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Haohan Wang, Da Sun, and Eric P. Xing. 2019b. What if we simply swap the two text fragments? A straightforward yet effective way to test the robustness of methods to confounding signals in nature language inference tasks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7136–7143.

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.

Ludwig Wittgenstein. 1953. *Philosophical investigations. Philosophische Untersuchungen.* Macmillan.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019. PAWS-X: A cross-lingual adversarial dataset for paraphrase identification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3687–3692, Hong Kong, China. Association for Computational Linguistics.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019. PAWS: Paraphrase adversaries from word scrambling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1298–1308, Minneapolis, Minnesota. Association for Computational Linguistics.

A Genbench evaluation card

Our work uses generalisation across senses to assess task understanding in LLMs. In Figure 3, we provide the GenBench eval card (Hupkes et al., 2023) of our experiments.

<table>
<tr><th colspan="12">Motivation</th></tr>
<tr><td colspan="3"><i>Practical</i></td><td colspan="3"><i>Cognitive</i></td><td colspan="3"><i>Intrinsic</i>
☒</td><td colspan="3"><i>Fairness</i></td></tr>
<tr><th colspan="12">Generalisation type</th></tr>
<tr><td colspan="2"><i>Compo-sitional</i></td><td colspan="2"><i>Structural</i></td><td colspan="2"><i>Cross Task</i></td><td colspan="2"><i>Cross Language</i>
☒</td><td colspan="2"><i>Cross Domain</i></td><td colspan="2"><i>Robust-ness</i>
☒</td></tr>
<tr><th colspan="12">Shift type</th></tr>
<tr><td colspan="3"><i>Covariate</i>
☒</td><td colspan="3"><i>Label</i></td><td colspan="3"><i>Full</i></td><td colspan="3"><i>Assumed</i></td></tr>
<tr><th colspan="12">Shift source</th></tr>
<tr><td colspan="3"><i>Naturally occuring</i>
☒</td><td colspan="3"><i>Partitioned natural</i></td><td colspan="3"><i>Generated shift</i></td><td colspan="3"><i>Fully generated</i></td></tr>
<tr><th colspan="12">Shift locus</th></tr>
<tr><td colspan="3"><i>Train–test</i></td><td colspan="3"><i>Finetune train–test</i></td><td colspan="3"><i>Pretrain–train</i></td><td colspan="3"><i>Pretrain– test</i>
☒</td></tr>
</table>

Figure 3: Our experiments assess cross-lingual generalisation for natural corpora, in pretrained LLMs, to assess LLM task understanding.

B Task instructions

Table 6 shows the task instructions for both tasks, in all languages. The table shows the original English, German, and Chinese instructions, as well as the model-internal translations of these instructions.[5]

C Translation instructions

We used the following instructions for model-internal translations:

- en→de/zh:
  Please translate the following text into German/Chinese: "[text]"

- de→en:
  Bitte übersetze den folgenden Text ins Englische: "[text]"

[5] Note that we also accept 不是 instead of 否 for I_{zh}.

