Holmes
Benchmark the Linguistic Competence of Language Models

Anonymous ACL submission

Abstract

We introduce Holmes, a benchmark to assess the *linguistic competence* of language models (LMs) – their ability to grasp linguistic phenomena. Unlike prior prompting-based evaluations, Holmes assesses the linguistic competence of LMs via their internal representations using classifier-based probing. In doing so, we disentangle specific phenomena (e.g., part-of-speech of words) from other cognitive abilities, like following textual instructions, and meet recent calls to assess LMs' linguistic competence in isolation. Composing Holmes, we review over 250 probing studies and feature more than 200 datasets to assess *syntax, morphology, semantics, reasoning*, and *discourse* phenomena. Analyzing over 50 LMs reveals that, aligned with known trends, their linguistic competence correlates with model size. However, surprisingly, model architecture and instruction tuning also significantly influence performance, particularly in *morphology* and *syntax*. Finally, we propose FlashHolmes, a streamlined version of Holmes designed to lower the high computation load while maintaining high-ranking precision.

holmes-benchmark.github.io

1 Introduction

Linguistic competence is the unconscious understanding of language, like grasping grammatical rules (Chomsky, 1965). As language models (LMs) are trained on simple tasks like next word prediction (Brown et al., 2020), one might naturally wonder: *What is the linguistic competence of LMs, and how do they differ?* To answer such questions, benchmarks estimate cognitive abilities by providing textual instructions and evaluate LMs' responses, as done for mathematical reasoning (Cobbe et al., 2021) or factual knowledge (Petroni et al., 2019, 2020). However, they conflate latent abilities (like following provided instructions) with those under test, such as understanding specific linguistic phenomena, e.g., syntactic structures (Liang et al., 2023). As this entanglement makes it infeasible to draw definitive conclusions about distinct abilities (Hu and Levy, 2023), recent studies call to assess the linguistic competence of LMs comprehensively and in isolation (Lu et al., 2023; Mahowald et al., 2024).

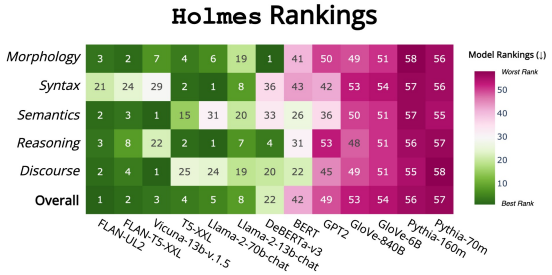

Figure 1: A subset of Holmes rankings (↓) for various evaluated LMs. FLAN-UL2 outperforms the others *overall*, while different LMs prevail for the five distinct types of linguistic phenomena.

In this work, we introduce the Holmes (Figure 2). A benchmark to assess the linguistic competence of LMs (Figure 1) regarding numerous linguistic phenomena. To fully disentangle the understanding of these phenomena and other abilities of LMs, we use classifier-based probing (Tenney et al., 2019a; Hewitt and Manning, 2019; Belinkov, 2022). A method that uses the LMs' internal representations of text inputs to train linear models (probes) to predict specific aspects of phenomena, such as words' part-of-speech (POS). We then ap-

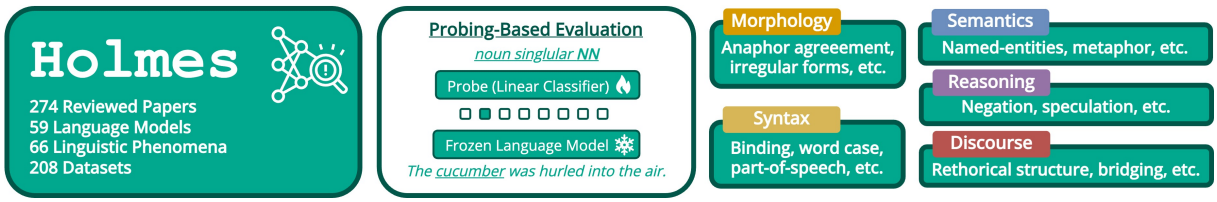

Figure 2: Overview of Holmes (left) with the five phenomena types (right) and an example of probing-based evaluations for part-of-speech: encoding the input tokens and predicting the POS tag for *cucumber*, here *NN*.

proximate the LMs' grasp of these phenomena using the probes' performance, rigorously verified using control tasks (Hewitt and Liang, 2019) and from an information theory perspective (Voita and Titov, 2020). With this particular and comprehensive scope, we thoroughly address the initially raised questions as follows:

Meta-Study (§ 3) The review of over 270 probing studies reveals a gap in comprehensively evaluating linguistic competence. Despite covering over 200 probing tasks and 150 LMs, individual studies focus on particular tasks and LMs. As a result, only three LMs were probed on over 20% of the tasks, and one single task was evaluated for more than 20% of the reviewed LMs. Notably, recent large LMs are significantly underrepresented.

