

From Form(s) to Meaning: Probing the Semantic Depths of Language Models Using Multisense Consistency

Xenia Ohmer

Osnabrueck University

Institute for Cognitive Science

xenia.ohmer@uni-osnabrueck.de

Elia Bruni*

Osnabrueck University

Institute for Cognitive Science

elia.bruni@uni-osnabrueck.de

Dieuwke Hupkes*

Meta

dieuwkehupkes@meta.com

The staggering pace with which the capabilities of large language models (LLMs) are increasing, as measured by a range of commonly used natural language understanding (NLU) benchmarks, raises many questions regarding what “understanding” means for a language model and how it compares to human understanding. This is especially true since many LLMs are exclusively trained on text, casting doubt on whether their stellar benchmark performances are reflective of a true understanding of the problems represented by these benchmarks, or whether LLMs simply excel at uttering textual forms that correlate with what someone who understands the problem would say. In this philosophically inspired work, we aim to create some separation between form and meaning, with a series of tests that leverage the idea that world understanding should be consistent across presentational modes—inspired by Fregean senses—of the same meaning. Specifically, we focus on consistency across languages as well as paraphrases. Taking GPT-3.5 as our object of study, we evaluate multisense consistency across five different languages and various tasks. We start the evaluation in a controlled setting, asking the model for simple facts, and then proceed with an evaluation on four popular NLU benchmarks. We find that the model’s multisense consistency is lacking and run several follow-up analyses to verify that this lack of consistency is due to a sense-dependent task understanding. We conclude that, in this aspect, the understanding of LLMs is still quite far from being consistent and human-like, and

* Shared senior authorship.

Action Editors: Marianna Apidianaki, Abdellah Fourtassi, and Sebastian Padó. Submission received: 16 December 2023; revised version received: 18 April 2024; accepted for publication: 19 June 2024.

<https://doi.org/10.1162/coli.a.00529>

deliberate on how this impacts their utility in the context of learning about human language and understanding.

1. Introduction

In the past ten years, the abilities of neural language models (LMs) have developed at a—for most—unimaginable pace. This progress has aroused much excitement among both scientists and applied researchers, and it comes with a range of interesting questions in various domains. One category of such questions pertains to the type of (linguistic) intelligence that neural networks possess and how studying them may help us make progress on scientific questions related to linguistics, cognitive science, and human language processing (e.g., Baroni 2023; Linzen and Baroni 2021; Hupkes 2020; Pavlick 2023). Specifically, recurrent neural networks (Elman 1990), which were originally proposed as alternative theories of human sequential processing, have been examined in this context, primarily with respect to topics in syntax and morphology (among many others, Dankers et al. 2021; Lakretz et al. 2021; Jumelet et al. 2021; Malouf 2017; Van Schijndel and Linzen 2018; Abnar et al. 2019). More recently, their attention-based counterparts have also gained popularity in exploring human linguistic processing (e.g., Timkey and Linzen 2023; Lakretz et al. 2022). In the fields of cognitive science and psychology, neural networks have, among other things, taken on an important role in the debate about syntactic nativism. In particular, later generations of neural networks, which show strong command of natural language syntax (for an overview, see Chang and Bergen 2023, Section 3), are by some considered to provide a counter-argument to the claim that innate biases are required to learn natural languages (Contreras Kallens, Kristensen-McLachlan, and Christiansen 2023; Piantadosi 2023; Mahowald et al. 2023, i.a.).

While the debate on this has hardly been resolved¹—and likely will not be for a long time—LMs have arrived at a stage where their mastery of syntax is almost undisputed, as they obtain nearly perfect scores on syntactic datasets that are challenging even for humans (Wang et al. 2019; Kocijan et al. 2023; Liang et al. 2023). In recent times, research exploring the capabilities of (large) language models—(L)LMs—has therefore shifted to their ability to correctly process semantics. In this vein, many datasets have been developed to quantify the extent to which LMs are able to conduct a range of different natural language understanding (NLU) tasks (e.g., Wang et al. 2018, 2019; Hendrycks et al. 2021). In the literature, there is considerable discussion about the extent to which these datasets accurately measure what they claim to measure. Commonly used arguments center around the concept of construct validity, and are supported by findings that datasets contain biases (Gururangan et al. 2018; Bencheekroun et al. 2023), can be solved with heuristics rather than understanding (McCoy, Pavlick, and Linzen 2019; Saxon et al. 2023; Sen and Saffari 2020; Niven and Kao 2019), or do not agree with other datasets claiming to measure the same skill (Sun, Williams, and Hupkes 2023). A much less frequently discussed topic is what this new wave of models, which according to many learn under vastly different circumstances than humans, can still teach us about human language (processing).

1 A frequently mentioned critique of this ability is that LMs require vastly more data than humans to arrive at this level of performance (see, e.g., Dupoux [2018] or Warstadt and Bowman [2022] for a discussion). Therefore, more and more research is being carried out to study which syntactic skills language models can learn from smaller amounts of data (Zhang et al. 2021), or even amounts comparable to what children have ingested (Warstadt et al. 2023).

While training on inconceivable amounts of data likely makes modern LLMs less suitable to study questions related to syntactic processing and grammar, their new-found NLU abilities open the door to studying a new realm of questions, related to the nature of meaning and how language expresses it. Some have argued that it is a priori not possible to learn meaning from form alone (e.g., Bender and Koller 2020), yet others disagree or argue that the training signal for at least some LLMs goes beyond form (e.g., Piantadosi and Hill 2022; Mollo and Millière 2023; Pavlick 2023; Mandelkern and Linzen 2023). Here, we take a different stance: Although our approach is embedded in theoretical arguments about the concept of meaning, we propose an empirical method to investigate the notion of meaning acquired when (mostly) being exposed to form. Our focus is not on explaining *how* meaning is acquired from form, but rather on individuating necessary criteria for grasping meaning and developing a metric to quantify this in LLMs.

Our method is inspired by the seminal works of Frege (1892) and Wittgenstein (1953), who both put forward influential philosophical theories of meaning. Frege’s work starts from the observation that if the meaning of a word or phrase were uniquely determined by what it denotes, this would imply that the statements “ $a=a$ ” and “ $a=b$ ” were equally informative, which is evidently not the case, even if a and b refer to the same object. To solve this apparent paradox, Frege introduced the key concept of the **sense** (Sinn) of an expression, which conveys the mode of presentation by which a particular phrase denotes a referent. As such, Frege’s work acknowledges and formalizes the idea that different linguistic expressions can share the same referent. We combine Frege’s notion of sense with Wittgenstein’s idea that the meaning of language is defined by the effect it has on the world (Wittgenstein 1953), which thus functions as an anchor for diverse linguistic forms. Put together, this suggests that having a genuine understanding of language entails understanding its relation to the world, which would in turn imply consistency among different linguistic expressions that pertain to the same entities within the world. As LLMs are trained without direct access to the anchor that is the world, we propose that their understanding can be tested by investigating if they—nevertheless—have constructed their representational space such that they respond consistently across different forms with the same meaning.

We translate this idea into a method to probe the semantic depths of the form-driven meaning acquired by LLMs, which we call **multisense consistency**.² Crucially, we do not presuppose that particular linguistic expressions have the same meaning, but we ask the model itself to generate meaning-preserving expressions, thus focusing more on whether a model has acquired a notion of meaning than on whether that notion is exactly aligned with ours. If a model generates consistent responses when prompted with these expressions, this would suggest it might be linking them to their common underlying meaning. We apply our consistency-based test to investigate one of the currently most advanced models: GPT-3.5.³ In a series of experiments, beginning

2 It is worth pointing out that, according to Frege, different linguistic expressions with the same referent may also have the *same* sense. Our borrowing of the term is, in that sense, loose.

3 Note that GPT-3.5 was trained on more than form. While the details are unknown, the training involved Reinforcement Learning from Human Feedback (Ouyang et al. 2022), which arguably provides additional information such as communicative intent. It has also been argued that, even without this additional training stage, typical training corpora contain information beyond form, for example, written computer programs and the outputs they generate (Bender and Koller 2020). Detecting inconsistencies thus suggests that even this kind of additional information does not give rise to a meaning-based understanding. Beyond that, multimodal LLMs, which we do not consider here, encounter more explicit information about form-meaning mappings during training.

with the evaluation of basic truth-conditional statements and progressing to more complex ones, we discover numerous instances where the LLM responds inconsistently across different, meaning-preserving expressions, even in scenarios as straightforward as reiterating a fact. This is true both when meaning-preserving senses are *paraphrases* and *translations*. Our results, which we substantiate with several follow-up analyses, illustrate that even one of the best-performing LLMs does not seem to have meaning-preserving representations that align with what a Fregean theory of meaning may consider true meaning. While this may come as no surprise to many, it still begs the question of what the conclusion would have been if the model *did* pass this consistency-based test, and if there is anything that could convince us that an LLM has—in fact—truly acquired meaning. We elaborate on this in our discussion.

Outline. In the remainder of this article, we will first take a closer look at Frege’s theory on sense and reference, which provides the framework for our approach (§ 2). We will then give a high-level overview of how multisense consistency can be used to study the discrepancy between competence in form and competence in meaning (§ 3) before providing more details on our experiments, such as the model and the senses considered (§ 4). We discuss results for two different types of datasets—simple hand-crafted probes of factual knowledge and popular NLU benchmarks (§ 5 and § 6, respectively), following up with several analyses to study when and why inconsistencies arise (§ 7). Finally, we position our contribution in the context of related work (§ 8) and discuss our findings within the broader scope of using LLMs as models of meaning (§ 9).

2. Philosophical Background

Our study draws inspiration from philosophical notions of meaning, in particular the one put forth by Frege (1892). Here, we provide a short discussion of this philosophical backbone and its relevance to evaluating LLMs.

Sense and Reference. Before Frege, theories of meaning often struggled to explain the relationship between words and the world they describe, typically approaching this relationship in a linear and simplistic way. These theories faced difficulties in explaining how language could meaningfully refer to non-existent entities, define the meaning of statements that cannot be easily mapped to a truth value, or handle identity statements where two different expressions appear to refer to the same object. Frege’s introduction of the concepts of **sense** (Sinn) and **reference** (Bedeutung) offered a solution to these problems. The reference of an expression is the actual entity or concept the expression corresponds to in the real world and is decisive in determining the truth value of a sentence. The sense of an expression, in contrast, comprises the way in which this reference is presented. For example, the morning star and the evening star refer to the same celestial body, Venus, but have different senses (see Figure 1). Not only can the same reference be presented through different senses, but the same sense can also be realized through different expressions—with some surface level variations (Frege [1918–1919] mentions injections such as “alas” or “thank God” as examples). If two forms (expressions) have the same sense, it is possible to determine a priori that they map to the same referent. However, if two forms have different senses, learning that they have the same referent provides an extension of our knowledge. The distinction between sense and reference is vital for understanding identity statements and language paradoxes, where the same reference may be approached through distinct senses. Furthermore, it implies that language is not just a tool for naming or describing things

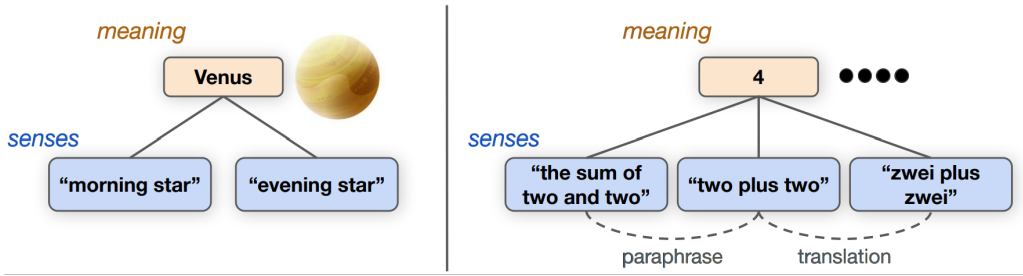


Figure 1 Illustration of the relationship between sense and meaning for the classical Fregean example of “morning star” and “evening star” (left) and for the ADDITION task in our experiments (right).⁴

but serves as a window into how speakers conceptualize and engage with their environment. By distinguishing between sense and reference, Frege provided a framework that could handle the subtleties of language use, such as ambiguity, metaphor, and the context-dependent nature of meaning. This framework, now central to the philosophy of language, underscores that a certain reference can be expressed and conceptualized in different ways.

Relevance to LLMs. Making use of the conceptual groundwork laid by Frege, we posit that true linguistic understanding in LLMs should be evident not just in processing the surface form of text but in grasping the reference that underlies this text. Our methodology leverages this principle by examining the model’s consistency across different expressions that refer to the same underlying meaning. By using the model itself to generate the alternative forms, we ensure that it should—in principle—“know” that they have the same meaning. Taking the example above, if a person is not aware that “evening star” and “morning star” have the same reference (or “two plus two” and “the sum of two and two” for that matter), their response to these two expressions will likely not be the same. However, if a person knows that the two expressions can be used interchangeably, they should be able to answer the same facts about Venus regardless of the choice of expression. By testing across languages and paraphrases, we essentially probe whether LLMs can discern that different textual forms (or senses) may converge on the same reference or meaning, thus revealing a more profound understanding of language beyond mere textual mimicry.

Adopting a loose interpretation of Frege’s notion of “sense”, our multisense consistency method applies to the more general case of different senses as well as the more specific case of different forms expressing the same sense. At the same time, considering translations and paraphrases as potentially involving shifts from one sense to another acknowledges the complexity and richness of language. Different languages and (paraphrased) expressions can present the same referent (or truth value) in diverse ways, capturing the many-sided nature of human thought and culture. Regardless of shifts in sense, the crucial factor is the preservation of the reference—the actual object

⁴ The illustration of Venus was taken from <https://www.universiteitleiden.nl/leven-in-het-heelal/over-leven/venus>.

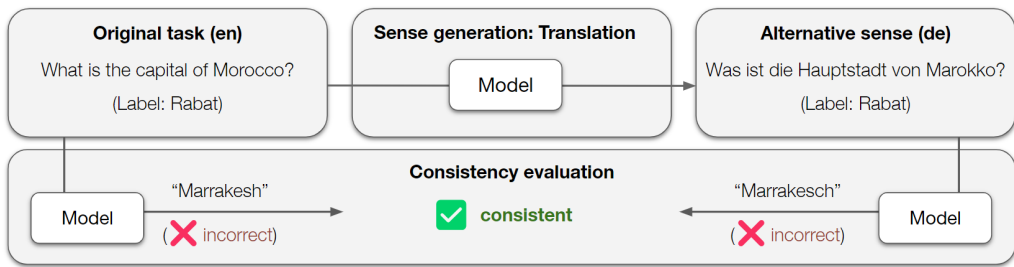


Figure 2

Illustration of the multisense consistency paradigm. We use a model to generate alternative meaning-preserving senses of the original input, and then evaluate whether the same model gives consistent responses to the original input and alternative sense. In this example, the task is to answer a simple factual question, and the model is asked to generate an alternative sense through translation (from English to German). The example illustrates that accuracy and consistency are distinct. Even though the model's responses are incorrect (Marrakesh/Marrakesch instead of Rabat), they are consistent because they refer to the same city.

or truth condition the expressions pertain to. This approach is consistent with Frege's emphasis on the importance of reference in determining the truth value of sentences.

3. Evaluating Multisense Consistency

Concretely speaking, we investigate whether LLMs can be considered to have a form-independent notion of meaning by constructing a test that quantifies whether their understanding is consistent across different expressions with the same meaning. In what follows, we refer to those tuples of expressions as *senses*. Before diving into our experiments, we first give a high-level overview of the main components of this idea. We discuss how we generate different senses (§ 3.1), what data we start from to do so (§ 3.2), and our method for computing multisense consistency (§ 3.3). We provide a schematic in Figure 2.

3.1 Generating Different Senses

The first important component of our paradigm comprises the senses: tuples of expressions that express the same meaning in different manners. Senses could be generated in several ways. In this work, we consider two different methods: translation and paraphrasing, which we will denote by the superscripts T and P , respectively. Importantly, we use the model under investigation to generate meaning-preserving senses, with the idea that *if* the model has a meaning-based understanding and is proficient at generating alternative senses (which we control for in § 7), these senses should have the same meaning according to the model and should thus elicit consistent responses. On the contrary, if a model's meaning is tied to a specific form, there is no reason to assume the response to two senses that have the same meaning should be the same. Thus, using the model to generate the senses controls for subjective meaning-consistency. This approach mirrors Frege's seminal distinction between sense and reference (Frege 1892) emphasizing that true understanding transcends linguistic form to grasp the underlying meaning. Just as Frege illustrated how different expressions can denote the same reference, our paradigm tests whether LLMs can discern and maintain this crucial distinction in a computational context.

3.2 The “Base” Data

The second component of our paradigm is a “base” dataset, to generate different senses from. While the multisense consistency paradigm can in theory be applied to any data, generating senses that have the same meaning may be more or less difficult depending on the initial data and the sense-generation procedure. In this article, we work with two types of datasets. The first type comprises synthetically constructed datasets with simple facts. Because we can be certain that their meanings are consistent across languages, they allow us to test form-independent meaning in a very controlled way. We describe this data as well as our experiments with this data in § 5. Secondly, we consider benchmarks commonly used to evaluate understanding in LLMs. Specifically, we include four different benchmarks covering four different types of NLU tasks: PAWS (Zhang, Baldridge, and He 2019) for paraphrase detection, the English portion of XNLI (Conneau et al. 2018) for natural language inference, COPA (Roemmele, Bejan, and Gordon 2011) for commonsense (causal) reasoning, and Belebele (Bandarkar et al. 2023) for reading comprehension. We describe this data as well as our experiments with this data in § 6.

3.3 Measuring Self-consistency

Lastly, given two senses with the same meaning and two model responses to those senses, we need to define when those two responses are considered to be the same. In other words, we need to specify a method to compute consistency. Consistency is distinct from accuracy or other performance metrics, in that the model’s responses to one sense are evaluated against its responses to the other sense, rather than the ground truth (see Figure 2). Whether responses count as consistent depends both on the task and the way that different senses are generated. For instance, if senses are generated through paraphrasing and the task is a classification task where a model has to pick an answer from a predefined list (e.g., “yes”/“no”), exact match is a good candidate to quantify consistency. If senses are generated through translation, however, model answers will likely be given in different languages, and may look completely different but still share a meaning (e.g., “yes” in English, “ja” in German). In that case, a more custom consistency function is required to judge consistency across senses. For open-ended generation tasks, it can be complicated to define consistency. In such cases, one option is to ask the model itself to judge whether its two answers have the same meaning. In our experiments, we use different methods to evaluate consistency, which we elaborate upon in the respective sections.