Task	Language	Instruction
PAWS-X	I_{en}	*Do the following sentences have the same meaning? Sentence 1: “[sentence_1]” Sentence 2: “[sentence_2]” Please answer with “yes” or “no”.*
	I_{de}	*Haben die folgenden Sätze die gleiche Bedeutung? Satz 1: “[sentence_1]” Satz 2: “[sentence_2]” Bitte antworte mit “ja” oder “nein”.*
	I_{zh}	下面的句子有着相同的含义吗？句子*1: “[sentence_1]”* 句子*2: “[sentence_2]”* 请用“是”或者“否”回答。
	$I_{en\rightarrow de}$	*Haben die folgenden Sätze die gleiche Bedeutung? Satz 1: “[sentence_1]” Satz 2: “[sentence_2]” Bitte antworten Sie mit “Ja” oder “Nein”.*
	$I_{en\rightarrow zh}$	以下句子的意思相同吗？句子*1: “[sentence_1]”* 句子*2: “[sentence_2]”* 请用“是”或“不是”回答。
	$I_{de\rightarrow en}$	*Do the following sentences have the same meaning? Sentence 1: “[sentence_1]” Sentence 2: “[sentence_2]” Please respond with ‘yes’ or ‘no’.*
	$I_{zh\rightarrow en}$	*Does the following sentence have the same meaning? Sentence 1: “[sentence_1]” Sentence 2: “[sentence_2]” Please answer with ‘yes’ or ‘no’.*
XNLI	I_{en}	*Given the following sentence pair, which one of the following is true: (A) the first sentence entails the second sentence, (B) the first sentence contradicts the second sentence, or (C) neither of the two? Sentence 1: “[sentence_1]” Sentence 2: “[sentence_2]” Please answer with “A”, “B”, or “C”.*
	I_{de}	*Welche dieser Aussagen trifft auf das folgende Satzpaar zu: (A) der erste Satz impliziert den zweiten Satz, (B) der erste Satz widerspricht dem zweiten Satz, oder (C) keines von beiden? Satz 1: “[sentence_1]” Satz 2: “[sentence_2]” Bitte antworte mit “A”, “B” oder “C”.*
	I_{zh}	对于给出的一对句子，以下哪一个选项是正确的：（*A*）第一个句子涵盖了第二个句子（*B*）第一个句子与第二个句子矛盾（*C*）两者都不？句子*1: “[sentence_1]”* 句子*2: “[sentence_2]”* 请用*“A”*、*“B”*或*“C”*来回答。
	$I_{en\rightarrow de}$	*Angesichts des folgenden Satzpaares, welche der folgenden Aussagen ist wahr: (A) Der erste Satz impliziert den zweiten Satz, (B) Der erste Satz widerspricht dem zweiten Satz oder (C) Keines von beiden? Satz 1: “[sentence_1]” Satz 2: “[sentence_2]” Bitte antworten Sie mit “A”, “B” oder “C”.*
	$I_{en\rightarrow zh}$	给定以下句子对，哪一个是正确的：（*A*）第一句蕴含第二句，（*B*）第一句与第二句相矛盾，还是（*C*）两者都不是？句子*1: “[sentence_1]”* 句子*2: “[sentence_2]”* 请用*“A”*、*“B”*或*“C”*回答。
	$I_{de\rightarrow en}$	*Which of these statements applies to the following pair of sentences: (A) the first sentence implies the second sentence, (B) the first sentence contradicts the second sentence, or (C) neither of the above? Sentence 1: “[sentence_1]” Sentence 2: “[sentence_2]” Please reply with “A”, “B”, or “C”.*
	$I_{zh\rightarrow en}$	*For a given pair of sentences, which of the following options is correct: (A) The first sentence covers the second sentence. (B) The first sentence contradicts the second sentence. (C) Neither of them? Sentence 1: “[sentence_1]” Sentence 2: “[sentence_2]” Please answer with “A”, “B”, or “C”.*

Table 6: Task instructions in different languages. The original instructions in English, German, and Chinese are given by I_{en}, I_{de}, and I_{zh}. The model-internal translations of these instructions (from source to target language) are given by $I_{source\rightarrow target}$.

	label	
	true	false
ground truth	0.45	0.55
T_{en}	0.58	0.42
$T_{en\to de}$	0.62	0.38
$T_{en\to zh}$	0.69	0.31
T_{de}	0.65	0.34
$T_{de\to en}$	0.62	0.38
T_{zh}	0.78	0.21
$T_{zh\to en}$	0.53	0.47

Table 7: Ground truth and predicted label distributions for PAWS-X.

	label		
	entail	neutral	contradict
ground truth	0.33	0.33	0.33
T_{en}	0.48	0.21	0.30
$T_{en\to de}$	0.62	0.18	0.20
$T_{en\to zh}$	0.42	0.36	0.22
T_{de}	0.69	0.26	0.05
$T_{de\to en}$	0.54	0.15	0.31
T_{zh}	0.52	0.25	0.22
$T_{zh\to en}$	0.40	0.31	0.30

Table 8: Ground truth and predicted label distributions for XNLI.