Benchmark (§ 4) To address this identified deficiency, Holmes offers a structured framework to assess the English linguistic competence of LMs comprehensively. It features 208 distinct datasets covering *morphology*, *syntax*, *semantics*, *reasoning*, and *discourse* phenomena, including previously underrepresented ones like negation or rhetoric in text (Liang et al., 2023).

Results and Analysis (§ 5) From assessing 59 LMs (Figure 1), we find that no single one consistently excels the others and that their linguistic competence is more pronounced for *morphology* and *syntax* than the other phenomena types. Instead, we find **model size**, **model architecture**, and **instruction tuning** fundamentally affect their linguistic competence.

First, LMs' linguistic competence, particularly for *morphology* and *syntax*, scales with their **model size**. This generalizes previous findings (Tenney et al., 2019b; Zhang et al., 2021) beyond LMs with 350 million parameters. Second, contrary to prompting evaluations (Lu et al., 2023) and aligned with other work (Waldis et al., 2024a; Gautam et al., 2024), **model architecture** is critical. The linguistic competence of decoder-only LMs is less pronounced, and even 70 billion does not allow them to encode linguistic phenomena of words with comparable strength to encoder-only LMs of a similar size. Third, while previous studies focused on aligning LMs with human interactions through **instruction tuning** (Ouyang et al., 2022; Touvron et al., 2023; Zhou et al., 2023), we show for the first time its effect on their linguistic competence. It improves *morphology* and *syntax* but has mixed effects for the other types of phenomena. Lastly, we contrast the results of Holmes with OpenLLM (Beeching et al., 2023), an extensive LM benchmark focusing on user-centered applications like mathematical reasoning. We find that Holmes provides a unique but supplementary perspective, as rankings partly align, especially for reasoning-related phenomena.

Efficiency (§ 6) Finally, to mitigate the heavy computational burden of evaluating a new LM on Holmes, we form the streamlined version FlashHolmes by selectively excluding samples not significantly influencing overall rankings (Perlitz et al., 2023). Specifically, FlashHolmes approximates Holmes rankings with high precision while requiring only ~3% of the computation.

We summarize our contributions as follows:

- **Benchmark.** Holmes comprehensively and thoroughly assesses the linguistic competence of LMs in isolation, providing substantial ground for advancements in NLP.
- **Empirical insights.** Extensive experiments reveal that LMs' linguistic competence is more pronounced for *morphology* and *syntax*, and size, architecture, and instruction tuning are crucial for LM differences.
- **Ease of use.** We provide tools to interactively explore Holmes results and straightforward code to evaluate upcoming LMs with efficiency in mind (FlashHolmes)[1].

[1]Find resources at https://holmes-benchmark.github.io

2 Preliminaries

Language Models (LMs) Language Models compute probabilities for word sequences i, enabling tasks such as classifying i, textual comparisons between i and another sequence i', and text generation based on i. We consider LMs as any model producing representations of input i, regardless of their specific type: **sparse** like bag-of-words (Harris, 1954); **static** such as GloVe (Pennington et al., 2014); or **contextualized** transformer-based LMs (Devlin et al., 2019; Raffel et al., 2020).

Linguistic Competence Following Chomsky (1965), linguistic competence is defined as the unconscious knowledge of language, encompassing the understanding of specific linguistic phenomena, including word dependencies and their distinct parts of speech (POS).

Linguistic Phenomena We define the linguistic competence of LMs as their ability to understand a diversity of linguistic phenomena. Specifically, we focus on five phenomena types: *morphology*, the structure of words; *syntax*, the structure of sentences; *semantics*, the meaning of words; *reasoning*, the use of words in logical deduction and other related phenomena like negation or speculation; *discourse*, the context in text like rhetorical structure. Following Mahowald et al. (2024), we categorize these phenomena types into two groups: *morphology* and *syntax* are **formal** phenomena, which include understanding grammatical rules and statistical patterns, while **functional** ones (*semantics*, *reasoning*, and *discourse*) focus on practical abilities like interpreting text sentiment or detecting the existence of speculation.

Datasets We define a dataset as text examples and labels covering a specific aspect of a linguistic phenomenon, like words and their POS tag. Typically, these labels are highly unambiguous to assess the specific aspect under test in isolation.

Probes Using probes, we empirically assess the linguistic competence of LMs regarding the featured linguistic phenomena in `Holmes`. To this end, we employ probing tasks using the widely recognized classifier-based probing method (Tenney et al., 2019a; Hewitt and Manning, 2019; Belinkov, 2022), or known as diagnostic classifiers (Veldhoen et al., 2016; Giulianelli et al., 2018). Running such a probing task involves training a probe (linear model) using the specific dataset to test a distinct aspect of a linguistic phenomenon in isolation. Therefore, we feed the text examples, encoded with a given LM, as training inputs. Subsequently, we use the probe's performance to approximate how an LM understands the specific linguistic phenomenon under test. With a higher score, we assume the embeddings embody patterns relevant to this phenomenon, which enhances the accuracy (Tenney et al., 2019b).