3.4 Summary of the Procedure

Overall, our procedure can be summarized as follows. Given a model \mathcal{M} and a task \mathcal{T} , which consists of datapoints $\mathcal{T} = \{x_1, \dots, x_n\}$,

1. Collect the model’s responses on \mathcal{T} : $R = (r_1, \dots, r_n)$, with $r_i = \mathcal{M}(x_i)$.
2. Use the model to generate an alternative sense \mathcal{T}^* of the task, using a specific prompt p : $\mathcal{T}^* = \{x_1^*, \dots, x_n^*\}$, with $x_i^* = \mathcal{M}(p, x_i)$.

3. Collect the model's responses on \mathcal{T}^* : $R^* = (r_1^*, \dots, r_n^*)$, with $r_i^* = \mathcal{M}(x_i^*)$.
4. Calculate the consistency between R and R^* according to some function: $C(R, R^*) = \frac{1}{n} \sum_{i=1}^n f(r_i, r_i^*)$.

The resulting consistency value C expresses multisense consistency.

4. Experimental Details

Before coming to our experiments, we provide some basic details about the setup that all experiments share.

4.1 Model

We investigate GPT-3.5-TURBO-0613, a specific snapshot of GPT-3.5-TURBO from 13 June 2023. We use the default parameters but set the temperature to 0.2. The sampling temperature can be chosen between 0 and 2, and 0.2 is considered a low value, leading to more deterministic and focused output (see also the OpenAI API documentation⁵). In our case, a small temperature yields model responses that closely match the template answers for benchmarking, as well as model translations that closely preserve the meaning of the source sentences.

4.2 Senses

In all our experiments, our starting point is an English dataset, which we denote with *en*. We consider model-generated paraphrases of that data and model-generated translations into other languages. For some datasets, we also have external translations, which we use for saliency checks and comparisons. Target languages include German (*de*), Italian (*it*), Dutch (*nl*), and Swedish (*sv*). We use the current common crawl statistics⁶ to compute an estimate of how low- or high-resource these languages are in Web-based corpora. Of this corpus, English constitutes 46% of the data, German 5.8%, Italian 2.7%, Dutch 2.2%, and Swedish 0.7%. We assume that the GPT-3.5 training data qualitatively follows a similar pattern for these languages, from higher- to lower-resource. The multisense evaluation method only works if the model is able to accurately paraphrase and translate the inputs. Therefore, we do not include even-lower-resource languages. With our selection of languages, we aim to cover some range in the amount of training data without compromising translation quality.

4.3 Same-sense Baseline

We report multisense consistency next to a same-sense baseline consistency. The baseline consistency is the consistency between two generations with the exact same English input (*id*). In other words, the two inputs underlying the baseline consistency do not even differ in form. Differences in model responses on these inputs can thus be attributed to inherent model stochasticity (possible because of the non-zero sampling temperature). The baseline consistency therefore serves as a reference, which can be

⁵ <https://platform.openai.com/docs/api-reference/chat/create>.

⁶ <https://commoncrawl.github.io/cc-crawl-statistics/plots/languages>, CC-MAIN-2023-40.

used to estimate the degree to which inconsistencies between different senses can be attributed to differences in form rather than such inherent stochasticity.

5. Multilingual Factual Consistency

In our first set of experiments, we test the model’s form-dependency when answering simple questions about facts. To do so, we generate datasets that assess a model’s consistency in representing basic factual information from various knowledge domains. The power of these datasets lies in their simplicity. There is little room for nuances in wording across different senses that could cause the model to assign a different meaning. Factual knowledge—in contrast to more complex aspects such as expressions of sentiment—is easy to keep stable across senses, because the meaning of factual statements collapses to their truth value. To give an 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. Along the same lines, the model should generate consistent responses when asked about the kinds of simple facts considered here. Given that the fact-based questions leave hardly any room for ambiguity, inconsistent responses point straight to a form-dependent “understanding”.

5.1 Methods

Our SIMPLE FACTS dataset consists of five distinct datasets, each containing one or more subtasks.

Dataset Creation. Table 1 provides an overview of the datasets and subtasks, including information on the dataset size and examples. Each dataset comprises a single template with specific content fields masked out. During dataset creation, different entities (names, dates, etc.) are inserted into these fields. For instance, the WRITERS dataset is based on the template “In what year was the writer [WRITER] born?” and in each

Table 1

Simple facts datasets. In this table, we provide the templates we used to generate the SIMPLE FACTS datasets, and the total number of examples in each dataset (N). For each template, we provide an example in which the mask(s) are populated with an example datapoint (in bold) from our datasets.

dataset	subtask	N	template / example
ARITHMETICS	–	500	“What is three hundred seventy-five plus twenty-three ?”
ELEMENTS	FROM-ELEMENT	90	“What is the atomic number of the chemical element He ?”
	FROM-POSITION	90	“What is the atomic number of the chemical element in period 5 and group 7 ?”
OLYMPICS	100M	148	“Who won the gold medal in the men’s 100 meters at the 2000 Summer Olympics?”
	DOWNHILL	117	“Who won the bronze medal in the women’s downhill competition at the 1976 Winter Olympics?”
WRITERS	–	186 × 5 = 930	“In what year was the writer Friedrich Schiller born?”
COMPANIES	–	100 × 5 = 500	“In what city does Airbus SE have its headquarters?”

datapoint, [WRITER] is replaced by the name of a different writer. For both WRITERS and COMPANIES, we ensure—with some simplification—that the writers and companies are evenly distributed over countries in which the languages we consider constitute the dominant language.⁷ More details on each dataset can be found in Appendix A.

Sense Generation. We prompt the model to generate different senses for each (sub)task by asking it to paraphrase or translate the corresponding template. Because only the template changes, we can evaluate the quality of the generated paraphrases and translations by hand. Details on the instructions used for generating different senses can be found in Appendix B and the original instructions and the model’s translations can be found in Appendix C.

Model Instructions. To facilitate the performance and consistency evaluations, we always instruct the model to respond with a single entity (e.g., the name of the athlete for OLYMPICS) or number (e.g., “4754” for ARITHMETICS).⁸ On the ARITHMETICS dataset, the model is further instructed to reply with the numerical answer, even though the two summands are spelled out.

Consistency Evaluation. The SIMPLE FACTS datasets contain a set of correct answers A_d for each datapoint d . For example, the answer sets for COMPANIES cover all variations in city names for the languages we work with (e.g., for the city of Berlin, $A_d = \{\text{“Berlin”, “Berlijn”, “Berlino”}\}$). To give another example, the answer sets for OLYMPICS contain different variations of the athletes’ names (e.g., for the winner of the men’s hundred meters in 1920, $A_d = \{\text{“Charlie Paddock”, “Charles Paddock”, “Charles William Paddock”}\}$) as well as multiple names if there is more than one winner. Model responses are always normalized by lowercasing and removing surrounding white spaces and punctuation. Given the normalized model responses, R and R^* , the consistency

$$C(R, R^*) = \frac{1}{n} \sum_{i=1}^n f(r_i, r_i^*)$$

(see § 3.4, step 4) is calculated as

$$f(r, r^*) = \begin{cases} \mathbb{1}_{r \in A \ \& \ r^* \in A}, & \text{if } \exists A : r \in A \text{ or } r^* \in A, \\ \mathbb{1}_{r=r^*}, & \text{else;} \end{cases}$$

where A is a set of possible answers for a datapoint in \mathcal{T} (which are the same as the answer sets for \mathcal{T}^*) and $\mathbb{1}$ is the indicator function. In other words, if an answer set is available that contains the model’s response r or r^* , both of the responses have to be in that set to be consistent. If no such set exists, consistency is approximated by exact match.

⁷ The languages we consider are spoken in different countries, but we tend to focus on one country each. For example, for COMPANIES, we consider an equal amount of US, German, Italian, Dutch, and Swedish companies, establishing a rough correspondence between prompt languages and factual information.

⁸ We double-checked if the model sometimes indicates that it does not know the correct answer, if it is not instructed to respond in these particular ways. On all datasets but WRITERS, it does so very rarely ($\leq 1\%$). Additionally, a comparison of the model’s responses to WRITERS in *en* and *de*^T showed that even if the model indicates that it does not know the correct answer, it does not do so consistently between senses.

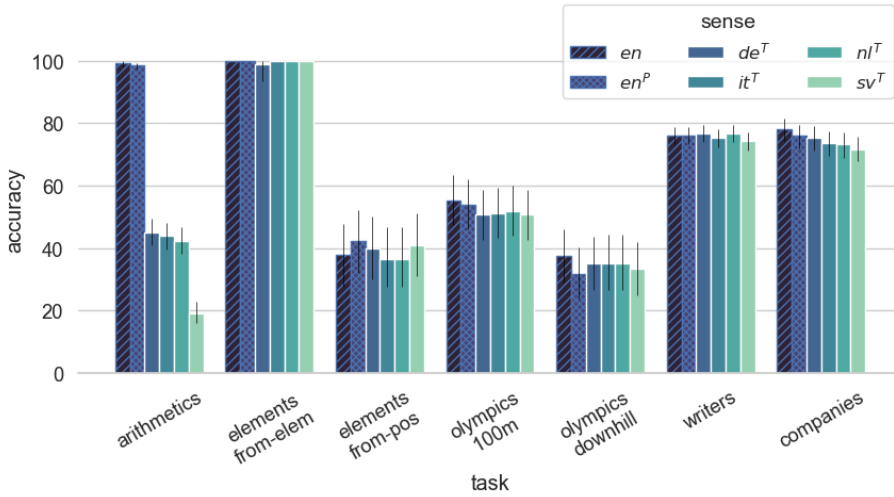


Figure 3 Accuracy (%) for the SIMPLE FACTS datasets, with 95% confidence intervals. Apart from the arithmetics task, the accuracy scores are generally similar across different senses. Numerical scores can be found in Table 7.

5.2 Results

Before studying the model’s consistency, we consider its ability to correctly answer the factual questions. The model’s performance helps us put its consistency into perspective because it sets an upper and a lower bound for the consistency. For instance, if a model reaches maximal performance across senses on some task, it will also be perfectly consistent.

Performance. We compute the accuracy (exact match) scores across datasets and senses.⁹ For some datapoints there are several correct answers; the model’s response counts as correct if it corresponds to one of them. The set of correct answers contains variations in naming (e.g., “Charles Paddock”, “Charlie Paddock”, “Charles William Paddock”), including variations between the languages we use (e.g., “Berlin”, “Berlino”, “Berlijn”). The full list of equivalent answers can be found in our repository.¹⁰ In Figure 3, we can see that the difficulty of the tasks and subtasks varies strongly. For instance, accuracies on ELEMENTS-FROM-ELEMENT are uniformly close to 100% whereas accuracies on OLYMPICS-DOWNHILL are below 38%. However, the model’s performance within subtasks is relatively consistent across the different senses, except for ARITHMETICS, where performance in English is vastly higher than performance for other languages.

The differences in accuracy for ARITHMETICS are striking. We double-checked if the model fails to reply with a numerical answer in some of the languages but this was not the case. In Swedish, the model sometimes responds with the entire equation instead of

⁹ The model is instructed to reply with the correct entity and no additional words. In the large majority of the cases the model follows this instruction, such that there is little difference between counting responses as correct when they *contain* the right answer instead of being an *exact match*. For details, see Appendix E.

¹⁰ https://github.com/facebookresearch/multisense_consistency.

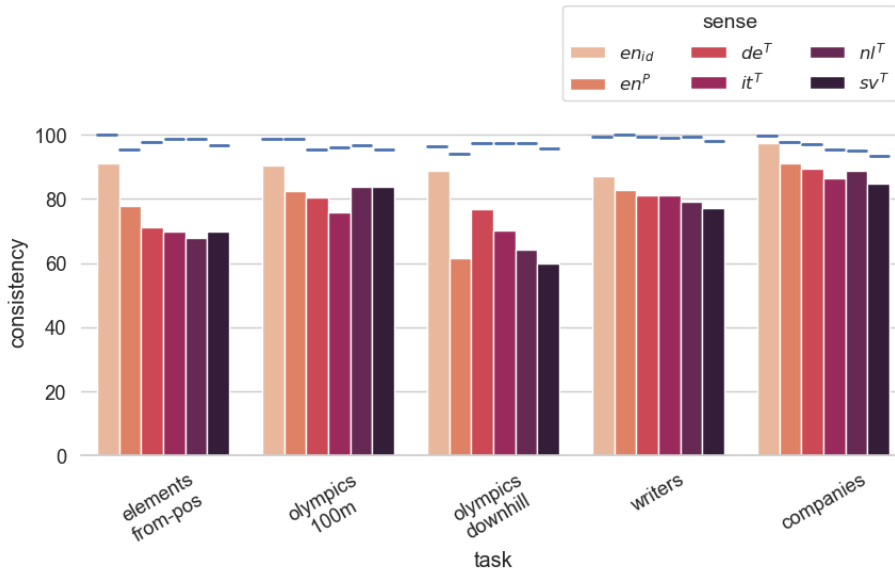


Figure 4

Consistency (%) for the SIMPLE FACTS datasets. None of the senses have a consistency close to the maximum possible given the difference in accuracy between the two senses (indicated by the horizontal blue lines), indicating that the models are inconsistent even beyond those differences. Numerical scores can be found in Table 9.

the correct sum (e.g., “342 + 122 = 464” instead of “464”) but accuracy only increases by 2% when accounting for these cases. It could be that spelled-out numbers are rare in the training corpus such that high versus low-resource effects get magnified, which could explain why there is a big drop from en to $de/it/nl$, and then another one to se .

Consistency. Next, we consider how consistent the model’s representations are across senses. We report the results in Figure 4. Because the generation process is stochastic at non-zero temperature, asking the same question twice may lead to different responses. We exploit this to report also same-sense consistency between two en -runs (denoted with id). Note that if a model has a maximal accuracy on one of the senses, its consistency score equals the accuracy of the other sense, without providing any evidence for form-independent meaning representations. We therefore exclude the ARITHMETICS and ELEMENTS-FROM-ELEM task from our consistency results. More generally, given a difference in accuracy between two senses, $\Delta(\text{Acc})$, the consistency cannot be higher than $1 - \Delta(\text{Acc})$.¹¹ We indicate these upper bounds in the figure with blue lines above each bar. While consistency and accuracy are thus not independent, as long as accuracies are not at 100%, they are clearly distinct. Even if the differences between the accuracies are small, the consistency may vary wildly.

¹¹ For example, if the model is 80% correct on one sense and 60% correct on another sense, the maximal consistency is achieved when the respective overlap between correct and incorrect responses is maximal: The same 60% of the datapoints are correct on both senses, and the same 20% of the datapoints are incorrect on both senses, resulting in $100\% - (80\% - 60\%) = 80\%$ consistency.

In Figure 4, we can see a manifestation of this statement: Although the accuracy scores across senses are all comparable (see Figure 3), there is not a single case where the consistencies are near-maximal. This is remarkable given the simplicity of the tasks and instructions. Even for English paraphrases, consistency can be as low as 61.5% at a 88.9% baseline (see OLYMPICS-DOWNHILL). In this case, almost all inconsistencies arise because the model replies with the names of different athletes, usually winners of other medals in the same competition or winners of other competitions. For example, when asked for the female bronze medallist in 1988, the model gives the correct answer to the original prompt (“Brigitte Oertli”) but replies with the name of the world champion of 1989 to the paraphrased prompt (“Karin Dedler”). More examples can be found in Appendix G. The baseline scores (*id*) show that the inconsistencies are not (primarily) caused by the model assigning equal probabilities to possible answers, leading to different outputs on different senses. While the baseline scores are not maximal, they are much higher than what would be expected in such a case.¹² In other words, most inconsistencies cannot be attributed to the lack of a clear winner, in which case the model would sample from several roughly equally low probabilities.

6. Natural Language Understanding Benchmarks

Our results with the SIMPLE FACTS datasets point to substantial form-dependencies in the LLM’s representation of factual knowledge. Next, we investigate how the model behaves on a set of different NLU tasks in which meaning and task understanding are more complex than merely reiterating knowledge.

6.1 Methods

For our continued evaluation of consistency across more complicated scenarios, we consider four different benchmarks covering four different types of NLU tasks.

Datasets. First, we consider PAWS (Zhang, Baldridge, and He 2019), a paraphrase dataset where sentence pairs were adversarially created by word-swapping, resulting in negative pairs that have clearly distinct meanings but high lexical overlap (see, for instance, the example in Table 2). Second, we consider (mainly the English portion of) XNLI (Conneau et al. 2018), a language inference task containing sentence pairs that either entail or contradict each other, or have a neutral relationship. Third, we use COPA (Roemmele, Bejan, and Gordon 2011), a dataset containing tuples of a premise and two alternatives, where the task is to select the alternative that more plausibly has a relation with the premise. Lastly, Belebele (Bandarkar et al. 2023) is a reading comprehension task with multiple choice questions where an answer should be given based on a text passage. We run all our evaluations on the test split of the respective datasets. Note that all tasks correspond to classification problems; we standardize the model’s responses and map them onto the corresponding class labels. Furthermore, for some of the languages we consider, parallel data for the tasks exist either in the original corpus (in the case of Belebele and XNLI)

¹² The simple facts datasets are open QA tasks. When the model is asked for an entity (e.g., a city), it can potentially choose its answer from the set of all entities in the relevant category (e.g., all cities). If the model assigned similar probabilities to many answers in this set, it would likely be inconsistent whenever it is incorrect. In that case, the baseline consistency would be less than or at the maximum (when there is a perfect overlap between correct responses) equal to the model’s accuracy on *en*.