- zh→en:
  请将下面的文字翻译成英语*: "[text]"*

D Elaborations on the inconsistencies in the main experiment

Tables 7 and 8 display the distributions of the ground truth labels and the predicted labels for different representations of PAWS-X and XNLI, respectively. Regarding PAWS-X (see Table 7), the model consistently overestimates the number of paraphrases across all task representations. At the same time, the predicted label distributions vary – sometimes strongly – between the original task versions (T_{en}, T_{de}, T_{zh}) and their model-internal translations. For example, the amount of predicted paraphrases increases from 58% in English to 62% when translating to German and 68% when translating to Chinese. More extremely, the model predicts 78% paraphrases on the Chinese task version but only 53% on its translation to English. These distributions suggest the presence of language-dependent biases in the model's assessment of whether two sentences convey the same meaning or not. In particular, if the model translates from a certain source language to a certain target language, the predicted label distribution for the model-internal translation ($T_{source\to target}$) becomes more similar to that of the "model-external" translation (T_{target}). In other words, if the model predicts more or fewer paraphrases on the target language (T_{target}) compared to the source language (T_{source}), the predictions on the model-internal translation tend to increase or decrease accordingly.

These patterns are reflected in the types of inconsistencies observed when comparing the model's responses on the original task version to those on the model-internal translation. When translating from English to German, 60% of the inconsistencies are cases where the model classifies a sentence pair as a paraphrase in German but not in English. When translating from English to Chinese (with an even higher proportion of predicted paraphrases), these cases account for 0.72% of the inconsistencies. Conversely, when translating from German or Chinese to English, most inconsistencies are cases where the model classifies a sentence pair as a paraphrase in the source language but not in English (60% for German and 92% for Chinese).

For XNLI (see Table 8), the model consistently overestimates the number of entailments and, correspondingly, tends to underestimate the number of contradicting and neutral sentence pairs. Especially notable are the high amounts of predicted entailments for T_{de} (69%) and $T_{en\to de}$ (62%), which are further explored in the qualitative analysis provided in Appendix E. Despite this general trend, the predicted distributions exhibit significant variations between the source language and the model-internal translation. For example, while the model predicts only 48% entailments on T_{en}, it predicts 62% on $T_{en\to de}$. Conversely, while it predicts 69% entailments on T_{de}, it predicts only 54% on $T_{de\to en}$.

Compared to PAWS-X, it is more challenging to identify patterns in the inconsistencies for XNLI. Firstly, there are more interactions between inconsistencies as there are three class labels instead of two. Secondly, the more complex task instruction may have a stronger influence, leading to mixed effects from differences in task interpretation and execution. However, for translations between English and German (which are also of higher quality than translations between English and Chinese), some patterns can still be identified. Most inconsistencies when translating from English to German involve sentences where the model switches from *neutral* (33%) or *contradiction* (35%) to *entailment*, together accounting for 68% of the inconsistencies. When translating from German to English, in turn, a large proportion of the sentence pairs formerly classified as *entailment* are now classified as *contradiction*, constituting 51% of the inconsistencies (with 9% for switching from *entailment* to *neutral*). These inconsistencies might be considered a particularly strong argument against a genuine task understanding by the model, as it regularly switches interpretation between the contrasting concepts of "entailment" and "contradiction", rather than mostly transitioning between *neutral* and the other two categories.

E Qualitative analysis for model-internal translations from English to German

We conduct a qualitative analysis of the model's inconsistencies when translating from English to German. We examine 100% of the inconsistencies on PAWS-X (329 data points) and the first 50% of the inconsistencies on XNLI (664 data points).

To begin with, our focus lies on verifying whether the model's change in response is indeed due to a change in sense (but not meaning) or whether there might be an alternative explanation. For that purpose, we classify the data points into two categories: category (1) if no alternative explanation can be identified and category (2) if an alternative explanation can be identified. After reviewing the examples, we define the following alternative explanations for category (2):

(2.1) Ambiguity

- Source ambiguities: the source sentence contains ambiguous expressions and the model a) switches interpretation or b) resolves this ambiguity.

- Target ambiguities: the target sentence contains ambiguous expressions that were not ambiguous in the source sentence.

(2.2) Translation quality

- The translation does not preserve meaning.

- The translation is of poor linguistic quality, potentially making the task more difficult.

(2.3) Identical sentences

- The translations of the input sentences are identical, which confuses the model.

Note that this is a very conservative encoding. Firstly, some of these cases should arguably not cause inconsistencies. For example, if the model "understands" what it means for two sentences to have the same meaning, it should also understand that two identical sentences have the same meaning (subcategory 2.3). Secondly, even if there is ambiguity in the source or target language, or the linguistic quality is subpar, it is not clear whether the model changes its response because of these factors.