3 Meta-Study

In this section, we survey 274 studies (§ 3.1), probing LMs' linguistic competence. We analyze these studies regarding their evolution, covered probing tasks and LMs (§ 3.2), and identify the apparent need for consolidating existing resources (§ 3.3).

3.1 Scope

We analyze 28k papers (P) from 2015 to August 2023 of major NLP conferences (TACL, ACL, AACL, COLING, EACL, EMNLP, NAACL, and corresponding workshops) expanded with selected work from other venues such as ICLR. To identify relevant work, we follow a semiautomatic approach. First, we automatically select papers based on their meta-data and full text.[2] We select a total of 493 candidate papers matching at least one of the following three criteria ($P' = \{\forall p \in P | p \in P_1 \cup p \in P_2 \cup p \in P_3\}$):

$\mathbf{P_1}$: papers contain *probing* or *probe* in the title.

$\mathbf{P_2}$: papers contain *probing* or *probe* in the abstract and at least five times in the main content.

$\mathbf{P_3}$: papers contain *probing* or *probe* at least ten times in the main content.

We manually verified these automatically curated candidates (P') and found 274 relevant papers (P_r).We selected them as they either evaluate LMs regarding one or more linguistic phenomena as part of the analysis or as a main contribution. This involves filtering papers using the term *probing* in other senses, such as *probing hash tables* (Bogoychev and Lopez, 2016).

3.2 Analysis

Next, we analyze these 274 relevant studies (P_r).

i) Scattered evolution calls for consolidation. First, we analyze the evolution of the relevant studies. Figure 3 relates how these studies cite each other (**probing citations** C_p) compared to other

[2]We use PyPDF2 v3.0.0, DBLP and semanticscholar API.

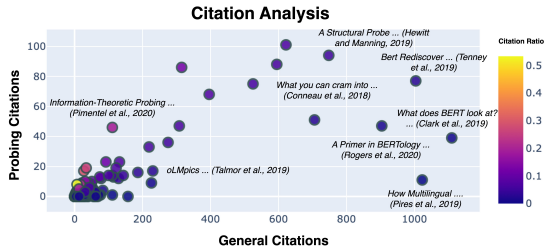

Figure 3: Citation analysis considering *probing citations* originating from the set of relevant work and every other citation (*general citations*). The color scale indicates the ratio (α) between them.

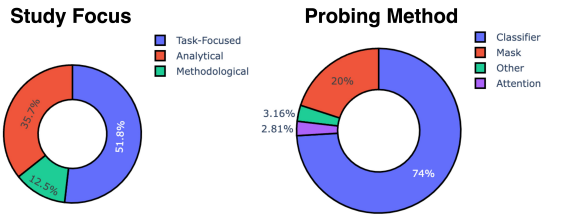

Figure 4: Categorization of the selected studies by their focus and their conducted probing method.

gathered citations (**general citations** C_g). Colorized, we show the ratio α between these two measures $\alpha = \frac{|C_p|+1}{|C_g|+1}$. First, only a fraction of the works gained general attention, as 16 papers exceeded 200 general citations. Further, probing works cite each other rather sparsely, with an average probing citation ratio of $\alpha = 0.1$. Therefore, we see other fields are paying little attention to the linguistic competence of LMs. Paired with scattered citation patterns, we identify the need to consolidate existing resources to solidly ground research in this field.

ii) Probing work prioritizes tasks and analytics over methods. We categorize the selected work according to their probing focus: **methodological**, new methods, like control tasks (Hewitt and Liang, 2019) or minimum description length (Voita and Titov, 2020); **task-focused** assessing specific linguistic phenomena as main contributions, such as discourse relations in text (Koto et al., 2021); and **analytical** using probing tasks to analyze LMs, such as the impact of pre-training data (Zhang et al., 2021). Figure 4 shows: the majority (51.8%) of studies focus on specific probing tasks like numeric scales (Zhang et al., 2020), or morphosyntactic (Shapiro et al., 2021); 35.7% use probing as a supplementary analytical tool, for example, analyzing the effect of fine-tuning (Mosbach et al., 2020a; Zhu et al., 2022a); 12.5% address methodological problems related to probing (Wu et al., 2020; Immer et al., 2022; Zhu et al., 2022b).