Table 2

Instructions and example inputs for the benchmark data. We provide an example for each benchmark dataset in our experiments. The example input is given in bold, the instructions in normal font.

dataset	template / example
paws	Do the following two 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 reply with a single word, either “yes” or “no”.
xnli (en)	Given the following premise and hypothesis, please identify whether the premise entails the hypothesis, contradicts the hypothesis, or neither of the two. Premise: “Well, I wasn’t even thinking about that, but I was so frustrated, and, I ended up talking to him again.” Hypothesis: “I haven’t spoken to him again.” Please reply with a single word: “entailment” if the premise entails the hypothesis, “contradiction” if the premise contradicts the hypothesis, and “neutral” if the premise neither entails nor contradicts the hypothesis.
copa	Given the following premise, which of the two alternatives is more plausible? Premise: “The item was packaged in bubble wrap.” Alternative 1: “It was fragile.” Alternative 2: “It was small.” Please answer with a single word: “Alternative-1” if alternative 1 is more plausible and “Alternative-2” if alternative 2 is more plausible.
belebele	Virtually all computers in use today are based on the manipulation of information which is coded in the form of binary numbers. A binary number can have only one of two values, i.e., 0 or 1, and these numbers are referred to as binary digits - or bits, to use computer jargon. According to the passage, which of the following is an example of a five bit binary number? Option A: 1010 Option B: 12001 Option C: 10010 Option D: 110101 Please reply with “A”, “B”, “C”, or “D” to indicate the correct answer. Your reply should be a single letter and should not contain any additional words.

or in multilingual versions of the corpus (PAWS-X and XCOPIA [Yang et al. 2019; Ponti et al. 2020, respectively]). While our paradigm does not require parallel multilingual datasets, we use them in § 7 to run additional analyses.

Sense Generation and Model Instructions. For each dataset, we write an English instruction which together with the task input data forms the prompt presented to the model (see Table 2). We ask the model to paraphrase and translate the instruction and the input data separately, and we recombine the two outputs to generate the alternative sense. Individual datapoints in the benchmarks comprise several components, for example, a premise and a hypothesis in the case of XNLI. We provide all these components within the same prompt when asking the model to paraphrase or translate. Combining the components for each datapoint has the advantage that the resulting paraphrases/translations will be more consistent (e.g., the model will resolve ambiguities or make certain translation choices in the same way across components). We compared this method to paraphrasing/translating each component separately, and it resulted in slightly higher task accuracies on the generated senses. More details on the sense generation can be found in Appendix B, and the model’s translations and paraphrases of the instructions can be found in Appendix C.

Consistency Evaluation. The model’s responses for the benchmark data are standardized and mapped onto the corresponding class label. Standardization involves lower-casing,

removing surrounding whitespaces and punctuation. The model generally conforms to the instruction and responds only with the correct answer. However, if necessary, additional words are also removed (automatically). For example, if the model replies “The answer is ‘yes’.” in English and “Ja” in German, both responses will be standardized (“yes”, “ja”) and then mapped onto the corresponding class labels, $l(r) = 1$ and $l(r^*) = 1$. Consistency (see § 3.4, step 4) is then calculated as

$$C(R, R^*) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{l(r_i)=l(r_i^*)}$$

where $\mathbb{1}$ is again the indicator function.

6.2 Results

We discuss our results, again starting with accuracy and then continuing with consistency scores.

Performance. We plot the accuracy scores in Figure 5; horizontal blue lines indicate chance accuracy. We excluded the results for paraphrases of Belebele, because the model consistently failed to paraphrase this task—sometimes it ignored the text passage and sometimes it answered the question instead of paraphrasing. The accuracies for COPA and Belebele are relatively high ($\geq 79\%$) across senses, followed by PAWS and then XNLI. Performance on Belebele is particularly high, considering that there are four answer possibilities, compared to three for XNLI, and two for COPA and PAWS. Performance on XNLI is particularly low, raising the question of whether this task is

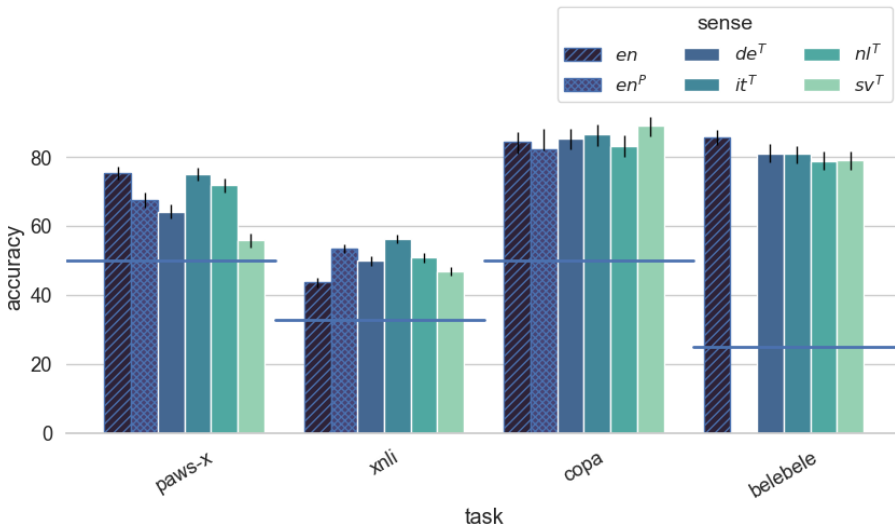


Figure 5 Accuracy (%) for the benchmark datasets, with 95% confidence intervals. For Belebele, we have no en^P score, because the model did not provide useable paraphrases. Horizontal lines indicate chance accuracy. Numerical scores can be found in Table 8.

perhaps simply not suited for zero-shot evaluation. Looking into the task in more detail, we suggest that the task may be very prompt-sensitive, with different preferences in different model versions. For instance, we observed much higher performances with an older GPT-3.5-TURBO snapshot as well as GPT-4 on this task. This may indicate that XNLI is a task that is particularly form-tied, making it an interesting candidate for evaluating multisense consistency. Overall, we observe that for each task, performance can vary strongly across senses, with up to 19.7% points on PAWS and up to 12.7% points on XNLI.

Consistency. Next, we look at the consistency. We plot the results in Figure 6, again against the *en* same-sense baseline (*id*). Horizontal blue lines indicate the maximal possible consistency when accounting for differences in accuracy. Overall, model consistency is much lower on some tasks than on others. With regard to the accuracy scores above, the model tends to be more consistent on tasks it can solve well. For example, consistency is as low as 51.2% on the German translation of XNLI whereas it is above 84% for all task versions of COPA. This is not entirely unsurprising because the model can also be consistent when it has a form-dependent task understanding but has learned to generate the correct response for each form (separately). If the model makes a mistake, however, it is much less likely that it will generate the same mistake in another form, if the generated responses are form-dependent. The fact that the model overall has a higher consistency on tasks with higher accuracy thus suggests that at least part of its consistency is not due to a form-independent understanding of meaning. We further investigate this difference in § 7.3. We also see that consistency can vary strongly between senses, ranging from 51.2% to 82.8% on XNLI, and 67.9% to 82.4% on PAWS. A comparison against the baseline scores confirms that inconsistencies go

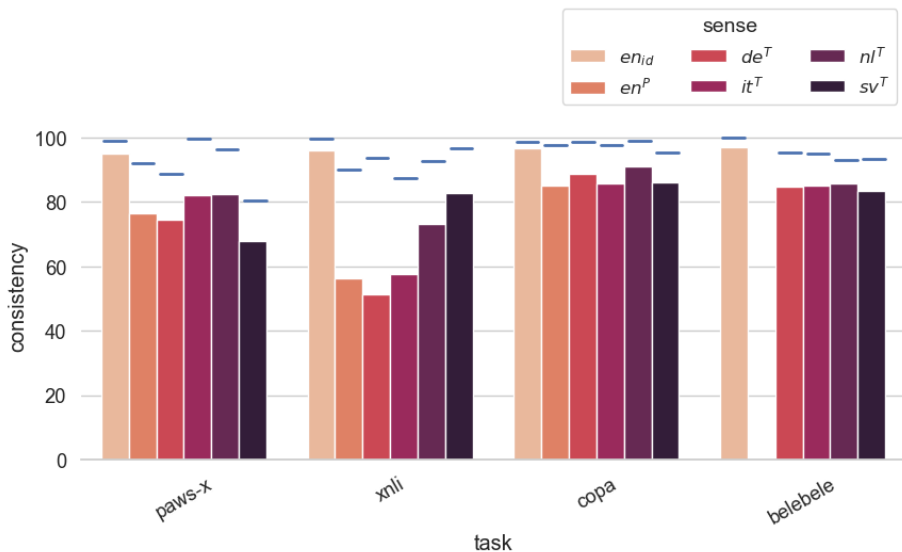


Figure 6 Consistency scores (%) for the benchmark datasets. None of the consistencies between original and alternative sense are close to the maximum possible given the difference in accuracy between the two senses (indicated by the horizontal blue lines), indicating that the models are inconsistent even beyond those differences. Numerical scores can be found in Table 10.

beyond stochasticity inherent to the model. Considering the results for both SIMPLE FACTS and benchmark data, it seems that accuracy and consistency tend to decrease slightly from higher- to lower-resource languages. Given that this effect is small, most of the inconsistencies are likely not driven by the choice of senses or the process of generating these senses with the model (see § 7.1 for a detailed analysis). In sum, the systematic benchmark evaluation provides evidence across larger and more diverse datasets than the SIMPLE FACTS evaluation. The results are in line with our earlier observation that GPT-3.5 is not very self-consistent.

7. Analysis

The results in the previous sections suggest that the meaning representations of the model we investigate are strongly tied to form. The main evidence for that is the model's inconsistencies across senses. In this section, we aim to better understand when and why inconsistencies arise. More specifically:

1. We evaluate whether inconsistencies stem from the model's inability to generate meaning-preserving senses, that is, it does not have the ability to adequately paraphrase or translate (§ 7.1).
2. We evaluate whether the model is inconsistent in its task interpretation, in its task execution, or in both (§ 7.2).
3. We evaluate consistency conditioned on correctness of the model's responses, because—as we argue below—consistently incorrect responses provide stronger evidence for a form-independent task understanding than consistently correct ones (§ 7.3).
4. We study if there is a connection between requested information and prompt language that could provide direct evidence for form dependency as a source of inconsistency (§ 7.4).

We present these analyses for the simple facts, the benchmarks, or both, as appropriate.

7.1 Quality of Alternative Senses

The metric we propose conflates task understanding of the “primary” sense and ability to generate different senses: If a model is not able to generate adequate translations or paraphrases, this may give rise to inconsistencies even if it has a form-independent understanding of meaning. While both are important qualities, and the metric favors models that do well across the board, it makes sense to consider the two parts separately as well. Differences in task understanding for high-quality senses point to a form-dependent task understanding whereas, as pointed out earlier, a failure to translate or paraphrase may not. For example, while a poor task understanding can lead to a bad translation, a poor translation might also arise from a poor command of the target or source language, or an inability to translate. To examine if inconsistencies are due to one of the latter causes, we investigate the quality of the paraphrases and translations.

Table 3

Translation quality. We consider the quality of the translations of the input data to different senses, according to different commonly used metrics. All scores are comparatively high, suggesting that the model’s inconsistencies are not driven by an inability to translate.

		bleu	rouge1	rouge2	rouge-l	comet-22
paws	de ^T	57.5	0.81	0.65	0.77	0.85
xnli	de ^T	41.9	0.69	0.49	0.66	0.84
copa	it ^T	40.9	0.66	0.45	0.64	0.86
belebele	de ^T	41.1	0.69	0.46	0.63	0.84
	it ^T	38.1	0.69	0.44	0.61	0.85
	nl ^T	34.3	0.68	0.40	0.57	0.85
	sv ^T	44.0	0.73	0.53	0.68	0.86

Translation and Paraphrase Quality. First, we check the quality of the translations and paraphrases for both SIMPLE FACTS and benchmark data. To evaluate the instruction data, we ask native speakers of each language, who are also fluent in English, to verify whether the paraphrases and translations are correct and meaning-preserving. For the SIMPLE FACTS data, we consider the templates; for the benchmark data, we consider the task instructions (see Appendix C for a full list of these). For both types of data, the instructions were largely judged to be grammatically correct and meaning-preserving, although they tend to stay relatively close to the English original, such that a native speaker might prefer a slightly different wording.

Next, we automatically evaluate whether the numbers for the ARITHMETICS task are translated correctly. Each datapoint consists of a pair of numbers (see Table 1) and the translation counts as correct if both numbers are translated correctly. We find that the translations are highly accurate for German (99.6%) and Dutch (99.4%), but less so for Italian (89.2%) and Swedish (81.0%). Still, the proportion of wrong translations is significantly smaller than the proportion of inconsistencies across all languages, and can thus explain only a small part of the inconsistencies for that task.

For the benchmark data, we further evaluate the quality of the translations of the *task input data*, by comparing them to reference data, available either in the benchmark itself (in the case of Belebele) or in the multilingual benchmark versions we use. We report BLEU (Papineni et al. 2002), ROUGE (Lin 2004), and COMET-22 (Rei et al. 2022) scores, all commonly adopted measures of translation quality, in Table 3. All metrics indicate that the model’s translations are of high quality across tasks and languages. The high scores suggest that, for most of the considered source-target language combinations, inconsistencies can largely not be ascribed to changes in meaning induced by the translation.

Translation Quality vs Consistency. To investigate the relationship between translation quality and consistency in more detail, we run several follow-up analyses. First, we calculate the Pearson correlation between consistency and COMET scores. The correlation for XNLI is negative ($\rho = -0.06$), and for COPA ($\rho = 0.07$) and PAWS ($\rho = 0.11$) it is relatively small. For Belebele the correlations are also rather small (ρ between 0.08–0.13), with a somewhat higher value for Swedish ($\rho = 0.21$). Second, we evaluate the consistencies for a subset of the best translations, considering only datapoints with COMET scores greater than 0.80. Relative to the original scores across all datapoints,

consistency scores change between -2.7 and 2.0 percentage points across datasets and languages; based on a two-sided t-test this difference is not significant ($p > 0.9$). Finally, we evaluate the model’s consistency when replacing the self-translated input data with the ground truth references for each language. When reference data is available, we pair the model’s translation of the instruction with the benchmark data for the corresponding target language (e.g., de^T instruction and de input data). It turns out that the model’s consistency *decreases* in six out of seven cases (by up to -5.2%) and increases in one case (by 0.7%). In other words, the model tends to be more consistent when the alternative sense is self-generated. This result also highlights the importance of using the model’s own translations and paraphrases: Despite imperfect translations and paraphrases, the model treats self-generated senses as slightly more meaning-equivalent than externally generated ones. These additional analyses show that translation quality can affect consistency but is not a major driver of the inconsistencies observed in our experiments.

7.2 Interpretation versus Execution

Next, we investigate if, when a model is inconsistent across senses, this inconsistency stems from an inadequate understanding of what the task is or from an inadequate execution of that task in that specific language.¹³ To exemplify this, compare the scenario in which you are asked to judge whether one English sentence implies the other, but the request is made in a language that you do not have a great command of with the scenario in which the question is asked in English, but the sentences to be judged are in a language you do not understand well. Because the SIMPLE FACTS does not have separate instruction and task data, we analyze this only for the benchmark data.

To disentangle the impact of changing the sense of the task instruction and the task input data, we run an ablation experiment. Specifically, we assess the model’s consistency when paraphrasing/translating only the instruction while keeping the original input data (condition *I*), as well as its consistency when paraphrasing/translating only the input data while keeping the original instruction (condition *X*). The resulting consistency scores are displayed in Table 4 and the corresponding accuracies in Appendix H. Neither consistencies for translating only the instructions nor those for translating only the input data are at their maximum, indicating that the model is inconsistent in both interpretation and execution. Whether inconsistencies in execution or interpretation are more pronounced depends largely on the task. In particular for XNLI, where the instruction is very complex, consistencies are higher when using the same instruction compared to using the same input data. For tasks with comparatively simple instructions, the pattern is at least partially reversed. Consistency is always lower when using the same instruction but different input data for Belebele and COPA, and in some cases also for PAWS. When paraphrasing/translating both instructions and input data (cf. Figure 6 / Table 10) consistencies are mostly lower than for either ablation. Thus, inconsistencies seem to be driven by differences in both task interpretation and execution. Differences in execution are more pronounced unless the task is difficult to interpret.

¹³ This distinction is related to the fact that we evaluate the model’s understanding with different *tasks*.

Based on Frege’s observation that different senses can have the same meaning, we need to create an interface that allows us to test whether LLMs actually assign the same meaning to different senses. In our case, this interface consists of the task that the model is supposed to carry out on a given input. Thus, the analysis can also be considered a way to disentangle the model’s meaning understanding of the input sentences from its meaning understanding of the instructions.

Table 4

Consistency scores (%) for the ablation experiments. We analyze whether consistencies mainly arise from differences in task interpretation or execution, by considering ablations in which we translate/paraphrase only the instruction (columns *I*) or only the input data (columns *X*). Where inconsistencies are more pronounced depends largely on the task. Mostly for XNLI, interpreting the (comparatively) complex instruction appears to be more challenging than understanding the sentence.

	paws		xnli		copa		belebele	
	<i>I</i>	<i>X</i>	<i>I</i>	<i>X</i>	<i>I</i>	<i>X</i>	<i>I</i>	<i>X</i>
en ^P	89.5	78.4	64.0	86.7	90.2	87.0	–	94.4
de ^T	77.8	81.1	57.9	88.5	94.0	88.6	94.1	84.7
it ^T	91.2	82.0	60.9	88.9	91.8	86.2	94.4	83.3
nl ^T	86.4	83.3	77.9	88.6	93.2	90.0	94.1	86.7
sv ^T	72.7	80.3	82.4	88.6	91.0	87.4	94.2	84.9

7.3 Consistency vs. Correctness

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 consistency of meanings beyond form. If a model gives consistently correct answers for an example, it is possible that it has inferred those correct answers independently from the data for the respective languages. In that case, consistency does thus not necessarily point to a form-independent understanding of the particular question. This is much less likely the case for *incorrectly consistent* examples, as it would require that the data the model was trained on contained the same error for both languages. Being consistently incorrect across two examples thus 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.