Here are examples illustrating the subcategories. An example of ambiguity is the following sentence pair from PAWS-X: "*The film stars Oscar Nunez, Rob Huebel, Timothée Chalamet, Lily Rabe, Anthony Quintal, and Lili Reinhart.*", "*Film stars Oscar Nunez, Rob Huebel, Timothée Chalamet, Lily Rabe, Anthony Quintal, and Lili Reinhart.*" The first sentence is ambiguous as to whether *stars* is a verb or part of the compound noun *film stars*. In German, it is translated as a verb ("*Der Film hat Oscar Nunez, [...]*"), and as a result the sentence pair is classified as a paraphrase in English but not in German. An example of an inaccurate translation is the following sentence pair from XNLI: "*Smaller boats for local jaunts can be rented at Sea Horse Boat Rentals, Marsh Harbour, Abacos (Tel.,*", "*You can rent one passenger boats.*" Due to the missing hyphen between *one* and *passenger* in the premise, *one passenger boats* is interpreted as *one passenger boat* in the German translation ("*Sie können ein Passagierboot mieten.*"). The model correctly predicts that the sentences are *neutral* in German but predicts *entailment* in English. Finally, here is an example of identical sentences from PAWS-X: "*The first series was recorded by critics better than the second .*", "*The first series was better received by critics than the second .*"; which are both accurately translated to "*Die erste Staffel wurde von Kritikern besser aufgenommen als die zweite.*" While the model predicts that the sentences have the same meaning in English, it only replies that the sentences are identical in German ("*Die beiden Sätze sind identisch.*").

Despite the conservative encoding, a majority of the inconsistencies – 78% for PAWS-X and 86% for XNLI – fall into category (1), which means that none of the alternative explanations are applicable.

For PAWS-X, 6% of the inconsistencies may be related to ambiguities, 10% to translation quality, and 4% to identical sentence pairs.[6] For XNLI, it is 7% for ambiguities, 7% for translation quality, and 0% for identical sentence pairs. Alongside the analyses in § 4.3 and § 4.4, this less general but more in-depth analysis provides further evidence that the model's responses are sense-dependent.

Examining examples from category (1) can help understand how a sense-dependent task understanding might lead to inconsistencies. In most cases, it remains unclear why the model makes different predictions. Especially for PAWS-X, it is surprising how the model is sometimes fooled by the adversarial nature of the sentences in one language but not the other. For example, given the sentence pair "*The Tabaci River is a tributary of the River Leurda in Romania .*", "*The Leurda River is a tributary of the Tabaci River in Romania .*"; and the correct German translations "*Der Fluss Tabaci ist ein Nebenfluss des Flusses Leurda in Rumänien.*", "*Der Fluss Leurda ist ein Nebenfluss des Tabaci-Flusses in Rumänien.*"; the model identifies that the sentences have different meanings in English but not in German. The reverse case where the model is fooled in English but not in German also exists.

In some cases, one can speculate that certain informational content of the sentences is more readily available to the model in one language than the other, which might influence its response. Take for example the following sentence pair from XNLI: "*Among the many jazz clubs are the famed Jazz Bakery in Culver City, the Catalina Bar and Grill in Hollywood, and the Baked Potato in North Hollywood.*", "*There are no famous jazz clubs in Los Angeles.*" The model correctly predicts *contradiction* in English but predicts *entailment* in German, possibly because the information that Culver City is part of Los Angeles appears more often in English text than in German text. This example constitutes an important exception because, unlike for most other examples, the ability to make a correct prediction is knowledge-dependent. As such, it illustrates the very situation where the model should give the same response in both languages. The fact that the model apparently knows that Culver City is in LA when asked in English but not when asked in German thus provides powerful evidence for a sense-dependent task understanding.

An influence of prior knowledge may also lead to mistakes as in the following example from PAWS-X: "*Stipsits was born in Korneuburg , Germany and spent his childhood in Stammersdorf , Vienna .*", "*Stipsits was born in Korneuburg , and spent his childhood in Stammersdorf , Vienna .*" The model correctly classifies these as paraphrases in English, but argues that the sentences do not have the same meaning in German because Korneuburg is in Austria and not in Germany ("*Nein. Satz 1 ist inkorrekt, da Korneuburg in Österreich liegt und nicht in Deutschland [...]*"). It has very recently been established that LLMs (including ChatGPT) use prior knowledge for language inference, for example, they judge the truth of the hypothesis based on information in the training data rather than information in the premise (McKenna et al., 2023). Our results are in line with this finding and further suggest that the interfering information is language-dependent.