iii) The dominance of classifier-based probing. Next, we analyze the specific employed probing method: **classifier**, using linear or shallow models to probe internal representations of LMs, as demonstrated in Tenney et al. (2019a); **mask**, letting LMs fill gaps to verify linguistic phenomena, as shown in Talmor et al. (2020) or Warstadt et al. (2020); **attention**, which relies on attention patterns, as used in Pandit and Hou (2021) for bridging; and **other**, methods not belonging to the previous three categories, such as dimension selection (Torroba Hennigen et al., 2020). Most studies utilize the classifier-based probing method (74%), 20% conduct mask-based probing, and only a minority of work ($\sim$ 3%) considers attention patterns or other approaches.

iv) Tasks and LMs are barely broadly evaluated. Finally, we analyze which tasks and LMs the relevant probing studies consider. For example, Tenney et al. (2019b) considers BERT and probes POS tagging, semantic-role labeling (SRL), and other ones. Aggregated over all studies, we found a broad coverage of 289 unique tasks and 161 distinct LMs. Below, we delve into the details and highlight noteworthy findings.

We analyze how LMs and tasks are considered jointly in Figure 5. Despite the broad coverage, single studies, including fundamental ones, maintain a particular focus and consider only a fraction of LMs and tasks. For example, while most tasks (72%) were assessed on BERT, RoBERTa's coverage has already declined to 42%. Conversely, part-of-speech tagging (POS), the most probed task, was only evaluated on 23% of the LMs, for example, not covering prominent examples like BART (Lewis et al., 2020). Notably, more recently released larger and powerful LMs, like PYTHIA (Biderman et al., 2023), UL2 (Tay et al., 2023), or LLAMA-2 (Touvron et al., 2023), and instruction-tuned LMs (FLAN-T5 (Chung et al., 2022), LLAMA-2-Chat (Touvron et al., 2023), or TK-Instruct (Wang et al., 2022) are missing almost entirely, with single more recent exceptions (Hu and Levy, 2023; Waldis et al., 2024a). Again, these insights underscore the need to consolidate existing resources for more dense coverage. This is further evident when considering Figure 5, where we sort LMs and tasks according to how often they were mentioned in the relevant works. Then, we plot

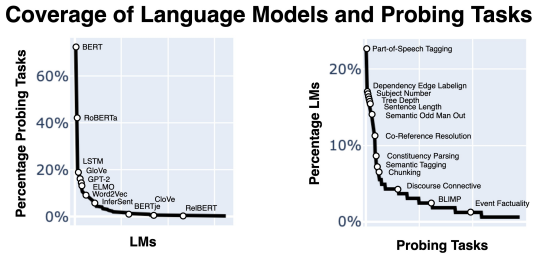

Figure 5: Overview of how many tasks single LMs cover and vice versa - single examples are highlighted.

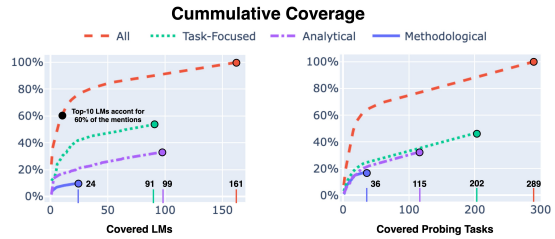

Figure 6: Cumulative coverage of LMs and tasks, considering all relevant studies and their focus.

their cumulative coverage concerning all mentions. For example, considering all studies (red line), the top-10 most mentioned LMs account for 80% of all LMs mentions (black dot). In contrast, the other 151 unique LMs account for only 40%. Comparing the paper focus, we see that methodological studies rely only on 24 LMs and 36 tasks. In contrast, task-focused and analytical work covers a similar number of LMs (91 and 99, respectively). However, due to their distinct focus, task-focused studies cover significantly more tasks (202) than analytical ones (115).

3.3 Summary

This meta-study emphasizes the need to consolidate existing resources for a comprehensive assessment of the linguistic competence of LMs — a manifold but rather blind spot in evaluation research. Apart from more thorough evaluations, such a stimulus can significantly boost future research, as happened in computer vision with ImageNet (Deng et al., 2009) or in NLP with GLUE and SuperGLUE (Wang et al., 2019a,b).

4 Holmes Benchmark

With `Holmes`, we provide an extensive ground to tackle these identified deficiencies in the existing literature and comprehensively investigate the English linguistic competence of LMs. Specifically, `Holmes` features 208 datasets addressing distinct aspects of 66 phenomena covering *morphology*, *syntax*, *semantic*, *reasoning*, and *discourse*.