Figure 7 shows the consistency scores conditioned on whether the model was correct on the source task (*en*), for both the SIMPLE FACTS (left) and the benchmark data (right). The scores are averaged across senses and the *id* baseline is given by the dotted lines. We can see that the model is always more consistent on correct responses than on incorrect responses, suggesting that the responses are—at least in some cases—consistent simply because they are independently correct in both languages. Given that the difference between the two conditions (correct vs. incorrect) is more pronounced for different senses than the same-sense baseline, it cannot solely be attributed to stochasticity for cases where the model’s distribution is relatively flat among the highest scoring answers but it can nevertheless not answer “I don’t know”. In conclusion, not only when answering simple factual questions, but also across a range of NLU tasks, the model seems to infer a significant amount of its responses separately for each sense.

7.4 Direct Evidence for Form-dependency

The analyses above all provide converging but indirect evidence for form dependencies in the model’s understanding. In this final analysis, we aim to establish a direct connection between the type of information the model is asked about and the form of the question. It is plausible that certain information is more often presented to the model

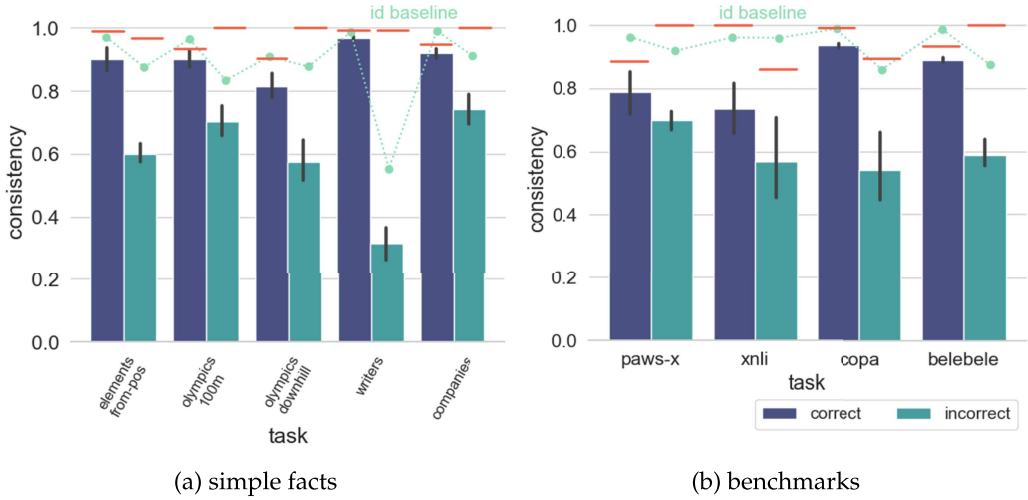


Figure 7 Consistency scores conditioned on correctness. Error bars indicate 95% confidence intervals. Examples that are consistent and *incorrect* provide stronger evidence for a form-independent meaning understanding than *consistent correct* examples, because it is less likely that incorrect information was inferred independently. The large differences between consistent correct and consistent incorrect in this plot thus indicate that—likely—some of the consistent correct examples were correct independently. As in the previous plots, upper-bound consistency based on the individual sense accuracies is given by horizontal lines. The dotted line indicates the *id* baseline (two runs in English).

in a certain form during training. For instance, information about Italian companies likely occurs more often in Italian text than in Swedish text and vice versa. If acquired meanings transcended the form they were acquired in, this should not matter: Once acquired, a fact should be accessible in any language mastered by the model. Thus, if a model scores comparatively better in the language that is related to the information requested, this points to a form-dependent question understanding. To test this hypothesis we exploit the controlled structure of the WRITERS and COMPANIES datasets. Both datasets comprise five subsets of equal size (see Table 1). Each subset contains facts that can be considered somewhat specific to one of our test languages, establishing two conditions of matching or mismatching prompt language and target information. Accordingly, we investigate whether prompting the model in the information-specific language yields higher accuracy compared to prompting it in another language.

In Figure 8, we plot (i) the absolute difference between the accuracy of the model when prompted in the language *matching* the data subset (e.g., asked about Dutch writers in Dutch) and the overall average accuracy for all languages on that subset, and (ii) the absolute difference between the accuracy of the model when prompted on *mismatched* subsets (e.g., asked about non-Dutch writers in Dutch) and the overall average for all languages for that same group of subsets. With the exception of Italian on the WRITERS task, the model is always comparatively (and sometimes absolutely) better on the language-matched subsets (plain blue bars) than on the mismatched subsets (hatched turquoise bars). For example, when prompting the model in Dutch on the Dutch WRITERS subset, accuracy is almost 4% higher compared to the average accuracy for this subset across prompts (including *nl*). A two-sided t-test between the deviations from the mean for cases with matching versus mismatching information and prompt

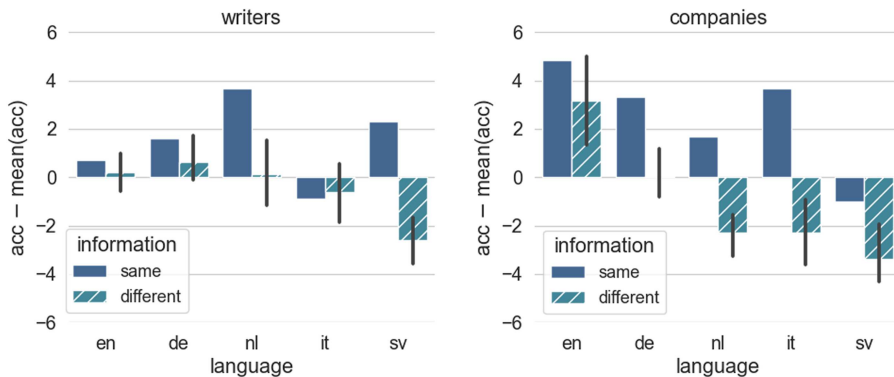


Figure 8

Language-dependent knowledge for the SIMPLE FACTS dataset. Error bars indicate 95% confidence intervals. For each language, we compute how its accuracy when asked about information matching that language compares to its accuracy when asked about information not matching that language (e.g., asking about Dutch writers in Dutch vs in Swedish), compared to the overall averages for those groups. Generally, the model has higher accuracy when the prompt language and requested information pertain to the same country (plain bars) than when it is asked about non-matching information (hatched bars).

languages is highly significant ($p = 0.001$). While this analysis covers only two datasets, the results provide direct, positive evidence for a form-dependent task understanding.

8. Related Work

In this work, we considered LLMs as explanatory models of meaning. Here, we discuss work related to the various aspects of our study. In particular, we discuss studies that have used LMs as explanatory models of language or language processing (§ 8.1); work that explicitly discusses form and meaning in LLMs (§ 8.2); and studies that have involved (multilingual) consistency in LLM evaluation protocols (§ 8.3).

8.1 LLMs as Explanatory Models

Despite the many differences between biological and artificial neural networks, the latter have been extensively investigated as explanatory models to further our understanding of human cognition, primarily in the domains of vision and natural language. In the field of natural language processing, these endeavors have spanned a large range of phenomena and questions. As some understanding of how neural networks behave or what they represent is a prerequisite for using them as explanatory models, such studies often interweave various interpretability methods with (psycho)linguistic theory. Here, we focus specifically on studies that use (modern) LMs and make an explicit attempt to reconnect their findings with human processing, linguistics, or cognition.¹⁴

¹⁴ There also exists quite some literature that aims to directly draw connections between the representations in neural networks and in the human brain. We consider that beyond the scope of this article, and will not further discuss it.

Nested Hierarchical Processing. One subject elaborately explored in linguistically inspired studies of LMs is their ability to process hierarchical structure in language. Starting from the work of Linzen, Dupoux, and Goldberg in 2016, a wave of studies have considered long-distance subject-verb agreement as a proxy for this ability (e.g., Gulordava et al. 2018; Giulianelli et al. 2018). The most clear-cut example of using subject-verb agreement in LLMs in an explanatory fashion is the series presented by Lakretz et al. (2019, 2021) and Baroni (2023), who used a psycholinguistic experiment to assess whether a mechanism for processing nested dependencies they found in LMs may be deployed by humans as well.

Inflectional Morphology. Another topic that has long been used as a testing ground for answering questions about linguistic generalization in humans and the viability of neural networks as models of cognition is inflectional morphology. The amount of literature on this topic is too vast to discuss in detail in this work; for a concise summary, we refer to the related work section of Dankers et al. (2021).

Processing Difficulty. Lastly, starting from Elman (1990), there is a long tradition of trying to link the performance of—mostly recurrent—neural networks to human processing difficulty (Christiansen and Chater 1999; Frank and Bod 2011; Futrell and Levy 2017, i.a.). Several such studies have considered surprisal (i.e., predictive difficulty) to study hypotheses regarding the role of retrieval and prediction in defining human processing difficulty. Among others, Wilcox et al. (2020), Van Schijndel and Linzen (2021), and Huang et al. (2023) show that surprisal in neural networks often differs strongly from human reading-time data, and that predictive difficulty is thus likely insufficient to explain processing difficulty. In a similar vein, several others have considered how LMs process garden path sentences (e.g., Ulmer, Hupkes, and Bruni 2019; Van Schijndel and Linzen 2018, 2021; Arehalli, Dillon, and Linzen 2022)—in psycholinguistics often studied to investigate if humans maintain multiple parses at once. Ryu and Lewis (2021), and recently Timkey and Linzen (2023), focus more on the retrieval side, and show positive results concerning the similarity of attention head behavior with effects observed in human experiments.

8.2 Form and Meaning in LLMs

Currently, the degree to which LLMs can and do have meaning-based, rather than mere form-based, knowledge and understanding is widely debated (e.g., Mitchell and Krakauer 2023; Raji et al. 2021). To begin with, there is no agreement in the community on whether LLMs can *in principle* learn meaning from text. While some argue that meaning cannot be learned from form alone (e.g., Bender and Koller 2020) others disagree or argue that the training signal for some LLMs goes beyond form (e.g., Piantadosi and Hill 2022; Mollo and Millièrè 2023; Pavlick 2023; Mandelkern and Linzen 2023). Importantly, current NLU benchmarks do not provide the means to disentangle the roles of form and meaning (e.g., Heineman 2023). If a model achieves a high score on a benchmark, it is not clear whether the model relies on specific lexical patterns or general principles when performing the task (e.g., Ray Choudhury, Rogers, and Augenstein 2022). In some cases, LLMs have been found to exploit spurious statistical patterns or rely on information memorized from the training, rather than a flexible and generalizable task understanding (e.g., Geva, Goldberg, and Berant 2019; McCoy, Pavlick, and Linzen 2019; McKenna et al. 2023). Adversarial datasets (e.g., Nie et al. 2020) are designed

precisely to expose such shortcut learning behaviors (for an overview of shortcut learning, see Du et al. 2023). Despite this uncertainty, it is common to construct “understanding” benchmarks without considering this question. Instead, “understanding” is typically reduced to generalization across many different tasks (e.g., Wang et al. 2018, 2019; Hendrycks et al. 2021). An evaluation of consistency can also be considered a generalization evaluation.¹⁵ However, by evaluating a model across different senses with the same meaning (i.e., different versions of the same task) rather than different meanings (i.e., different tasks), it is possible to uncover form dependencies that stand in contrast to a human-like task understanding.

8.3 Consistency in LLMs

Various studies have shown that inconsistencies are common in LLMs (and have suggested methods for improving consistency, which is not our focus). To begin with, investigations of model robustness have revealed that even minor (meaning-preserving) perturbations of the model input can strongly affect the generated output (e.g., Chakraborty, Kulkarni, and Li 2023; Weber, Bruni, and Hupkes 2023; Wang et al. 2023; Mizrahi et al. 2023; Podkorytov, Biś, and Liu 2021). Other than that, studies are mostly concerned with self-consistency in natural language inference (NLI) (e.g., Minervini and Riedel 2018; Wang, Sun, and Xing 2019; 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, Choi, and Durrett 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 of 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). Their evaluation of *Macaw* (Tafjord and Clark 2021), a fine-tuned T5 model, revealed significant inconsistencies in the model’s beliefs. In the same vein, various GPT models fail to generalize from statements of the form “A is B” to “B is A” (Berglund et al. 2023). More similar to our work, Elazar et al. (2021) studied whether factual knowledge in masked LMs is invariant to paraphrasing. To this end, they created PARAREL, a dataset containing cloze-style English paraphrases (e.g., *Homeland originally aired on [MASK], Homeland premiered on [MASK]*), which was, for example, recently used to reveal inconsistencies across various LLaMA (Touvron et al. 2023) and Atlas (Izacard et al. 2023) models (Hagström et al. 2023). In the studies mentioned here, 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, Kwon, and Lukasiewicz 2022) is a benchmark for evaluating these two types of consistency (logical and semantic) across various tasks. For each task, the benchmark provides an alternative version (e.g., for semantic consistency the inputs are paraphrased) to compare the model’s answers across task instances. 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). Except for Jang and Lukasiewicz (2023) and our own preliminary work (Ohmer, Bruni, and Hupkes 2023), consistency usually relies on different forms of the same meaning that are generated externally from the model. We focus on true *self*-consistency, where alternative senses are generated by the model under investigation,

¹⁵ See Appendix I for a GenBench eval card (Hupkes et al. 2023) that classifies our work in the context of generalization research.

to ensure that the model—if it can assign meaning—should assign the same meaning to the original and the derived sense.

Multilingual Consistency. Given that we generate different forms through translation, our approach is related to multilingual model evaluation. Multilingual benchmarks are usually generated from existing benchmarks through expert translations (for a more expansive overview, we refer to Hupkes et al. 2023, Appendix D). Prominent examples include PAWS-X (Yang et al. 2019), XCOPA (Ponti et al. 2020), and XNLI (Conneau et al. 2018). Furthermore, multilingual tasks have been combined to form multilingual multitask benchmarks (e.g., Hu et al. 2020; Ruder et al. 2021; Liang et al. 2020). All of these benchmarks reveal language-dependent differences in performance for current multilingual LLMs, which indicates that the models’ responses to the original and the translated task versions are not perfectly consistent. Recently, Qi, Fernández, and Bisazza (2023) combined consistency and multilingual evaluation by introducing a ranking-based consistency metric for evaluating knowledge consistency across languages independently from accuracy. They found that consistency correlates strongly with the sub-word vocabulary overlap between two languages, suggesting that knowledge transfer between languages relies on shallow features rather than a true understanding. In contrast to existing multilingual evaluation approaches, we aim to evaluate self-consistency by detecting language-dependent changes in model responses, relying on the model’s own translations.

9. General Discussion

In this article, we proposed a paradigm to investigate whether LLMs acquire form-independent notions of meaning, with the larger aim of assessing the viability of using them as explanatory models to better understand the concept of meaning. In this last section, we summarize the key aspects of our approach and the main findings from our experiments (§ 9.1), discuss the separation of form and meaning in humans versus LLMs in light of our findings (§ 9.2), and revisit the discussion on using LLMs as explanatory models of meaning, specifically considering the role of multisense consistency therein (§ 9.3).

9.1 Summary

Motivated by the successes of LLMs as explanatory models of *form*, we are interested in their potential as explanatory models of *meaning*. Our analysis takes inspiration from philosophy of language. Based on Frege’s distinction between sense and reference, we propose a paradigm to study if LLMs, trained on only forms, possess form-independent notions of meaning. Specifically, we evaluate the self-consistency of a model across different meaning-preserving forms (senses), generated by the model itself. The main idea underpinning this paradigm is that if a model’s understanding extends beyond form, it should produce consistent responses to different senses that express the same meaning—provided it understands the equivalency between these different senses.

Using this paradigm, we investigated the form-dependency of natural language understanding in GPT-3.5, a state-of-the-art language model. We conducted experiments with a novel benchmark with simple factual questions and different NLU benchmarks. The former provides unambiguous evidence of form dependency, while the latter speak to the extent of this form dependency across various NLU tasks. We detected inconsistencies for all tasks, across all generated senses, both in paraphrases and

translations. Our analyses control for explanations other than a form-dependent understanding: Inconsistencies are neither due to inherent stochasticity, nor due to changes in meaning in the sense-generation process. They also help us better understand the nature of the model's inconsistencies, by showing that the model is inconsistent in task interpretation and execution and that the inconsistencies are more pronounced in incorrect examples than in correct examples. These findings indicate that the model infers its responses separately for each sense and highlight the limitations of current LMs in capturing the true nature of meaning.

9.2 Form and Meaning in Humans versus LLMs

Form-independent meaning is critical to human understanding. Many tasks that we encounter share a common abstract structure. In solving familiar and novel tasks we can exploit this structure by accessing the same knowledge, reasoning process, or skill (e.g., Tenenbaum et al. 2011; Barsalou 2005; Gentner and Hoyos 2017). Furthermore, neurological evidence supports that the brain maintains abstract task representations which are used in generalization (e.g., Liu et al. 2019; McKenzie et al. 2014; Badre and Nee 2018; Vaidya et al. 2021). In our implementation, different forms of the same task correspond to different languages or paraphrases. Also for this specific instance, there is evidence for a form-independent understanding in humans. Studies with bilinguals and second-language learners collectively support the view that lexical-level representations (form) are independent whereas semantic-level representations (meaning) are shared (Kroll and De Groot 1997; Hernandez, Li, and Macwhinney 2005; Francis 2009). The multilingual inconsistencies observed in our experiments with ChatGPT suggest that the model does not possess such form-independent semantic-level representations. Further evidence for a form-dependent task understanding in LLMs comes from multilingual consistency evaluations with model-external translations. While these experiments do not guarantee that the different translations are meaning-equivalent according to the model, they still indicate that LLM responses seem to be largely driven by the lexical form of the input (Qi, Fernández, and Bisazza 2023).

To different degrees, both translations and paraphrases preserve the meaning of the original expression. In our work, we tested both translating and paraphrasing as sense-generation methods. However, translation equivalents and synonyms are treated differently in human cognition. For example, monolingual and bilingual children accept two names for the same object—violating the mutual exclusivity assumption—if the two names come from distinct languages but not if they come from the same language (Au and Glusman 1990). In particular, it seems that translation-equivalents have a closer cognitive status than within-language synonyms (Francis 2009). The model's consistency for translations versus paraphrases stands in contrast to the empirical evidence that changes in language have a more similar cognitive role than changes in wording. If anything, consistency tends to be higher for English paraphrases than translations (see for example Table 9). In conclusion, LLMs do not seem to separate between form and meaning in the way humans do.