Other cases where the model provides an explanation for its answer (deviating from the answer template) are also revealing. In particular, they show that the model generally struggles to interpret the German instruction for XNLI, consistent with the low accuracies for $T_{en \to de}$ (0.60) and T_{de} (0.48) in Table 4. For example, on one occasion the model responds "*Die richtige Antwort ist (C) Keines von beiden. Die beiden Sätze sind unabhängig voneinander und widersprechen sich nicht.*", on another one "*Die richtige Antwort ist (C) Keines von beiden. Die beiden Sätze haben keine direkte Beziehung zueinander und widersprechen sich auch nicht.*" These responses indicate that the model excludes the option of entailment because the sentences *are independent from each other* or *do not have a direct relationship*. Possibly, the model also applies these as positive criteria for entailment, which would explain why it significantly overestimates the number of entailments in German.

F Main experiment with BoolQ

BoolQ is a question answering dataset where each example consists of a passage and a yes/no question about that passage. We use the validation split of the dataset and prompt the model by providing the passage, followed by the question (capitalised and with a question mark), and the instruction *Please answer with "yes" or "no"*. We retrieve

[6] The remaining 2% are sentences that do not fall into category (1) or (2) because the original sentences are so ungrammatical that it is difficult to determine whether the translation is accurate.

Task version	Consistency	Accuracy
T_{en} (orig)	-	0.86
$T_{en\to de}$	0.89	0.82
$T_{en\to zh}$	0.81	0.78

Table 9: Consistency and accuracy for BoolQ. The first column provides the consistencies between the model's responses on the original task (T_{en}) and the model-internal translations of that task to German ($T_{en\to de}$) and Chinese ($T_{en\to zh}$). The second column provides the model's accuracy for each task version.

the model's responses for T_{en} and evaluate consistency with $T_{en\to de}$ and $T_{en\to zh}$, respectively. The translations of the input sentences are obtained by instructing the model to translate each passage and question in a separate request, using the English translation instruction (see Appendix C). Since the instruction for BoolQ corresponds to the instruction suffix for PAWS-X, we reuse these translations. The resulting consistencies are provided in Table 9, together with the accuracy for each task version. The consistencies follow the same pattern as those for PAWS-X and XNLI when translating from English to German and Chinese (see Table 1): The model is not perfectly consistent regardless of the target language, with lower consistency for the Chinese translation.

G Translation evaluation scores

We evaluate translation quality for the input sentences using BLEU, ROUGE, and COMET-22 scores (see Table 10).

H Inconsistencies for very high quality translations

We extend the analyses from Section 4.3 by calculating the inconsistencies for data points with a BLEU score of at least 50. Our focus remains on translations of the input data and the model is instructed in the original (source) language. Table 11 shows the amount of data (%) included in the analysis, along with the corresponding consistency. Importantly, the model's inconsistency persists, as it never achieves consistencies surpassing 0.87. Moreover, across the board, consistencies exhibit only a slight improvement compared to the original values (see *consistency orig*, same as in Table 1, column *X*). The only substantial increase in performance occurs for translations from Chinese to English on XNLI, with consistency rising from 0.72 to 0.80. This finding aligns with the observation that the translations from Chinese to English are of significantly lower quality than the other translations. Hence, bad translations may reduce consistency, but this phenomenon is only observed in one specific case.

I Performance on mixed languages for input data and instructions

We look at different ablations to understand the effect of using a language other than English for input data or instruction. Table 12 shows the model's accuracy on different combinations of languages for input sentences and instructions, always using the input sentences provided by the multilingual benchmark, and the English, German, or Chinese instructions developed for us by native speakers. Compared to T_{en}, with an accuracy of 0.77 on PAWS-X and 0.71 on XNLI (see Table 4), accuracy decreases when instruction or input data are changed from English to German or Chinese. Changing the language for both at the same time further decreases accuracy, as errors from each language change accumulate (see T_{de} and T_{zh} in Table 4). For PAWS-X there is a more substantial decrease when changing instructions or input data to Chinese compared to German. For XNLI, especially the use of the German instruction is detrimental, with accuracies dropping from 0.71 to 0.50. Testing alternative German instructions reveals that this effect does not only pertain to our specific formulation. While a decrease in performance may be expected for non-English inputs, the extent of this effect when changing only the task instruction is surprising. For example, changing the instruction for PAWS-X from English to Chinese leads to a 10% absolute decrease in accuracy, even though this instruction is very simple.