4.1 Datasets

To feature a total of 208 unique datasets, we leverage existing and established resources like OntoNotes (Weischedel et al., 2013), English Web Treebank (Silveira et al., 2014), or BLIMP (Warstadt et al., 2020) and create datasets addressing phenomena like the POS of words, their dependencies or determine the linguistic acceptability of sentences. Further, we include a range of less employed data, addressing contextualization of words (Klafka and Ettinger, 2020), reasoning (Talmor et al., 2020), semantic decomposition (White et al., 2016; Rudinger et al., 2018a,b; Govindarajan et al., 2019; Vashishtha et al., 2019), grammatical knowledge (Huebner et al., 2021), bridging (Pandit and Hou, 2021), and rhetorical (Carlson et al., 2001) and discourse (Webber et al., 2019) structure in text. Finally, we cover rarely probed phenomena like negation (Szarvas et al., 2008; Konstantinova et al., 2012; Vahtola et al., 2022), or word complexity (Paetzold and Specia, 2016).

4.2 Structure

Apart from the comprehensive scope, `Holmes` provides a clear structure for specific evaluations on different levels of aggregation. We first group the datasets according to the linguistic phenomena addressed. Then, we categorize these phenomena into their previously introduced type (see § 2) - *morphology*, *syntax*, *semantics*, *reasoning* and *discourse*. We rely on the categorization provided by the specific studies whenever given. The detailed categorization is given in § A.3.

4.3 Experimental Setup

`Holmes` evaluation follows the primarily used classifier-based probing paradigm, as described in § 2. Considering the internal representations allows us to maximally disentangle the understanding of distinct linguistic phenomena from each other and from other cognitive abilities (like following textual instructions). Further, this method allows us to assess any type of LMs, including sparse, static, or contextualized ones. Based on the specific dataset, we either select the embeddings of the specific input tokens (like single words for POS tagging) or average embeddings across a span or the whole sentence. We define a probing task as training a probe f_p (linear model without intermediate layers) using

these embeddings as inputs and the dataset labels as training signals. If not defined in the original data, we divide the dataset samples into train/dev/test split following a ratio of 70/10/20. We repeat this procedure five times using different random seeds and aggregate the results afterward.

4.4 Evaluations

We approximate how well an LM encodes specific linguistic phenomena using the absolute prediction performance of the probes. In addition, we rigorously evaluate the reliability of probing results using control tasks and from an information theory perspective (Voita and Titov, 2020; Hewitt and Liang, 2019). Different from commonly used prompting assessments, this particular evaluation protocol refrains from known fallacies in which the results and conclusions are sensible with specific instructions (Mizrahi et al., 2024; Min et al., 2022) or few-shot examples (Lu et al., 2023).

Task Score Metric Based on a dataset's specific task type, we use a corresponding performance measure, macro F_1 for classification or Pearson correlation for regression. In addition, we calculate the standard deviation σ of the probe across multiple seeds. A lower σ indicates a better encoding of a given linguistic phenomenon since the measurement is robust to noise. Further, we use the task score for ranking-based evaluation of all evaluated LMs $L = \{l_1, ..., l_m\}$ within `Holmes`. We calculate the mean winning rate mwr (in percentage), telling us how many times one LM l_1 wins against others (Liang et al., 2023). With a higher mwr, we assume an LM encodes tested linguistic phenomena better than others.

Compression Next, we evaluate the probes' reliability from an information-theoretic perspective. Following Voita and Titov (2020), we use the compression I to measure how well a probe compresses input data. A higher I means fewer bits are needed, indicating that the given linguistic phenomenon is more clearly encoded in the embeddings.

Selectivity A reliable probe should grasp patterns relevant to the tested phenomena in the internal representations of LMs but should not be able to learn anything else. Therefore, we expect high performance when evaluating the specific dataset but low performance when we randomize training signals. We check this using control tasks introduced in Hewitt and Liang (2019). Specifically, we calculate the selectivity S as the difference between the probe trained with the original labels y and the control task where we train the probe with randomly assigned labels y'. With a higher S, we assume the detected patterns are relevant for the specific phenomena under test, as random patterns do not lead to similar performance.

5 `Holmes` Results

Using `Holmes`, we evaluate a diverse collection of 59 LMs.[3] Using the results of these extensive experiments, we first answer the research question: *what is the linguistic competence of LMs?* In doing so, we discuss the reliability of results (i), the linguistic competence of LMs concerning the unique structure of `Holmes` (ii), and how these results relate to other downstream abilities (iii). Subsequently, we examine *how linguistic competence varies among LMs*, as we find LMs prevailing for different types of linguistic phenomena (Figure 1) and delve into the effects of model architecture (iv), size (v), and instruction tuning (vi).

i) The reliability of `Holmes`. First, we show the reliability of probing-based evaluation, using *deviation* σ, *compression* I, and *selectivity* S results in Figure 7. Single outliers are datasets that are too hard for all LMs, as the sample size is too small, or the linguistic phenomena under test are too complex, as the ability to detect spans causes speculations in a text. We average these metrics for every dataset across all LMs. Note, for *selectivity*, we consider only base-sized model (10m-200m parameters) for computational efficiency.