It is important to keep in mind that looking up a fact with an LLM is not as straightforward as looking up a fact in an encyclopedia. Our experiments show that LLM responses to factual questions may vary between different representational forms of the same input, even if the model judges these forms to be meaning-equivalent. LLMs might (at least partially) lack an anchor for the linguistic forms they encounter, which humans naturally find in the physical world and social interactions (Bisk et al. 2020). Their responses, especially to factual questions, should thus be considered with caution

and users should be aware that other knowledge sources are more reliable. Chang and Bergen (2023) suggest that many weaknesses of LLMs, including form-dependencies, can be framed as under- and over-generalization errors. When a model is sensitive to small, meaning-preserving changes to the input, when recalling facts, this can be considered an under-generalization of the underlying factual knowledge. The model may compensate for this failure by over-generalizing other patterns, thus falling back on certain heuristics to generate an answer. In general, it is important to keep in mind that LLMs and humans are shaped by different pressures when making a comparison. For example, while LLM accuracy is strongly influenced by the probability of the task to be performed, the probability of the target output, and the probability of the provided input, humans are likely better at generalizing their task understanding across such variations (McCoy et al. 2023).

9.3 LLMs as Explanatory Models of Meaning: The Role of Multisense Consistency

What are the consequences of our findings for the role of LLMs as explanatory models of semantic understanding in humans? Up until now, the discussion has largely revolved around their capacity to represent symbolic structure and to capture the nature of language use, including communicative intent and grounding in the world. While there are a priori arguments that LLMs fail on both these fronts, let us consider some arguments in favor of such capacities. Concerning symbolic structure, arguments come, for example, from interpretability studies that identified dedicated neurons for encoding specific knowledge (Dai et al. 2022), concepts (Geva et al. 2021), or skills (Wang et al. 2022) in transformer-based LMs. Concerning perceptual grounding, it has been argued that important aspects of meaning are captured by the role a certain concept plays, that is, how it relates to other concepts within a representational framework, rather than being defined by an external referent (Piantadosi and Hill 2022). When studying the internal representations of LLMs, the organization of concepts—measured through similarity relationships—indeed seems to match the ground-truth organization of perceptual concepts such as colors (Abdou et al. 2021) or spatial relations (Patel and Pavlick 2022). The lack of self-consistency revealed by our findings opens up a new dimension to be considered when making such arguments. For example, it is not only relevant whether LLMs can encode symbolic structure and whether they encode concepts in line with a human-like conceptual structure, but also whether these encodings are consistent across senses. In other words, to establish a strong correspondence between LLM and human concept encodings, these encodings should bear resemblance across different senses.

With that, we believe that measuring multisense consistency could be a useful addition to the toolkit used to evaluate the extent to which models can understand natural language. The method can be used to assess generalization 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 could serve as a complement to performance-based model evaluation. Reporting consistency next to standard evaluation metrics like accuracy, BLEU, or F1-scores will make model evaluation more meaningful in providing an estimate of how well the model understands a given task beyond its specific form. Our paradigm can be cheaply and easily expanded to include more languages, tasks, models, and notions of “sense”. Our choice to generate 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, senses could be generated through various word- and sentence-level perturbations (e.g., Wang et al. 2021), across accents or dialects, or across different modalities. 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. Therefore, we encourage other researchers to treat multisense consistency as an integral part of benchmarking.

The consistency evaluation is only interesting if the model does not master the task on each sense, in which case its responses are trivially consistent. Although it is usually impressive when a model achieves high scores on a benchmark that was challenging for the previous model generation, the community rarely concludes that this model has mastered the skill this benchmark is supposedly testing. As a result, benchmarks are usually replaced by more challenging successors when this happens. Thus, we think it is likely that challenging benchmarks, which can be used for a non-trivial consistency evaluation, will continue to be available. Still, it is important to mention that consistency should be evaluated in experiments where the main source of potential inconsistencies is form-dependency. Model mistakes and inconsistencies should not be enforced on purpose, for example through ambiguous instructions. Further analyses, such as controlling the quality of the generated senses or calculating the proportions of consistent correct versus incorrect responses (see § 7), can help to rule out alternative explanations.

Crucially, multisense consistency experiments can primarily provide *negative evidence*. After all, even if an LLM is perfectly self-consistent, it could be mastering each form independently without relying on a shared meaning. With that, our method can be grouped together with other methods probing for human-level understanding that, when successfully passed, provoke thought about what “human-level understanding” means, rather than providing a proof for it (e.g., Biever 2023; Johnson-Laird and Ragni 2023).

A. Simple Facts Datasets

We use five different datasets to test for factual knowledge. To facilitate the dataset curation, we focused on facts that are usually presented in a table format and can be queried with the same template question regardless of the exact datapoint. At the same time, we tried to cover different domains of factual knowledge, including arithmetics, science, sports, economy, and literature. Note that these datasets are not intended to serve as full-fledged benchmarks of factual knowledge but rather as a proof-of-principle. In the following, these datasets are described in detail. We describe only the base data. The corresponding instructions are given in Table 1 in the main text. The csv files for each dataset can be found in our repository.

Arithmetics

The ARITHMETICS dataset tests for the sum of two numbers. The two numbers are sampled randomly between 1 and 1,000 and, to make the questions more different between languages, we chose to spell out the numbers in words. We wrote functions to map numerals to spelled-out numbers in all the languages we consider (see our repository). The function for English was used to generate the original dataset once the integers were sampled. The functions for the other languages were used to evaluate whether the

model correctly translated the English (spelled-out) numbers when generating other senses. The model, in turn, is asked to reply in numerical form, such that the answers can easily be validated. For instance, one datapoint could be $d = (\textit{five hundred seventy-three, twenty-seven})$ and the corresponding set of correct answers would be $A_d = \{600\}$. We sample 500 pairs of numbers, giving us a total of 500 datapoints.

Elements

The ELEMENTS dataset tests for the atomic number of chemical elements. Each datapoint consists of a chemical element (denoted by its element symbol), as well as its position on the periodic table (given by period and group). For example, Helium, which is in period 1 and group 18, is given by $d = (\textit{He}, 1, 18)$. The dataset is used for two different tasks. In the FROM-ELEMENT subtask, the atomic number of an element has to be determined from its chemical symbol. In the FROM-POSITION subtask, the atomic number of an element has to be determined from its position in the periodic table. Hence, in both cases, the set of correct answers for the above datapoint is $A_d = \{2\}$. The model is instructed to reply with the correct number allowing for easy evaluation against the ground truth. We ignore the f-block of the periodic table, resulting in a total of 90 datapoints (per subtask).

Olympics

The OLYMPICS dataset tests for the names of Olympic medallists. It is used for two subtasks. The 100M subtask asks for the medallists in the 100m competition (Summer Olympics). The DOWNHILL subtask asks for the medallists in the downhill competition (Winter Olympics). Information on the medallists for these competitions can be found on various sites on the Internet, for example, <https://olympics.com/en/news/olympics-100-metres-winners-list-men-women-gold-medals-champions> (100m) and https://en.wikipedia.org/wiki/List_of_Olympic_medalists_in_alpine_skiing (downhill). The templates have to be adapted, depending on whether the model is asked about the men's or the women's competition. Taken together, each datapoint consists of the competition (100m or downhill), the year of the games, the subgroup (men or women), and the type of medal (gold, silver, bronze). For example, one datapoint is $d = (100m, 1968, \textit{men}, \textit{gold})$. Athletes are often called by their nicknames. We ensure that the set of correct answers contains the nickname as well as the real name(s). For example, the set of correct answers for the datapoint above is $A_d = \{\textit{Jim Hines}, \textit{James Hines}, \textit{James Ray Hines}\}$. Each year in which Summer Olympics or Winter Olympics were held generates 6 datapoints (3 types of medals, men and women). We consider games until 2022 and remove ambiguous cases, resulting in a total of 148 datapoints for 100M and 117 datapoints for DOWNHILL.

Writers

The WRITERS dataset tests for the year of birth of well-known writers. Thus, each datapoint is a writer and the set of correct answers contains their year of birth, for example, $d = (\textit{Friedrich Schiller})$ and $A_d = \{1759\}$. We tried to generate a dataset structure such that writers are sampled equally from the languages we consider. That is, one fifth of the data are English-language writers, one fifth are German-language writers, and so forth. However, we did not ensure that all countries in which these languages

are spoken are taken into account. Lists of writers for the five languages were taken from Wikipedia:

- English (American authors only): <https://de.wikipedia.org/wiki/Liste.amerikanischer.Schriftsteller>
- German: https://en.wikipedia.org/wiki/List_of_German-language_authors
- Italian: https://en.wikipedia.org/wiki/List_of_Italian_writers
- Dutch: https://en.wikipedia.org/wiki/List_of_Dutch-language_writers
- Swedish: https://en.wikipedia.org/wiki/List_of_Swedish-language_writers

The list of Swedish-language writers had 186 entries and was the shortest. Therefore, we randomly sampled 186 writers from each of the lists (without replacement) and used those $186 \times 5 = 930$ datapoints to compose the dataset.

Companies

The COMPANIES dataset tests for the headquarters locations of different companies. Similar to WRITERS, we try to cover five different countries (US, Germany, Italy, Netherlands, Sweden), such that each of the languages we work with is the dominant language in one of them. Each datapoint consists of a company, for example, $d = (Volvo AB)$, and the set of correct answers contains all relevant variations in the city name, for example, $A_d = \{Gothenburg, Göteborg, Gotemburgo, Gotenburg\}$. We took the 100 largest companies for each of these countries from different lists on the Internet:

- US: https://en.wikipedia.org/wiki/List_of_largest_companies_in_the_United_States_by_revenue
- Germany: https://de.wikipedia.org/wiki/Liste_der_gr
- Italy: <https://www.value.today/headquarters/italy>
- Netherlands: <https://www.value.today/headquarters/netherlands>
- Sweden: <https://www.stockblogs.se/sveriges-storsta-foretag/>

If possible, we extracted both company and headquarters location from these lists. When no location was given, we searched for it online. In total, the dataset contains 100×5 datapoints.

B. Sense Generation Prompts

Simple Facts

For all SIMPLE FACTS datasets, except ARITHMETICS, only the task instructions (corresponding to the templates in Table 1) need to be translated, since the input data does not change between languages. The prompt for translating is “Please translate the following text into [LANGUAGE]:\n[TEXT]”. The prompt for paraphrasing is “Please paraphrase the following text:\n[TEXT]”. The ARITHMETICS input data consists of

spelled-out numbers, which have to be translated as well. In the case of paraphrasing, these spelled-out numbers are not paraphrased but remain in their original version. In the case of translation, the model is instructed to translate each number separately using the translation prompt above.

Benchmark Data

We use the model to generate alternative senses, treating the task instruction and the input data separately. The prompt for translating is “Please translate the following text into [LANGUAGE]:\n[TEXT]”. [LANGUAGE] is replaced by the target language and [TEXT] by the instruction (for translating instructions) or each datapoint from the benchmark (for translating input data). For Belebele, it was necessary to explicitly instruct the model to translate everything *without answering the question*. The prompt for paraphrasing differs depending on whether task instructions or input data are paraphrased. The prompt for paraphrasing the task instruction is “Please paraphrase the following text:\n[TEXT]”. The prompt for paraphrasing the input data from the benchmarks is task-specific to help preserve the structure of the original task prompt:

- PAWS: “Please paraphrase the following two sentences (separately). Reply only with the paraphrased text and do not add any additional comments: \n[TEXT].”
- XNLI: “Please paraphrase the following premise and hypothesis (separately). Reply only with the paraphrased text and do not add any additional comments: \n[TEXT].”
- COPA: Please paraphrase the following premise and two alternatives (separately). Reply only with the paraphrased text and do not add any additional comments: \n[TEXT].”
- Belebele: “Please paraphrase the following text passage, question, and multiple-choice answer options (separately). Make sure to paraphrase everything, including the passage, and reply only with the paraphrased text and do not add any additional comments:\n[TEXT].”

C. Task Instructions and Alternative Senses

Simple Facts

Table 5 shows the original English (*en*) task instructions for the SIMPLE FACTS datasets as well as the model’s paraphrases (*en^P*) and translations (*de^T*, *it^T*, *nl^T*, *sv^T*) thereof. Native speakers of the corresponding languages judged the paraphrases and translations to be mostly accurate, although they tend to stay very close to the English original. In some cases, this tendency leads to some formal mistakes. For example, the Dutch instruction for ARITHMETICS is “Wat is [NUMBER1] plus [NUMBER2]? Antwoord alstublieft alleen met het juiste nummer [...]”), where “Hoeveel is [NUMBER1] plus [NUMBER2]? Antwoord alstublieft alleen met het juiste getal [...]” would be more correct. In addition, there is a grammatical mistake in the Swedish translation for ELEMENTS, where the definitive article of “the atomic number” should be expressed by a suffix on the noun “atomnummer”, resulting in “atomnumret”.

Table 5
SIMPLE FACTS task instructions.

Task	Language	Instruction
ARITHMETICS	<i>en</i>	What is [NUMBER1] plus [NUMBER2]? Please reply with only the correct number (in numerical form) and no additional words.
	<i>en^P</i>	What is the sum of [NUMBER1] and [NUMBER2]? Please respond with only the correct numerical answer and no extra words.
	<i>de^T</i>	Was ist [NUMBER1] plus [NUMBER2]? Bitte antworten Sie nur mit der korrekten Zahl (in numerischer Form) und ohne zusätzliche Wörter.
	<i>it^T</i>	Quanto fa [NUMBER1] più [NUMBER2]? Si prega di rispondere solo con il numero corretto (in forma numerica) e senza parole aggiuntive.
	<i>nl^T</i>	Wat is [NUMBER1] plus [NUMBER2]? Antwoord alstublieft alleen met het juiste nummer (in numerieke vorm) en geen extra woorden.
	<i>sv^T</i>	Vad är [NUMBER1] plus [NUMBER2]? Vänligen svara endast med det korrekta numret (i numerisk form) och inga ytterligare ord.
ELEMENTS-FROM-ELEMENT	<i>en</i>	What is the atomic number of the chemical element [ELEMENT]? Please reply with the number only and do not use any additional words.
	<i>en^P</i>	Please provide the atomic number of the chemical element [ELEMENT] using only the number and no extra words.
	<i>de^T</i>	Was ist die Ordnungszahl des chemischen Elements [ELEMENT]? Bitte antworten Sie nur mit der Zahl und verwenden Sie keine zusätzlichen Wörter.
	<i>it^T</i>	Qual è il numero atomico dell'elemento chimico [ELEMENT]? Si prega di rispondere solo con il numero e di non utilizzare altre parole aggiuntive.
	<i>nl^T</i>	Wat is het atoomnummer van het chemisch element [ELEMENT]? Antwoord alstublieft alleen met het nummer en gebruik geen extra woorden.
	<i>sv^T</i>	Vad är det atomnummer för grundämnet [ELEMENT]? Vänligen svara endast med numret och använd inga ytterligare ord.
ELEMENTS-FROM-POSITION	<i>en</i>	What is the atomic number of the chemical element in period [PERIOD] and group [GROUP]? Please reply with the number only and do not use any additional words.
	<i>en^P</i>	Please provide the atomic number of the element in period [PERIOD] and group [GROUP]. Respond with only the number and no extra words.
	<i>de^T</i>	Was ist die Ordnungszahl des chemischen Elements in Periode [PERIOD] und Gruppe [GROUP]? Bitte antworten Sie nur mit der Zahl und verwenden Sie keine zusätzlichen Wörter.
	<i>it^T</i>	Qual è il numero atomico dell'elemento chimico nel periodo [PERIOD] e nel gruppo [GROUP]. Si prega di rispondere solo con il numero e di non utilizzare altre parole aggiuntive.
	<i>nl^T</i>	Wat is het atoomnummer van het chemisch element in periode [PERIOD] en groep [GROUP]. Antwoord alstublieft alleen met het nummer en gebruik geen extra woorden.
	<i>sv^T</i>	Vad är det atomnummer för det kemiska elementet i period [PERIOD] och grupp [GROUP]. Vänligen svara endast med numret och använd inga ytterligare ord.
OLYMPICS-100M	<i>en</i>	Who won the [MEDAL] medal in the [GENDER] 100 meters at the [YEAR] Summer Olympics? Please reply with the name only and do not use any additional words.
	<i>en^P</i>	Please provide the name of the athlete who won the [MEDAL] medal in the [GENDER] 100 meters at the [YEAR] Summer Olympics, using only the name and no extra words.
	<i>de^T</i>	Wer hat die [MEDAL]-Medaille im [GENDER]-100-Meter-Lauf bei den Olympischen Sommerspielen [YEAR] gewonnen? Bitte antworten Sie nur mit dem Namen und verwenden Sie keine zusätzlichen Wörter.

Table 5
Continued.