J Task performance for model-internal translations

Table 13 shows the model's accuracies for all source languages (T_{src}) and the corresponding model-internal translations: instruction only ($I_{src\to tgt}$ / X_{src}), input sentences only (I_{src} / $X_{src\to tgt}$), or both ($T_{src\to tgt}$). In addition, we add accuracies for French and Spanish and their translations to English. § 4.4 shows that model-internal translations from German and Chinese to English increase the model's accuracy compared to the original T_{de} and T_{zh} tasks. The results for

	Src→Tgt	**BLEU**	**Rouge1**	**Rouge2**	**Rouge-l**	**COMET-22**
PAWS-X	en→de	56.5	0.80	0.64	0.77	0.89
	en→zh	49.2	0.68	0.42	0.62	0.86
	de→en	60.0	0.87	0.72	0.83	0.88
	zh→en	37.6	0.73	0.49	0.66	0.85
XNLI	en→de	41.4	0.71	0.52	0.68	0.88
	en→zh	43.5	0.66	0.39	0.62	0.87
	de→en	45.8	0.76	0.57	0.74	0.89
	zh→en	28.0	0.61	0.37	0.57	0.86

Table 10: Evaluation of the model-internal translation of the input data.

		en→de	en→zh	de→en	zh→en
PAWS-X	consistency orig	0.85	0.79	0.86	0.75
	consistency BLEU $>$ 50	0.86	0.82	0.87	0.78
	% included BLEU $>$ 50	56.6	40.1	67.1	20.6
XNLI	consistency (orig)	0.76	0.71	0.81	0.72
	consistency BLEU $>$ 50	0.77	0.72	0.82	0.80
	% included BLEU $>$ 50	35.6	32.3	39.6	10.5

Table 11: Only datapoints with BLEU scores of > 50 are included in this analysis. The table shows the percentage of included data points (*% included BLEU>50*), and the consistency of the model for these selected translations (*consistency BLEU>50*) compared to the original consistency (*consistency orig*) repeated from Table 1 (column *X*).

	X / I			
	en/de	en/zh	de/en	zh/en
PAWS-X	0.75	0.67	0.73	0.68
XNLI	0.50	0.60	0.65	0.59

Table 12: Accuracies on mixed-language combinations of original input data (*X*) and instructions (*I*).

French and Spanish show that translations from other languages to English can also increase accuracy. For instance, translating (input sentences and instruction) from Spanish to English raises the accuracy on PAWS-X from 0.72 to 0.73, and on XNLI from 0.60 to 0.65. Looking at the separate effects of translating the instructions or the input sentences to English suggests that the observed improvements can largely be ascribed to the translation of the instruction, regardless of the source language.

			Acc (orig)	**Acc (translation)**		
	Src	**Tgt**	T_{src}	$T_{src \to tgt}$	$I_{src \to tgt}$ / X_{src}	I_{src} / $X_{src \to tgt}$
PAWS-X	en	de	0.77	0.76	0.77	0.77
	en	zh		0.66	0.75	0.70
	de	en	0.71	0.73	0.72	0.70
	zh	en	0.60	0.68	0.67	0.63
	fr	en	0.72	0.72	0.72	0.71
	es	en	0.72	0.73	0.73	0.71
XNLI	en	de	0.71	0.60	0.63	0.67
	en	zh		0.60	0.63	0.62
	de	en	0.48	0.65	0.64	0.49
	zh	en	0.56	0.61	0.59	0.56
	fr	en	0.58	0.63	0.61	0.60
	es	en	0.60	0.65	0.67	0.60

Table 13: Accuracies on the original multilingual benchmark tasks (T_{src}) and the model-internal translations of these tasks from source (src) to target (tgt) language. We consider translations of both input data and instructions ($T_{src \to tgt}$), instruction only ($I_{src \to tgt}$ / X_{src}), and input data only (I_{src} / $X_{src \to tgt}$). Besides, we add translations from French and Spanish to English to further study whether translating to English can improve performance.