First, we found a low average deviation ($\sigma = 0.02$), indicating the high reliability of probes across random seeds. These results also highlight the stability of probing results, compared to prompting-based ones where results across many paraphrased prompts lead to a deviation of $\sigma = 0.07$ reported in Mizrahi et al. (2024). Next, substantial compression (average $I = 1.9$) and selectivity (average $S = 0.31$) further confirm the probes' reliability. Interestingly, one identifies two parallel trends for selectivity. Harder datasets with many labels, like POS tagging, are arranged around a selectivity of 0.1 to 0.4 and a task metric of 0.3. In contrast, for easier binary classification tasks (such as linguistic applicability), we observe selectivity around 0.2 to 0.5 and a task metric of 0.6 to 0.9.

[3] Find a complete list in Appendix § A.2.

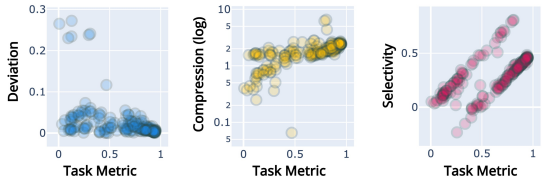

Figure 7: Reliability evaluation using *deviation*, *compression* (log), and *selectivity* on the y-axis for all 208 probing datasets. The x-axis represents the task metrics (either person correlation or macro F_1).

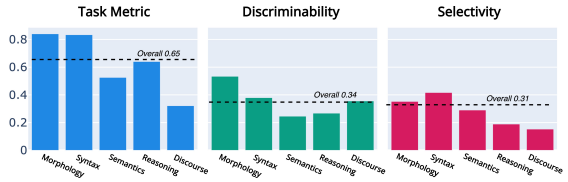

Figure 8: Average *task metric*, *difficulty*, and *discriminability* for each phenomena type. The dashed lines show the average measure over all datasets.

Further, we measure a significant ($p < 0.05$) positive correlation between the task metrics and the compression ($\tau = 0.64$) and selectivity ($\tau = 0.65$). This further confirms our reliability assumption and allows us to trust the task metric as the primary evaluation measure.

ii) The story of `Holmes`. We focus on what `Holmes` tells us in general and regarding formal and functional phenomena, as defined in § 2. We report in Figure 8 the *task metric*, *discriminability*, and *selectivity*, averaged for every phenomena type. Note, discriminability (Rodriguez et al., 2021) quantifies the alignment of LMs ranking of one specific dataset compared to the overall rankings using the Kendall Tau correlation. Considering these three metrics, all tested LMs strongly encode formal phenomena (*morphology* and *syntax*), which often depend on the local neighborhood of words. Therefore, we assume that LMs approximate these co-occurrences during pre-training with high precision. For example, the specific POS tag of a word, like *man* (*noun*), primarily depends on its surroundings, such as the frequent predecessor *the*. In contrast, LMs encode less information about functional phenomena (*semantics*, *reasoning*, and *discourse*) since they show a relatively low performance regarding the task metric. For these functional phenomena, we assume more complex co-occurrences are required to capture the broad context in language, such as the rhetorical relation of two distant text spans. Despite these differences between formal and functional phenomena types, they contribute to the benchmark in a balanced way. A low to medium discriminability indicates that none of these types of linguistic phenomena dominates the overall LM rankings.

This balanced influence of the five phenomena types is further visible when considering their ranking correlations (Figure 9, left). A high average correlation of 67.8 ± 6.6 with the overall results (last column/row) hints that they are facets of a broader occurrence but share common characteristics. Still, breaking into categories is meaningful, as the phenomena types (first five columns/rows) are medium correlated (average of 53.9 ± 14.5). Analyzing the results of phenomena types further highlights the value of this distinction. While results of *morphology* and *syntax* are similarly correlated with the overall results (68.2 and 70.2), their direct correlation (69.1) indicates their supplementary nature. Further, *discourse* results show the lowest correlation with others (44.8 ± 16.1), indicating the particular scope.

iii) The companions of `Holmes`. We analyze how the results of `Holmes` and those from other evaluations focusing on downstream applications align (Figure 9, right). We select the OpenLLM benchmark (Beeching et al., 2023), as it covers a wide range of open LMs, in contrast to others like HELM (Liang et al., 2023). First, `Holmes` and OpenLLM results of jointly evaluated LMs are medium correlated, hinting that the linguistic competence of LMs is partly aligned with their downstream abilities. The nature of this alignment is further evident when focusing on *morphology*, *reasoning*, and *discourse*. Interestingly, and in contrast to *syntax* and *semantics*, their correlation to the OpenLLM and `Holmes` overall results is similar. Therefore, these three phenomena presumably represent skills that are more tested in the general benchmarks. These correlation patterns are consistent across the three most meaningful OpenLLM datasets (*MMLU*, *TruthfulQA*, and *GSM8K*). As *TruthfulQA* shows lower correlations with the linguistic phenomena and other datasets within OpenLLM, we presume this dataset captures distinctly different skills (possibly knowledge).[4] These insights show how different benchmarks provide a different scope and supplement themselves simulta-

[4]Further, it's also known that we need to expect this dataset to be fully leaked (Balloccu et al., 2024).