Task	Language	Instruction
	<i>it</i> ^T	Chi ha vinto la medaglia [MEDAL] nei 100 metri [GENDER] alle Olimpiadi estive del [YEAR]? Si prega di rispondere solo con il nome e di non utilizzare altre parole aggiuntive.
	<i>nl</i> ^T	Wie heeft de [MEDAL] medaille gewonnen op de 100 meter voor [GENDER] tijdens de Zomerspelen van [YEAR]? Antwoord alstublieft alleen met de naam en gebruik geen extra woorden.
	<i>sv</i> ^T	Vem vann [MEDAL] medaljen i [GENDER] 100 meter vid sommar-OS [YEAR]? Vänligen svara med namnet endast och använd inga ytterligare ord.
OLYMPICS-DOWNHILL	<i>en</i>	Who won the [MEDAL] medal in the [GENDER] downhill competition at the [YEAR] Winter Olympics? Please reply with the name only and do not use any additional words.
	<i>en</i> ^P	Please provide the name of the athlete who won the [MEDAL] medal in the [GENDER] downhill competition at the [YEAR] Winter Olympics, without using any extra words.
	<i>de</i> ^T	Wer hat die [MEDAL]-Medaille im [GENDER]-Abfahrtsrennen bei den Olympischen Winterspielen [YEAR] gewonnen? Bitte antworten Sie nur mit dem Namen und verwenden Sie keine zusätzlichen Wörter.
	<i>it</i> ^T	Chi ha vinto la medaglia [MEDAL] nella gara di discesa libera [GENDER] alle Olimpiadi invernali del [YEAR]? Per favore, rispondi solo con il nome e non utilizzare altre parole aggiuntive.
	<i>nl</i> ^T	Wie heeft de [MEDAL] medaille gewonnen in de [GENDER] afdaling wedstrijd op de [YEAR] Olympische Winterspelen? Antwoord alstublieft alleen met de naam en gebruik geen extra woorden.
	<i>sv</i> ^T	Vem vann [MEDAL] medaljen i [GENDER] störtloppstävling vid vinter-OS [YEAR]? Vänligen svara med namnet endast och använd inga ytterligare ord.
WRITERS	<i>en</i>	In what year was the writer [AUTHOR] born? Please reply with the correct year only and do not use any additional words.
	<i>en</i> ^P	What is the birth year of the author [AUTHOR]? Please respond with only the correct year and avoid using extra words.
	<i>de</i> ^T	In welchem Jahr wurde der Schriftsteller / die Schriftstellerin [AUTHOR] geboren? Bitte antworten Sie nur mit dem korrekten Jahr und verwenden Sie keine zusätzlichen Wörter.
	<i>it</i> ^T	In che anno è nato lo scrittore / è nata la scrittrice [AUTHOR]? Per favore, rispondi solo con l'anno corretto e non utilizzare altre parole aggiuntive.
	<i>nl</i> ^T	In welk jaar is de schrijver [AUTHOR] geboren? Uw antwoord mo et alleen bestaan uit het juiste jaartal.
	<i>sv</i> ^T	I vilket år föddes författaren [AUTHOR]? Ditt svar ska bara bestå av det korrekta året.
COMPANIES	<i>en</i>	In what city does [COMPANY] have its headquarters? Please reply only with the name of the city and no additional words.
	<i>en</i> ^P	Where is the headquarters of [COMPANY] located? Please respond with only the city name, without any extra words.
	<i>de</i> ^T	In welcher Stadt hat [COMPANY] seinen Hauptsitz? Bitte antworten Sie nur mit dem Namen der Stadt und ohne zusätzliche Wörter.
	<i>it</i> ^T	In quale città ha sede [COMPANY]? Si prega di rispondere solo con il nome della città e senza parole aggiuntive.
	<i>nl</i> ^T	In welke stad heeft [COMPANY] zijn hoofdkantoor? Antwoord alstublieft alleen met de naam van de stad en geen extra woorden.
	<i>sv</i> ^T	I vilken stad har [COMPANY] sitt huvudkontor? Vänligen svara endast med stadens namn och inga ytterligare ord.

Benchmark Data

Table 6 lists the original English (*en*) task instructions for the benchmark datasets as well as the model’s paraphrases (*en^P*) and translations (*de^T*, *it^T*, *nl^T*, *sv^T*) thereof. Native speakers of the corresponding languages judged the paraphrases and translations to be generally accurate but some sentences contained minor mistakes or aspects that the native speakers would have translated differently. Points that were mentioned are that (1) the model translates “premise” to “presupposto” in Italian (COPA and XNLI) even though “premissa” is more appropriate and (2) the repeated use of “noch” in the Dutch XNLI instruction is incorrect and the correct sentence should end with something like “als de premisse de hypothese noch impliceert nog tegenspreekt”.

Table 6
Benchmark data task instructions.

Task	Language	Instruction
PAWS	<i>en</i>	Do the following two sentences have the same meaning? Sentence 1: “[SENTENCE1]” Sentence 2: “[SENTENCE2]” Please reply with a single word, either “yes” or “no”.
	<i>en^P</i>	Are the meanings of the following two sentences the same? Sentence 1: “[SENTENCE1]” Sentence 2: “[SENTENCE2]” Please respond with either “yes” or “no”.
	<i>de^T</i>	Haben die folgenden beiden Sätze die gleiche Bedeutung? Satz 1: “[SENTENCE1]” Satz 2: “[SENTENCE2]” Bitte antworten Sie mit einem einzigen Wort, entweder “ja” oder “nein”.
	<i>it^T</i>	Le seguenti due frasi hanno lo stesso significato? Fraser 1: “[SENTENCE1]” Fraser 2: “[SENTENCE2]” Rispondi con una sola parola, “sì” o “no”.
	<i>nl^T</i>	Hebben de volgende twee zinnen dezelfde betekenis? Zin 1: “[SENTENCE1]” Zin 2: “[SENTENCE2]” Antwoord alstublieft met één woord, ofwel “ja” ofwel “nee”.
	<i>sv^T</i>	Har de följande två meningarna samma betydelse? Mening 1: “[SENTENCE1]” Mening 2: “[SENTENCE2]” Svara med ett enda ord, antingen “ja” eller “nej”.
XNLI	<i>en</i>	Given the following premise and hypothesis, please identify whether the premise entails the hypothesis, contradicts the hypothesis, or neither of the two. Premise: “[PREMISE]” Hypothesis: “[HYPOTHESIS]” Please reply with a single word: “entailment” if the premise entails the hypothesis, “contradiction” if the premise contradicts the hypothesis, and “neutral” if the premise neither entails nor contradicts the hypothesis.
	<i>en^P</i>	Please determine if the premise and hypothesis are related. Premise: “[PREMISE]” Hypothesis: “[HYPOTHESIS]” If the premise supports the hypothesis, indicate “entailment”. If the premise contradicts the hypothesis, indicate “contradiction”. If there is no clear relationship between the two, indicate “neutral”.
	<i>de^T</i>	Angesichts der folgenden Prämisse und Hypothese, bitte identifizieren Sie, ob die Prämisse die Hypothese impliziert, der Hypothese widerspricht oder weder das eine noch das andere. Prämisse: “[PREMISE]” Hypothese: “[HYPOTHESIS]” Bitte antworten Sie mit einem einzigen Wort: “Implikation”, wenn die Prämisse die Hypothese impliziert, “Widerspruch”, wenn die Prämisse der Hypothese widerspricht, und “neutral”, wenn die Prämisse weder die Hypothese impliziert noch ihr widerspricht.

Table 6
Continued.

Task	Language	Instruction
	<i>it</i> ^T	Dato il seguente presupposto e ipotesi, per favore identifica se il presupposto implica l'ipotesi, contraddice l'ipotesi o né implica né contraddice l'ipotesi. Presupposto: "[PREMISE]" Ipotesi: "[HYPOTHESIS]" Per favore rispondi con una sola parola: "implicazione" se il presupposto implica l'ipotesi, "contraddizione" se il presupposto contraddice l'ipotesi e "neutrale" se il presupposto né implica né contraddice l'ipotesi.
	<i>nl</i> ^T	Gegeven de volgende premisse en hypothese, identificeer alstublieft of de premisse de hypothese impliceert, de hypothese tegenspreekt, of geen van beide. Premisse: "[PREMISE]" Hypothese: "[HYPOTHESIS]" Antwoord alstublieft met één woord: "implicatie" als de premisse de hypothese impliceert, "tegenspraak" als de premisse de hypothese tegenspreekt, en "neutraal" als de premisse noch de hypothese impliceert noch tegenspreekt.
	<i>sv</i> ^T	Givet följande premiss och hypotes, vänligen ange om premissen innebär hypotesen, motsäger hypotesen eller varken innebär eller motsäger hypotesen. Premiss: "[PREMISE]" Hypotes: "[HYPOTHESIS]" Vänligen svara med ett enda ord: "innebär" om premissen innebär hypotesen, "motsäger" om premissen motsäger hypotesen och "neutral" om premissen varken innebär eller motsäger hypotesen.
COPA	<i>en</i>	Given the following premise, which of the two alternatives is more plausible? Premise: "[PREMISE]" Alternative 1: "[CHOICE1]" Alternative 2: "[CHOICE2]" Please answer with a single word: "Alternative-1" if alternative 1 is more plausible and "Alternative-2" if alternative 2 is more plausible.
	<i>en</i> ^P	Based on the provided premise, which of the two options is more likely? Premise: "[PREMISE]" Option 1: "[CHOICE1]" Option 2: "[CHOICE2]" Please respond with either "Option-1" if option 1 is more likely, or "Option-2" if option 2 is more likely.
	<i>de</i> ^T	Angesichts der folgenden Prämisse, welche der beiden Alternativen ist plausibler? Prämisse: "[PREMISE]" Alternative 1: "[CHOICE1]" Alternative 2: "[CHOICE2]" Bitte antworten Sie mit einem einzigen Wort: "Alternative-1", wenn Alternative 1 plausibler ist, und "Alternative-2", wenn Alternative 2 plausibler ist.
	<i>it</i> ^T	Dato il seguente presupposto, quale delle due alternative è più plausibile? Presupposto: "[PREMISE]" Alternativa 1: "[CHOICE1]" Alternativa 2: "[CHOICE2]" Per favore, rispondi con una sola parola: "Alternativa-1" se l'alternativa 1 è più plausibile e "Alternativa-2" se l'alternativa 2 è più plausibile.
	<i>nl</i> ^T	Gegeven de volgende premisse, welke van de twee alternatieven is waarschijnlijker? Premisse: "[PREMISE]" Alternatief 1: "[CHOICE1]" Alternatief 2: "[CHOICE2]" Antwoord alstublieft met één woord: "Alternatief-1" als alternatief 1 waarschijnlijker is en "Alternatief-2" als alternatief 2 waarschijnlijker is.
	<i>sv</i> ^T	Givet följande premiss, vilket av de två alternativen är mer troligt? Premiss: "[PREMISE]" Alternativ 1: "[CHOICE1]" Alternativ 2: "[CHOICE2]" Svara med ett enda ord: "Alternativ-1" om alternativ 1 är mer troligt och "Alternativ-2" om alternativ 2 är mer troligt.

Table 6
Continued.

Task	Language	Instruction
BELEBELE	<i>en</i>	[PASSAGE]
		[QUESTION] Option A: [ANSWER1] Option B: [ANSWER2] Option C: [ANSWER3] Option D: [ANSWER4] Please reply with "A", "B", "C", or "D" to indicate the correct answer. Your reply should be a single letter and should not contain any additional words.
	<i>en^P</i>	[PASSAGE]
		[QUESTION] A) [ANSWER1] B) [ANSWER2] C) [ANSWER3] D) [ANSWER4] Please respond with the letter corresponding to the correct answer choice. Your response should be a single letter and should not include any extra words.
	<i>de^T</i>	[PASSAGE]
		[QUESTION] Option A: [ANSWER1] Option B: [ANSWER2] Option C: [ANSWER3] Option D: [ANSWER4] Antworten Sie bitte mit "A", "B", "C", oder "D", um die richtige Antwort anzugeben. Ihre Antwort sollte nur ein einzelner Buchstabe sein und keine zusätzlichen Wörter enthalten.
	<i>it^T</i>	[PASSAGE]
		[QUESTION] Opzione A: [ANSWER1] Opzione B: [ANSWER2] Opzione C: [ANSWER3] Opzione D: [ANSWER4] Rispondi con "A", "B", "C" o "D" per indicare la risposta corretta. La tua risposta deve essere una singola lettera e non deve contenere parole aggiuntive.
	<i>nl^T</i>	[PASSAGE]
		[QUESTION] Optie A: [ANSWER1] Optie B: [ANSWER2] Optie C: [ANSWER3] Optie D: [ANSWER4] Antwoord alstublieft met "A", "B", "C" of "D" om het juiste antwoord aan te geven. Uw antwoord moet uit één letter bestaan en mag geen extra woorden bevatten.
	<i>sv^T</i>	[PASSAGE]
		[QUESTION] Alternativ A: [ANSWER1] Alternativ B: [ANSWER2] Alternativ C: [ANSWER3] Alternativ D: [ANSWER4] Vänligen svara med "A", "B", "C", eller "D" för att ange det korrekta svaret. Ditt svar ska vara en enda bokstav och får inte innehålla några ytterligare ord.

D. Accuracy Scores

Simple Facts

Table 7
Accuracy (%) on the simple fact datasets, with 95% confidence intervals.

	arithmetics	elements		olympics		writers	companies	total
	-	elem	pos	100m	downhill	-	-	avg
en	99.4±1.0	100.0±nan	37.8±10.0	55.4±8.1	37.6±9.4	76.2±2.9	78.2±3.8	73.5
en ^P	98.6±1.4	100.0±nan	42.2±10.0	54.1±8.1	31.6±9.4	76.2±2.8	76.0±3.8	72.2
de ^T	45.2±4.6	98.9±5.6	40.0±10.0	50.7±8.1	35.0±9.4	76.8±2.8	75.4±3.8	68.7
it ^T	44.0±4.4	100.0±nan	36.7±10.0	51.4±8.1	35.0±9.4	75.3±2.9	73.6±4.0	67.4
nl ^T	42.4±4.4	100.0±nan	36.7±10.0	52.0±8.1	35.0±8.5	76.8±2.9	73.2±4.2	67.7
sv ^T	19.2±3.6	100.0±nan	41.1±11.1	50.7±8.1	33.3±8.5	74.3±2.9	71.8±4.0	65.0

Benchmark Data

Table 8
Accuracy (%) on the benchmark datasets, with 95% confidence intervals.

	paws	xnli	copa	belebele	avg
en	75.6±1.9	43.7±1.4	84.4±3.4	85.9±2.3	72.4
en ^P	67.6±2.1	53.5±1.4	82.2±3.4	-	-
de ^T	64.3±2.1	50.0±1.4	85.6±3.2	81.2±2.7	70.3
it ^T	75.1±2.0	56.4±1.4	86.6±3.2	81.0±2.7	74.8
nl ^T	71.9±2.0	50.9±1.4	83.4±3.4	79.0±2.7	71.3
sv ^T	55.9±2.2	47.0±1.4	89.2±2.8	79.1±2.8	67.8

E. Accuracy Based on Containment versus Exact Match

On the SIMPLE FACTS datasets, the model is instructed to reply with the correct entity (and no additional words), which we then use to quantify consistency. Hence, it is important that the model actually follows that instruction across all senses. Otherwise, it could be that the model replies with “Friedrich Schiller was born in 1759” when prompted for a writer in English but “1759” when prompted in German. While a failure to follow the instruction in one language but not the other could be considered an unwanted inconsistency, the meaning of both answers is arguably the same, and we would like to differentiate between both cases.

If the model replies correctly but not in one word, the response *contains* the right answer but does not *exactly match* it. Figure 9 shows the distribution of the difference in accuracy based on containment versus exact match. The scores for COMPANIES and WRITERS are calculated separately for each language-specific subgroup of samples (i.e., US companies, German companies, ...) to obtain more detailed information. In most cases, the “containment” score is not at all or only slightly higher than the exact match score. The only exception occurs for Dutch companies when prompted with

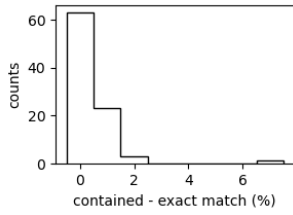


Figure 9
Containment score minus exact match score across tasks and senses.

en^P, with a 7% difference in accuracy. This mismatch arises because the model—while otherwise replying with only the city name—always responds with a full sentence when the correct answer is “The Hague” (e.g., “The headquarters of Shell PLC is located in The Hague.”). Thus, except for this curious case, inconsistencies can largely not be attributed to a failure to express a response in the correct form.

F. Consistency Scores

Simple Facts

Table 9
Consistency (%) on the simple fact datasets.

	arithmetics		elements		olympics		writers	companies	total
	–	elem	position	100m	downhill	–	–	avg	
id	100.0	100.0	91.1	90.5	88.9	87.1	97.4	92.8	
<i>en^P</i>	99.0	100.0	77.8	82.4	61.5	82.7	91.2	86.0	
<i>de^T</i>	45.0	98.9	71.1	80.4	76.9	81.2	89.4	81.7	
<i>it^T</i>	44.2	100.0	70.0	75.7	70.1	81.2	86.4	79.8	
<i>nl^T</i>	42.2	100.0	67.8	83.8	64.1	79.1	88.8	79.8	
<i>sv^T</i>	19.4	100.0	70.0	83.8	59.8	77.3	84.8	76.2	

Benchmark Data

Table 10
Consistency (%) on the benchmark datasets.

	paws	xnli	copa	belebele	avg
id	95.2	96.0	96.8	97.1	96.3
<i>en^P</i>	76.5	56.3	85.2	–	–
<i>de^T</i>	74.7	51.2	88.8	84.7	74.8
<i>it^T</i>	82.2	57.5	85.8	85.1	77.6
<i>nl^T</i>	82.4	73.1	91.0	85.8	83.1
<i>sv^T</i>	67.9	82.8	86.0	83.3	80.0

G. Examples of Inconsistent Responses

Table 11
Examples of inconsistencies for SIMPLE FACTS. We report the first ten inconsistent samples per dataset and sense.

Task	Senses	Examples
ARITHMETICS	$(en en)$	-
	$(en en^P)$	(540 340), (1770 1778), (237 Two hundred thirty-seven.), (1173 One thousand one hundred seventy-three), (1013 One thousand thirteen.)
	$(en de^T)$	(778 678), (618 1008), (926 526), (115 915), (924 524), (1535 1035), (1693 1689), (1437 1337), (1151 1248), (1248 1448)
	$(en it^T)$	(778 678), (858 788), (1471 1437), (926 836), (924 923), (577 577 + 300 = 877), (1535 1335), (1693 1683), (1437 1497), (1151 1051)
	$(en ni^T)$	(778 678), (965 865), (858 958), (926 726), (115 109), (1535 935), (1693 1689), (1437 1338), (1151 846), (1248 848)
$(en sv^T)$	(778 784 + 94 = 878), (965 929), (858 792), (1471 1465), (926 733 + 163 = 896), (1277 1170), (1304 645), (924 923), (577 577 + 300 = 877), (1535 825)	
ELEMENTS-FROM-ELEMENT	$(en en)$	-
	$(en en^P)$	-
	$(en de^T)$	(114 9)
	$(en it^T)$	-
	$(en ni^T)$	-
$(en sv^T)$	-	
ELEMENTS-FROM-POSITION	$(en en)$	(22 20), (13 31), (17 107), (107 104), (106 46), (86 14), (33 51), (17 53)
	$(en en^P)$	(16 8), (19 37), (36 26), (28 39), (13 31), (23 55), (22 38), (45 46), (33 51), (16 34)
	$(en de^T)$	(12 4), (16 8), (19 11), (35 17), (36 26), (28 39), (48 30), (33 15), (38 12), (22 23)
	$(en it^T)$	(2 1), (12 4), (13 5), (7 15), (16 8), (35 23), (28 35), (28 40), (48 40), (13 31)
	$(en ni^T)$	(12 4), (13 5), (16 8), (21 13), (35 23), (28 39), (28 40), (48 40), (33 15), (38 12)
$(en sv^T)$	(2 1), (13 5), (16 8), (17 9), (21 23), (36 26), (28 39), (28 40), (48 40), (13 31)	
OLYMPICS-100M	$(en en)$	(Charley Paddock Harold Abrahams), (Arthur Jonath Eddie Tolan), (Lloyd LaBeach Herbert McKenley), (Lloyd LaBeach Herbert McKenley), (Herb McKenley Hector Hogan), (Ben Johnson Calvin Smith), (Kim Collins Justin Gatlin), (Ethel Smith Elizabeth Robinson), (Shirley Strickland Marjorie Jackson), (Shirley Strickland Marlene Mathews)
	(en, en^P)	(Francis Lane Frank Lane), (Fay Moulton Frank Jarvis), (Nathaniel Cartmell Reggie Walker), (Nate Cartmell Reggie Walker), (Arthur Porritt Percy Williams), (Arthur Jonath Percy Williams), (Barney Ewell Herb McKenley), (Lloyd LaBeach Herb McKenley), (Lloyd LaBeach Herb McKenley), (Herb McKenley Hector Hogan)
	$(en de^T)$	(Francis Lane Frank Lane), (Nate Cartmell Reggie Walker), (Arthur Porritt Arthur Jonath), (Arthur Porritt Arthur Jonath), (Arthur Jonath Eddie Tolan), (Arthur Jonath Eddie Tolan), (Arthur Jonath Ralph Metcalfe), (Lloyd LaBeach Barney Ewell), (Lloyd LaBeach Herb McKenley), (Enrique Figuerola Bob Hayes)
	$(en it^T)$	(Fay Moulton Frank Castle), (Nathaniel Cartmell Reginald Walker), (Nate Cartmell Reginald Walker), (Charley Paddock Harold Abrahams.), (Arthur Porritt Percy Williams.), (Arthur Porritt Arthur Jonath), (Arthur Jonath Eddie Tolan.), (Arthur Jonath Eddie Tolan.), (Arthur Jonath Ralph Metcalfe.), (Lloyd LaBeach Barney Ewell.)
	$(en ni^T)$	(Fay Moulton Frank Waller), (Arthur Porritt Arthur Jonath), (Arthur Porritt Arthur Jonath), (Lloyd LaBeach Barney Ewell), (Lloyd LaBeach Herb McKenley), (Herb McKenley Thane Baker.), (Enrique Figuerola Edwin Roberts.), (Valeriy Borzov Valeri Borzov), (Ben Johnson Calvin Smith.), (Ato Boldon Maurice Greene)
$(en sv^T)$	(Francis Lane Frank Lane), (Ralph Craig Donald Lippincott), (Arthur Porritt Arthur Jonath), (Lloyd LaBeach Barney Ewell.), (Lloyd LaBeach Herb McKenley), (Valeriy Borzov Valeri Borzov), (Ben Johnson Calvin Smith.), (Linford Christie Carl Lewis), (Ato Boldon Maurice Greene.), (Kim Collins Asafa Powell)	

Table 11
Continued.

Task	Senses	Examples
OLYMPICS-DOWNHILL	(en en)	(Egon Zimmermann Guy Périllat), (Franz Klammer Bernhard Russi), (Franck Piccard Franz Heinzer), (Didier Défago Didier Defago), (Beat Feuz Kjetil Jansrud), (Hedy Schlunegger Trude Beiser-Jochum), (Andrea Mead-Lawrence Trude Beiser-Jochum), (Trude Beiser-Jochum Hanni Wenzel), (Christl Haas Christine Goitschel), (Brigitte Oertli Vreni Schneider)
	(en en ^P)	(Zeno Colò Andreas Molterer), (Christian Pravda Anton Sailer), (Christian Pravda Andreas), (Guy Périllat Jean Vuarnet.), (Egon Zimmermann Jean-Claude Killy), (Bernhard Franz Klammer), (Leonhard Stock Bill Johnson), (Anton Steiner Bill Johnson), (Franck Piccard Kjetil André Aamodt), (Hans Knauss Hermann Maier.)
	(en de ^T)	(Egon Zimmermann Guy Périllat), (Bernhard Bernhard Russi), (Anton Steiner Peter Müller), (Franz Franz Heinzer), (Hermann Maier Lasse Kjus.), (Hans Knauss Hermann Maier), (Antoine Dénériaz Michael Walchhofer), (Kjetil André Aamodt Michael Walchhofer), (Bode Miller Aksel Lund Svindal), (Kjetil Jansrud Aksel Lund Svindal)
	(en it ^T)	(Egon Zimmermann Jean-Claude Killy.), (Egon Zimmermann Guy Périllat.), (Bernhard Bernhard Russi.), (Leonhard Stock Peter Wirnsberger.), (Franz Franz Heinzer.), (Tommy Moe Markus Wasmeier.), (Hans Knauss Lasse Kjus.), (Fritz Strobl Lasse Kjus.), (Antoine Dénériaz Michael Walchhofer.), (Kjetil André Aamodt Michael Walchhofer.)
	(en nl ^T)	(Zeno Colò Andrea Mead-Lawrence), (Egon Zimmermann Guy Périllat), (Egon Zimmermann Guy Périllat), (Bernhard Bernhard Russi), (Leonhard Stock Peter Müller), (Leonhard Stock Peter Müller), (Franz Franz Heinzer), (Franz Heinzer Franck Piccard), (Franck Piccard Franz Heinzer), (Tommy Moe Patrick Ortlieb)
	(en sv ^T)	(Egon Zimmermann Guy Périllat), (Egon Zimmermann Guy Périllat.), (Bernhard Bernhard Russi), (Anton Steiner Bill Johnson), (Franz Franz Heinzer.), (Franck Piccard Franz Heinzer), (Tommy Moe Markus), (Hermann Maier Lasse Kjus.), (Hans Knauss Lasse Kjus.), (Antoine Dénériaz Fritz Strobl.)
WRITERS	(en en)	(1978 1977), (1903 1912), (1911 1901), (1965 1968), (1982 1984), (1975 1980), (1962 1939), (1880 1891), (1930 1936), (1851 1871)
	(en en ^P)	(1952 1944), (1978 1977), (1940 1925), (1992 I'm sorry, but I don't have access to personal information about individuals unless it has been shared with me in the course of our conversation.), (1903 1912), (1945 1939), (1956 1961), (1935 1923), (1955 1949)
	(en de ^T)	(1952 1949), (1978 1977), (1932 1941), (1955 1953), (1903 1922), (1943 1956), (1940 1939), (1956 1961), (1935 1923), (1911 1901), (1965 1962)
	(en it ^T)	(1992 1985), (1929 1932), (1903 1921), (1945 1935), (1940 1943), (1956 1961), (1935 1923), (1965 1962), (1982 1986), (1975 1969)
	(en nl ^T)	(1978 1977), (1940 1910), (1992 1987), (1903 1922), (1945 1950), (1941 1935), (1940 Het juiste jaartal van de geboorte van schrijver Jeannette Howard Foster is niet beschikbaar.), (1980 Ik heb geen informatie over een schrijver genaamd Eric San Juan.), (1956 1961), (1935 1923)
	(en sv ^T)	(1943 1938), (1941 1932), (1952 1949), (1978 1976), (1932 Bob McGrath föddes år 1932.), (1992 Jag beklagar, men jag har ingen information om författaren Zach Hughes och när han föddes.), (1903 1921), (1890 1871), (1963 Jag beklagar, men jag kan inte hitta information om författaren Susan Wrights födelseår.), (1943 1956)
COMPANIES	(en en)	(Dear user, the headquarters of Ford Motor Company is located in Dearborn. Dear user, Ford Motor Company has its headquarters in Dearborn.), (Mayfield Village Cleveland), (Berlin Munich), (Frankfurt Mannheim), (Milan Genoa), (Rome Milan), (Rome Milan), (Hilversum Amsterdam), (Hoofddorp Hague), (Apeldoorn Heerenveen)
	(en en ^P)	(Issaquah Seattle), (North Chicago Chicago), (Fort Worth Dallas), (Cologne Frankfurt), (Stuttgart Munich), (Salzgitter Salzgitter AG is headquartered in Salzgitter.), (Jena Oberkochen), (Nuremberg Frankfurt), (San Donato Milanese Milan), (Bergamo Stezzano)
	(en de ^T)	(Issaquah Seattle), (Dear user, the headquarters of Ford Motor Company is located in Dearborn. Dear user, Ford Motor Company has its headquarters in Dearborn.), (Purchase New York City), (Kenilworth New York), (Mayfield Village Cleveland), (Fort Worth Dallas), (Cologne Frankfurt), (Selm Lünen), (Munich Ehningen), (Stuttgart Ismaning)

Table 11
Continued.

Task	Senses	Examples
	(<i>en</i> <i>it</i> ^T)	(Irving Houston), (Chesterbrook Philadelphia), (Issaquah Seattle), (Dublin Dublin), (Bloomfield Philadelphia), (Dear user, the headquarters of Ford Motor Company is located in Dearborn. Detroit), (Purchase New York City.), (White Plains New York.), (Chicago Seattle), (North Chicago Chicago.)
	(<i>en</i> <i>nl</i> ^T)	(Minnetonka Minneapolis), (Issaquah Seattle), (Dear user, the headquarters of Ford Motor Company is located in Dearborn. Dear user, Ford Motor Company heeft zijn hoofdkantoor in Dearborn.), (Deerfield De stad waar Walgreens Boots Alliance Inc. zijn hoofdkantoor heeft, is Deerfield.), (Purchase New York City), (White Plains New York City), (Mayfield Village Cleveland), (Fort Worth Dallas), (Cologne Frankfurt), (Selm Lünen)
	(<i>en</i> <i>sv</i> ^T)	(Minnetonka Minneapolis.), (Dear user, the headquarters of Ford Motor Company is located in Dearborn. Dear user, Ford Motor Company's headquarters is located in Dearborn.), (Chicago San Francisco.), (Purchase New York.), (White Plains Hamilton), (Austin San Francisco.), (Fremont Toronto), (North Chicago Chicago), (Stamford St. Louis), (Mayfield Village Cleveland.)

H. Accuracy Scores for Ablation Experiments

Table 12

Accuracy scores for the ablation experiments. We consider ablations in which we translate/paraphrase only the instruction (columns *I*) or only the input data (columns *X*).

	paws		xnli		copa		belebele	
	<i>I</i>	<i>X</i>	<i>I</i>	<i>X</i>	<i>I</i>	<i>X</i>	<i>I</i>	<i>X</i>
en	75.6		43.7		84.4		85.9	
en ^P	71.8	69.7	54.2	44.5	88.2	80.6	87.4	–
de ^T	62.5	78.7	49.2	38.0	88.0	82.6	86.2	80.0
it ^T	72.8	76.8	57.2	38.4	91.0	81.8	86.0	78.0
nl ^T	68.7	78.9	48.1	36.8	86.4	84.8	85.0	79.2
sv ^T	59.1	76.1	48.3	37.1	93.0	81.4	85.6	79.8

I. Genbench Evaluation Card

Our work uses generalization across senses to assess task understanding in LLMs. In Figure 10, we provide the GenBench eval card (Hupkes et al. 2023) of our experiments.

Motivation					
<i>Practical</i>	<i>Cognitive</i>	<i>Intrinsic</i>	<i>Fairness</i>		
		☒			
Generalisation type					
<i>Compositional</i>	<i>Structural</i>	<i>Cross Task</i>	<i>Cross Language</i>	<i>Cross Domain</i>	<i>Robustness</i>
			☒		☒
Shift type					
<i>Covariate</i>	<i>Label</i>	<i>Full</i>	<i>Assumed</i>		
☒					
Shift source					
<i>Naturally occurring</i>	<i>Partitioned natural</i>	<i>Generated shift</i>	<i>Fully generated</i>		
☒					
Shift locus					
<i>Train–test</i>	<i>Finetune train–test</i>	<i>Pretrain–train</i>	<i>Pretrain–test</i>		
					☒

Figure 10 Our experiments assess cross-lingual generalization for natural corpora, in pretrained LLMs, to assess LLM task understanding.

Acknowledgments

We would like to thank the anonymous reviewers and Mortimer von Chappuis for their detailed and helpful feedback on our first submission. We would also like to thank Marco Baroni and Ryan Nefdt for their valuable feedback on this project at an earlier stage. Finally, we would like to thank Henrik Löfberg for his help with the evaluation of the Swedish translations.

References

Abdou, Mostafa, Kulmizev Artur Hershovich Daniel Frank Stella, Ellie Pavlick and Anders Søgaard. 2021. Can language models encode perceptual structure without grounding? A case study in color. In *Proceedings of the 25th Conference on Computational Natural Language*

Learning, pages 109–132. <https://doi.org/10.18653/v1/2021.conll-1.9>

Abnar, Samira, Lisa Beinborn, Rochelle Choenni, and Willem Zuidema. 2019. Blackbox meets blackbox: Representational similarity & stability analysis of neural language models and brains. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 191–203. <https://doi.org/10.18653/v1/W19-4820>

Alberti, Chris, 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. <https://doi.org/10.18653/v1/P19-1620>

Arehalli, Suhas, Brian Dillon, and Tal Linzen. 2022. Syntactic surprisal from neural

- models predicts, but underestimates, human processing difficulty from syntactic ambiguities. In *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*, pages 301–313. <https://doi.org/10.18653/v1/2022.conll-1.20>
- Asai, Akari 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. <https://doi.org/10.18653/v1/2020.acl-main.499>
- Au, Terry K. and Mariana Glusman. 1990. The principle of mutual exclusivity in word learning: To honor or not to honor? *Child Development*, 61(5):1474–1490. <https://doi.org/10.1111/j.1467-8624.1990.tb02876.x>, PubMed: 2245739
- Badre, David and Derek Evan Nee. 2018. Frontal cortex and the hierarchical control of behavior. *Trends in Cognitive Sciences*, 22(2):170–188. <https://doi.org/10.1016/j.tics.2017.11.005>, PubMed: 29229206
- Bandarkar, Lucas, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2023. The belebele benchmark: A parallel reading comprehension dataset in 122 language variants. *ArXiv preprint*, arXiv:2308.16884. <https://doi.org/10.18653/v1/2024.acl-long.44>
- Baroni, Marco. 2023. On the proper role of linguistically oriented deep net analysis in linguistic theorising. In *Algebraic Structures in Natural Language*, pages 1–16, CRC Press. <https://doi.org/10.1201/9781003205388-1>
- Barsalou, Lawrence W. 2005. Abstraction as dynamic interpretation in perceptual symbol systems. In *Building Object Categories in Developmental Time*, pages 389–431, Lawrence Erlbaum Associates.
- Bencheikroun, Youssef, Megi Dervishi, Mark Ibrahim, Jean-Baptiste Gaya, Xavier Martinet, Grégoire Mialon, Thomas Scialom, Emmanuel Dupoux, Dieuwke Hupkes, and Pascal Vincent. 2023. Worldsense: A synthetic benchmark for grounded reasoning in large language models. *ArXiv preprint*, arXiv:2311.15930.
- Bender, Emily M. and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198. <https://doi.org/10.18653/v1/2020.acl-main.463>
- Berglund, Lukas, Meg Tong, Max Kaufmann, Mikita Balesni, Asa Cooper Stickland, Tomasz Korbak, and Owain Evans. 2023. The reversal curse: LLMs trained on “A is B” fail to learn “B is A”. *ArXiv preprint*, arXiv:2309.12288.
- Bieber, Celeste. 2023. ChatGPT broke the Turing test—the race is on for new ways to assess AI. *Nature (News Feature)*, 619:686–689. <https://doi.org/10.1038/d41586-023-02361-7>
- Bisk, Yonatan, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, and Turian Joseph. 2020. Experience grounds language. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8718–8735. <https://doi.org/10.18653/v1/2020.emnlp-main.703>
- Chakraborty, Mohna, Adithya Kulkarni, and Qi Li. 2023. Zero-shot approach to overcome perturbation sensitivity of prompts. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5698–5711. <https://doi.org/10.18653/v1/2023.acl-long.313>
- Chang, Tyler A. and Benjamin K. Bergen. 2023. Language model behavior: A comprehensive survey. *ArXiv preprint*, arXiv:2303.11504.
- Chen, Jifan, 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. <https://doi.org/10.18653/v1/2021.findings-emnlp.324>
- Christiansen, Morten H. and Nick Chater. 1999. Toward a connectionist model of recursion in human linguistic performance. *Cognitive Science*, 23(2):157–205. <https://doi.org/10.1207/s15516709cog2302.2>
- Conneau, Alexis, 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. <https://doi.org/10.18653/v1/D18-1269>
- Contreras Kallens, Pablo, Ross Deans Kristensen-McLachlan, and Morten H. Christiansen. 2023. Large language models demonstrate the potential of statistical learning in language. *Cognitive Science*, 47(3):e13256. <https://doi.org/10.1111/cogs.13256>, PubMed: 36840975
- Dai, Damai, Dong Li, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2022. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8493–8502. <https://doi.org/10.18653/v1/2022.acl-long.581>
- Dankers, Verna, Anna Langedijk, Kate McCurdy, Adina Williams, and Dieuwke Hupkes. 2021. Generalising to German plural noun classes, from the perspective of a recurrent neural network. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 94–108. <https://doi.org/10.18653/v1/2021.conll-1.8>
- Du, Mengnan, Fengxiang He, Zou Na, Dacheng Tao, and Hu Xia. 2023. Shortcut learning of large language models in natural language understanding. *ArXiv preprint*, arXiv:2208.11857. <https://doi.org/10.1145/3596490>
- Dupoux, Emmanuel. 2018. Cognitive science in the era of artificial intelligence: A roadmap for reverse-engineering the infant language-learner. *Cognition*, 173:43–59. <https://doi.org/10.1016/j.cognition.2017.11.008>, PubMed: 29324240
- Elazar, Yanai, 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. <https://doi.org/10.1162/tacl.a.00410>
- Elman, Jeffrey L. 1990. Finding structure in time. *Cognitive Science*, 14(2):179–211. [https://doi.org/10.1016/0364-0213\(90\)90002-E](https://doi.org/10.1016/0364-0213(90)90002-E)
- Francis, Wendy S. 2009. Bilingual semantic and conceptual representation. In *Handbook of Bilingualism*, pages 251–267, Oxford University Press. <https://doi.org/10.1093/oso/9780195151770.003.0014>
- Frank, Stefan and Rens Bod. 2011. Insensitivity of the human sentence-processing system to hierarchical structure. *Psychological Science*, 22(6):829–834. <https://doi.org/10.1177/0956797611409589>, PubMed: 21586764
- Frege, Gottlob. 1892. Über Sinn und Bedeutung [“On sense and reference”]. *Zeitschrift für Philosophie und philosophische Kritik*, 100(1):25–50.
- Frege, Gottlob. 1918–1919. The thought: A logical inquiry [“Der Gedanke. Eine logische Untersuchung”]. *Beiträge Zur Philosophie des Deutschen Idealismus*, 2:58–77.
- Futrell, Richard and Roger Levy. 2017. Noisy-context surprisal as a human sentence processing cost model. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 688–698. <https://doi.org/10.18653/v1/E17-1065>
- Gentner, Dedre and Christian Hoyos. 2017. Analogy and abstraction. *Topics in Cognitive Science*, 9(3):672–693. <https://doi.org/10.1111/tops.12278>
- Geva, Mor, Yoav Goldberg, and Jonathan Berant. 2019. Are we modeling the task or the annotator? An investigation of annotator bias in natural language understanding datasets. 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 1161–1166. <https://doi.org/10.18653/v1/D19-1107>
- Geva, Mor, Roei Schuster, Jonathan Berant, and Levy Omer. 2021. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5484–5495. <https://doi.org/10.18653/v1/2021.emnlp-main.446>
- Giulianelli, Mario, Jack Harding, Florian Mohnert, Dieuwke Hupkes, and Willem Zuidema. 2018. Under the hood: Using diagnostic classifiers to investigate and improve how language models track agreement information. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 240–248. <https://doi.org/10.18653/v1/W18-5426>
- Gulordava, Kristina, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco