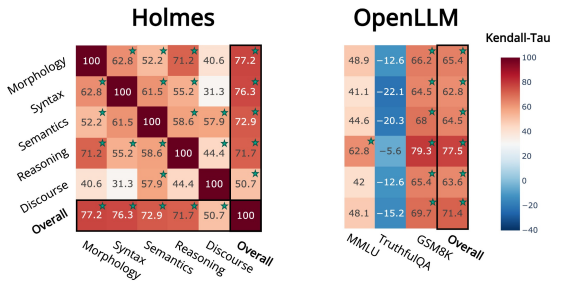

Figure 9: Kendall-tau correlation within `Holmes` (left) and compared to the OpenLLM benchmark (right). Green stars indicate significant correlations ($p < 0.05$).

neously. Further, the above analysis shows, again, the value of assessing the linguistic competence of LMs across different phenomena types, for fine-grained analyses.

iv) The effect of language model architecture. Next, we discuss the impact of model architecture on the linguistic competence of LMs. In Figure 11 (left), we compare encoder and decoder LMs. Due to the absence of big encoder LMs, we consider five *encoder* and six *decoder* LMs with up to 220m parameters. Encoder LMs show a higher mwr of 52% than decoder LMs (21%). This observation is the most saturated for *morphology* or *syntax*, encompassing a variety of token-level phenomena, like part-of-speech. We assume that the missing bi-directional encoding of decoder LMs causes this lower performance because the available context of one token heavily depends on its position. Thus, even common tokens, like *the*, have different potential representations - at the beginning or in the middle of a sentence. These instabilities are further evident when considering Figure 11 (right) which reports the accuracy for the top-20 most common POS tokens (such as *the*) based on the *pos*, *xpos*, *upos* dataset. Given their high frequency, one expects stable prediction performance. Surprisingly, encoder LMs (BERT and RoBERTa) show higher median accuracy and clearly lower deviations compared to the same-size decoder counterpart (GPT2). While scaling model size to 12B (Pythia) and 70B (Llama-2) allows for improved accuracy and lower deviations, decoder LMs do not match the encoder performance, even up to **700 times bigger**.

v) The effect of scaling parameters. We discuss how the number of parameters influences the linguistic competence of LMs. Given the variety of LMs of different sizes, we focus on the Pythia (decoder-only) and T5 (encoder-decoder) families. From Figure 10, we observe for both Pythia and T5 that the linguistic competence scales with model size, and it is particularly pronounced after exceeding 0.5B (Pythia) and 1.0B (T5) parameters. Again, model architecture is crucial, as T5 LMs (encoder-decoder) exhibit a clearly higher mean winning rate of $40 - 70\%$ than Pythia (decoder-only) ones with mwr of $20 - 60\%$. Further, we found formal phenomena evolving differently with increased model size than functional ones. Specifically, *morphology* and *syntax* start at a lower level, with an apparent performance jump after 0.5B (Pythia) and 1.0B (T5) parameters, followed by slow but steady growth. Differently, *semantics*, *reasoning*, and *discourse* start at a higher mwr, followed by a continuous improvement as the model size grows. From these results, we assume more parameters allow LMs to better approximate simpler co-occurrences in the near neighborhood of words to understand formal phenomena like word dependencies. In contrast, more parameters do not have the same pronounced effect on functional phenomena, like rhetorical relations, which require an LM to acquire more distant and complex word co-occurrences.

Model	Morphology	Syntax	Semantics	Reasoning	Discourse	Overall
Comparison against Llama-2 with 7 billion parameters						
Llama-2-Chat	-8%	+3%	-5%	-9%	-3%	-2%
Comparison against T5 with 11 billion parameters						
FLAN-T5	+10%	+2%	-2%	+6%	-2%	+1%
Comparison against Pythia with 12 billion parameters						
Dolly-v2	+4%	-3%	-9%	-3%	+4%	-4%
Comparison against Llama-2 with 13 billion parameters						
Tülu-2	+5%	+2%	-15%	0%	-30%	-8%
Orca-2	-1%	-3%	-4%	+4%	-5%	-2%
Llama-2-chat	+3%	+1%	-6%	+3%	-1%	-1%
Vicuna-v1.5	+23%	+7%	-3%	+6%	-6%	+4%
Comparison against UL2 with 20 billion parameters						
FLAN-UL2	**+40%**	**+16%**	**+7%**	**+13%**	+1%	**+13%**
Comparison against Mixtral with ~47 billion parameters						
Mixtral-Instruct	+4%	+3%	0%	+6%	-2%	+2%
Comparison against Llama-2 with 70 billion parameters						
Tülu-2	+15%	0%	-11%	-3%	0%	-2%
Llama-2-Chat	+23%	+14%	+2%	+4%	**+17%**	+10%
Average	*+10%*	*+4%*	*-3%*	*+4%*	*-2%*	*+1%*