- Baroni. 2018. Colorless green recurrent networks dream hierarchically. 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 1195–1205. <https://doi.org/10.18653/v1/N18-1108>
- Gururangan, Suchin, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112. <https://doi.org/10.18653/v1/N18-2017>
- Hagström, Lovisa, Denitsa Saynova, Tobias Norlund, Moa Johansson, and Richard Johansson. 2023. The effect of scaling, retrieval augmentation and form on the factual consistency of language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5457–5476. <https://doi.org/10.18653/v1/2023.emnlp-main.332>
- Heineman, David. 2023. Rethinking reasoning evaluation with theories of intelligence, pages 1–10. <https://davidheineman.com/reasoningevaluation.pdf>
- Hendrycks, Dan, 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)*, pages 1–27.
- Hernandez, A., P. Li, and B. Macwhinney. 2005. The emergence of competing modules in bilingualism. *Trends in Cognitive Sciences*, 9(5):220–225. <https://doi.org/10.1016/j.tics.2005.03.003>, PubMed: 15866148
- Hosseini, Arian, Siva Reddy, Dzmitry Bahdanau, R. Devon Hjelm, Alessandro Sordani, 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. <https://doi.org/10.18653/v1/2021.naacl-main.102>
- Hu, Junjie, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4411–4421.
- Huang, Kuan Jung, Suhas Arehalli, Mari Kugemoto, Christian Muxica, Grusha Prasad, Brian Dillon, and Tal Linzen. 2023. Surprisal does not explain syntactic disambiguation difficulty: Evidence from a large-scale benchmark. *PsyArXiv Preprint*, z38u6. <https://doi.org/10.31234/osf.io/z38u6>
- Hupkes, Dieuwke. 2020. *Hierarchy and Interpretability in Neural Models of Language Processing*. Ph.D. thesis, University of Amsterdam.
- Hupkes, Dieuwke, 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. A taxonomy and review of generalization research in NLP. *Nature Machine Intelligence*, 5(10):1161–1174. <https://doi.org/10.1038/s42256-023-00729-y>
- Izacard, Gautier, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2023. Atlas: Few-shot learning with retrieval augmented language models. *Journal of Machine Learning Research*, 24(251):1–43.
- Jang, Myeongjun, 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.
- Jang, Myeongjun and Thomas Lukasiewicz. 2023. Consistency analysis of ChatGPT. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15970–15985. <https://doi.org/10.18653/v1/2023.emnlp-main.991>
- Johnson-Laird, Philip N. and Marco Ragni. 2023. What should replace the Turing Test? *Intelligent Computing*, 2:1–2. <https://doi.org/10.34133/icomputing.0064>

- Jumelet, Jaap, Milica Denic, Jakub Szymanik, Dieuwke Hupkes, and Shane Steinert-Threlkeld. 2021. Language models use monotonicity to assess NPI licensing. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4958–4969. <https://doi.org/10.18653/v1/2021.findings-acl.439>
- Kassner, Nora 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. <https://doi.org/10.18653/v1/2020.acl-main.698>
- Kassner, Nora, 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. <https://doi.org/10.18653/v1/2021.emnlp-main.697>
- Kocijan, Vid, Ernest Davis, Thomas Lukasiewicz, Gary Marcus, and Leora Morgenstern. 2023. The defeat of the Winograd Schema Challenge. *ArXiv preprint*, arXiv:2201.02387. <https://doi.org/10.1016/j.artint.2023.103971>
- Kroll, Judith F. and Annette M. B. De Groot. 1997. Lexical and conceptual memory in the bilingual: Mapping form to meaning in two languages. In *Tutorials in Bilingualism*. Lawrence Erlbaum, pages 169–199.
- Lakretz, Yair, Théo Desbordes, Dieuwke Hupkes, and Stanislas Dehaene. 2022. Can transformers process recursive nested constructions, like humans? In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3226–3232.
- Lakretz, Yair, Dieuwke Hupkes, Alessandra Vergallito, Marco Marelli, Marco Baroni, and Stanislas Dehaene. 2021. Mechanisms for handling nested dependencies in neural-network language models and humans. *Cognition*, 213:104699. Special Issue in Honour of Jacques Mehler, *Cognition's* founding editor. <https://doi.org/10.1016/j.cognition.2021.104699>
- Lakretz, Yair, German Kruszewski, Theo Desbordes, Dieuwke Hupkes, Stanislas Dehaene, and Marco Baroni. 2019. The emergence of number and syntax units in LSTM language models. 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 11–20. <https://doi.org/10.18653/v1/N19-1002>
- Li, Tao, Vivek Gupta, Maitrey Mehta, and Srikumar Vivek. 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. <https://doi.org/10.18653/v1/D19-1405>
- Liang, Percy, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yan Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Zhang Ce, Christian Alexander Cosgrove, Christopher D Manning, Christopher Re, Diana Acosta-Navas, Drew Arad Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri S. Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Andrew Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2023. Holistic evaluation of language models. *Transactions on Machine Learning Research*, pages 1–162.
- Liang, Yaobo, Nan Duan, Yeyun Gong, Wu Ning, 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, Wu Winnie, Shuguang Liu, Fan Yang, Daniel Campos, Rangan Majumder, and Ming Zhou. 2020. XGLUE: A new benchmark dataset for cross-lingual pre-training, understanding and generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6008–6018. <https://doi.org/10.18653/v1/2020.emnlp-main.484>
- Lin, Chin Yew. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81.
- Linzen, Tal and Marco Baroni. 2021. Syntactic structure from deep learning. *Annual*

- Review of Linguistics*, 7:195–212. <https://doi.org/10.1146/annurev-linguistics-032020-051035>
- Linzen, Tal, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535. https://doi.org/10.1162/tacl_a_00115
- Liu, Yunzhe, Raymond J. Dolan, Zeb Kurth-Nelson, and Timothy E. J. Behrens. 2019. Human replay spontaneously reorganizes experience. *Cell*, 178(3):640–652. <https://doi.org/10.1016/j.cell.2019.06.012>, PubMed: 31280961
- Mahowald, Kyle, Anna A. Ivanova, Idan A. Blank, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. 2023. Dissociating language and thought in large language models. *ArXiv preprint*, arXiv:2301.06627. <https://doi.org/10.1016/j.tics.2024.01.011>
- Malouf, Robert. 2017. Abstractive morphological learning with a recurrent neural network. *Morphology*, 27:431–458. <https://doi.org/10.1007/s11525-017-9307-x>
- Mandelkern, Matthew and Tal Linzen. 2023. Do language models refer? *ArXiv preprint*, arXiv:2308.05576.
- McCoy, R., Shunyu Yao, Dan Friedman, Matthew Hardy, and Thomas L. Griffiths. 2023. Thomas Embers of autoregression: Understanding large language models through the problem they are trained to solve. *ArXiv preprint*, arXiv:2309.13638. <https://doi.org/10.1073/pnas.2322420121>
- McCoy, Tom, 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. <https://doi.org/10.18653/v1/P19-1334>
- McKenna, Nick, Tianyi Li, Liang Cheng, Mohammad Javad Hosseini, Mark Johnson, and Mark Steedman. 2023. Sources of hallucination by large language models on inference tasks. *ArXiv preprint*, arXiv:2303.06273. <https://doi.org/10.18653/v1/2023.findings-emnlp.182>
- McKenzie, Sam, Andrea J. Frank, Nathaniel R. Kinsky, Blake Porter, Pamela D. Rivière, and Howard Eichenbaum. 2014. Hippocampal representation of related and opposing memories develop within distinct, hierarchically organized neural schemas. *Neuron*, 83(1):202–215. <https://doi.org/10.1016/j.neuron.2014.05.019>, PubMed: 24910078
- Minervini, Pasquale 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. <https://doi.org/10.18653/v1/K18-1007>
- Mitchell, Eric, Joseph Noh, Siyan Li, Will Armstrong, Ananth Agarwal, Patrick Liu, Chelsea Finn, and Manning Christopher. 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. <https://doi.org/10.18653/v1/2022.emnlp-main.115>
- Mitchell, Melanie 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. <https://doi.org/10.1073/pnas.2215907120>, PubMed: 36943882
- Mizrahi, Moran, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. 2023. State of what ar? A call for multi-prompt LLM evaluation. *ArXiv preprint*, arXiv:2401.00595. https://doi.org/10.1162/tacl_a_00681
- Mollo, Dimitri Coelho and Raphaël Millière. 2023. The vector grounding problem. *ArXiv preprint*, arXiv:2304.01481.
- Nie, Yixin, 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. <https://doi.org/10.18653/v1/2020.acl-main.441>
- Niven, Timothy and Hung-Yu Kao. 2019. Probing neural network comprehension of natural language arguments. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4658–4664. <https://doi.org/10.18653/v1/P19-1459>
- Ohmer, Xenia, 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.
- Ouyang, Long, Wu Jeffrey, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744.
- Papineni, Kishore, 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. <https://doi.org/10.3115/1073083.1073135>
- Patel, Roma and Pavlick, Ellie. 2022. Mapping language models to grounded conceptual spaces. In *International Conference on Learning Representations*, pages 1–21.
- Pavlick, Ellie. 2023. Symbols and grounding in large language models. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 381(2251):20220041. <https://doi.org/10.1098/rsta.2022.0041>, PubMed: 37271171
- Piantadosi, Steven. 2023. Modern language models refute Chomsky’s approach to language. *Lingbuzz preprint*, 7180.
- Piantadosi, Steven and Felix Hill. 2022. Meaning without reference in large language models. In *NeurIPS 2022 Workshop on Neuro Causal and Symbolic AI (nCSI)*, pages 1–8.
- Podkorytov, Maksim, Daniel Biś, and Xiuwen Liu. 2021. How can the [mask] know? The sources and limitations of knowledge in BERT. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. <https://doi.org/10.1109/IJCNN52387.2021.9534299>
- Ponti, Edoardo Maria, Gora 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. <https://doi.org/10.18653/v1/2020.emnlp-main.185>
- Qi, Jirui, Raquel Fernández, and Arianna Bisazza. 2023. Cross-lingual consistency of factual knowledge in multilingual language models. *ArXiv preprint*, arXiv:2310.10478. <https://doi.org/10.18653/v1/2023.emnlp-main.658>
- Raji, Inioluwa Deborah, 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*, pages 1–17.
- Ray Choudhury, Sagnik, 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.
- Rei, Ricardo, José G. C. 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.
- Roemmele, Melissa, Cosmin Adrian Bejan, and Andrew S. Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *Proceedings of the 2011 AAAI Spring Symposium Series*, pages 90–95.
- Ruder, Sebastian, 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. <https://doi.org/10.18653/v1/2021.emnlp-main.802>
- Ryu, Soo Hyun and Richard Lewis. 2021. Accounting for agreement phenomena in sentence comprehension with transformer language models: Effects of similarity-based interference on surprisal and attention. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 61–71. <https://doi.org/10.18653/v1/2021.cmc1-1.6>
- Saxon, Michael, Xinyi Wang, Xu Wenda, and William Yang Wang. 2023. PECO:

- Examining single sentence label leakage in natural language inference datasets through progressive evaluation of cluster outliers. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3061–3074. <https://doi.org/10.18653/v1/2023.eacl-main.223>
- Sen, Priyanka and Amir Saffari. 2020. What do models learn from question answering datasets? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2429–2438. <https://doi.org/10.18653/v1/2020.emnlp-main.190>
- Sun, Kaiser, Adina Williams, and Dieuwke Hupkes. 2023. The validity of evaluation results: Assessing concurrence across compositionality benchmarks. In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 274–293. <https://doi.org/10.18653/v1/2023.conll-1.19>
- Tafjord, Oyvind and Peter Clark. 2021. General-purpose question-answering with Macaw. *ArXiv preprint*, arXiv:2109.02593.
- Tenenbaum, Joshua B., Charles Kemp, Thomas L. Griffiths, and Noah D. Goodman. 2011. How to grow a mind: Statistics, structure, and abstraction. *Science*, 331(6022):1279–1285. <https://doi.org/10.1126/science.1192788>, PubMed: 21393536
- Timkey, William and Tal Linzen. 2023. A language model with limited memory capacity captures interference in human sentence processing. *ArXiv preprint*, arXiv:2310.16142. <https://doi.org/10.18653/v1/2023.findings-emnlp.582>
- Touvron, Hugo, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *ArXiv preprint*, arXiv:2302.13971.
- Ulmer, Dennis, Dieuwke Hupkes, and Elia Bruni. 2019. Assessing incrementality in sequence-to-sequence models. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepLanLP-2019)*, pages 209–217. <https://doi.org/10.18653/v1/W19-4324>
- Vaidya, Avinash R., Henry M. Jones, Johann Castillo, and David Badre. 2021. Neural representation of abstract task structure during generalization. *eLife*, 10:e63226. <https://doi.org/10.7554/eLife.63226>, PubMed: 33729156
- Van Schijndel, Marten and Tal Linzen. 2018. Modeling garden path effects without explicit hierarchical syntax. In *Proceedings of the 40th Annual Meeting of the Cognitive Science Society (CogSci)*, pages 2603–2608.
- Van Schijndel, Marten and Tal Linzen. 2021. Single-stage prediction models do not explain the magnitude of syntactic disambiguation difficulty. *Cognitive Science*, 45(6):e12988. <https://doi.org/10.1111/cogs.12988>, PubMed: 34170031
- Wang, Alex, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems (NeurIPS)*, pages 1–15.
- Wang, Alex, 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. <https://doi.org/10.18653/v1/W18-5446>
- Wang, Boxin, Xu Chejian, Shuohang Wang, Zhe Gan, Yu Cheng, Jianfeng Gao, Ahmed Hassan Awadallah, and Li Bo. 2021. Adversarial GLUE: A multi-task benchmark for robustness evaluation of language models. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, pages 1–13.
- Wang, Haohan, Sun Da, and Eric P. Xing. 2019. What if we simply swap the two text fragments? A straightforward yet effective way to test the robustness of methods to confounding signals in natural language inference tasks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7136–7143. <https://doi.org/10.1609/aaai.v33i01.33017136>
- Wang, Haoyu, Guozheng Ma, Y u Cong, Ning Gui, Linrui Zhang, Zhiqi Huang, Suwei Ma, Yongzhe Chang, Sen Zhang, Shen Li, Xueqian Wang, Peilin Zhao, and Dacheng Tao. 2023. Are large language models really robust to word-level perturbations? *ArXiv preprint*, arXiv:2309.11166.

- Wang, Xiaozhi, Kaiyue Wen, Zhengyan Zhang, Lei Hou, Zhiyuan Liu, and Li Juanzi. 2022. Finding skill neurons in pre-trained transformer-based language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11132–11152. <https://doi.org/10.18653/v1/2022.emnlp-main.765>
- Warstadt, Alex and Samuel R. Bowman. 2022. What artificial neural networks can tell us about human language acquisition. *ArXiv preprint*, arXiv:2208.07998. <https://doi.org/10.1201/9781003205388-2>
- Warstadt, Alex, Leshem Choshen, Aaron Mueller, Adina Williams, Ethan Wilcox, and Chengxu Zhuang. 2023. Call for papers – The BabyLM Challenge: Sample-efficient pretraining on a developmentally plausible corpus. *ArXiv preprint*, arXiv:2301.11796. <https://doi.org/10.18653/v1/2023.conll-babylm.1>
- Weber, Lucas, 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. <https://doi.org/10.18653/v1/2023.conll-1.20>
- Wilcox, Ethan, Jonathan Gottlieb, Jon Gauthier, Hu Jennifer, Peng Qian, and Roger Levy. 2020. On the predictive power of neural language models for human real-time comprehension behavior. *ArXiv preprint*, arXiv:2006.01912.
- Wittgenstein, Ludwig. 1953. *Philosophical investigations*. *Philosophische Untersuchungen*. Macmillan.
- Yang, Yinfei, 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. <https://doi.org/10.18653/v1/D19-1382>
- Zhang, Yian, Alex Warstadt, Xiaocheng Li, and Samuel R. Bowman. 2021. When do you need billions of words of pretraining data? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1112–1125. <https://doi.org/10.18653/v1/2021.acl-long.90>
- Zhang, Yuan, 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. <https://doi.org/10.18653/v1/N19-1131>