Table 1: Effect of instruction tuning on the mean winning rate compared to the pre-trained LMs.

vi) The effect of instruction tuning. Finally, we focus on how instruction tuning affects LMs' linguistic competence and compare the tuned LMs with their base models—for example, FLAN-UL2 vs. UL2. From results in Table 1, we note less saturated effects for the overall scope while being more pronounced for the five phenomenon types - again emphasizing the structured and comprehensive evaluation of linguistic competence. On average, we found instruction tuning has the highest

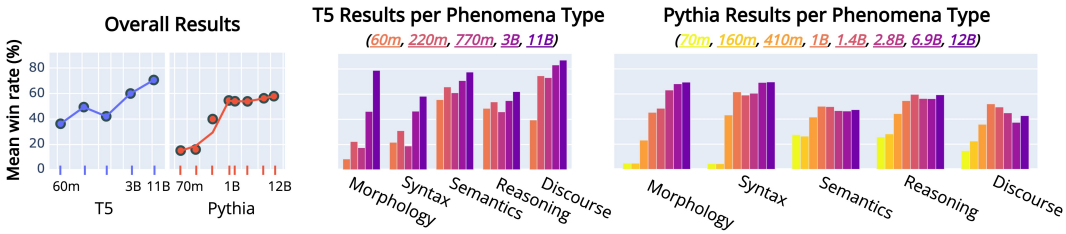

Figure 10: Effect of scaling LM parameters considering the T5 and Pythia model families providing eight and five different sizes. We address the overall scope (left) and the different types of linguistic phenomena (right).

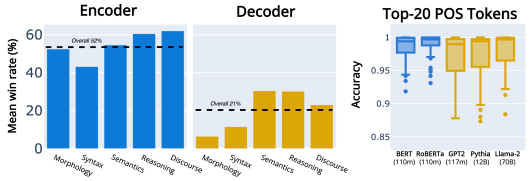

Figure 11: Comparison of the phenomenon types for encoder and decoder LMs (left) and on the right, the accuracy of the top-20 most common tokens of the three part-of-speech probing datasets for BERT, RoBERTa, GPT2, Pythia, and Llama-2.

effect on *morphology* (+10%) followed by *syntax* (+4%), *reasoning* (+4%), and a negative effect for *semantics* −3% and *discourse* −2%. These results confirm previous assumptions that instruction tuning updates are often superficial (Yadav et al., 2023; Hershcovitch et al., 2024; Sharma et al., 2023) and that LMs are better at mimicking language (formal phenomena) than understanding it, measured with functional phenomena (Mahowald et al., 2024). Further, larger models benefit more from instruction tuning. Llama-2-70b-Chat and FLAN-UL2 gain up to +23% and +40% for *morphology* and +10% and +13% on average. In addition, decoder-only LMs (Llama-2 and Pythia) tend to show less pronounced positive effects than encoder-decoder LMs (FLAN-T5-XXL and FLAN-UL2). However, they better understand *reasoning* phenomena. When comparing LMs based on Llama-2-13b, we see that specific fine-tuning methods shape the LMs differently. The top-ranked 13b LM for `Holmes` and OpenLLM, Vicuna, was trained on 125k instructions, less than other models. Thus, high quality is more important than the number of instructions for LMs' linguistic competence. Tülu loses performance while being trained on a large mixture of data (approx. 330k instructions), the same for its 70b version. Finally, the focus of Orca-2 on reasoning is also reflected in its embedding space. These insights show again that while providing a particular perspective, `Holmes` shows clear differences between LMs and allows us to map them to methodological decisions.

6 Efficiency

Seamless, easy, cost-effective integration of new LMs is crucial for widely adopting a benchmark. As `Holmes` covers many datasets and examples, it is computationally heavy in encoding text and training the probes. It takes approx. 6 GPU days to encode the 70 million tokens (∼230k pages of text) and 2 days to run the 208 probes for a 70b model. To account for this issue, we introduce `FlashHolmes`, a streamlined version of `Holmes`. It allows the evaluation of new LMs with a fraction of the compute while maintaining evaluation integrity.

Besides excluding licensed data (18 probing datasets), we analyze the effect of discarding training instances. As a result, we reduce the computation for encoding and the actual probing simultaneously. We follow Perlitz et al. (2023) and calculate the *rank resolution*, 95% CI of model rank difference. This measure indicates the maximum expected rank deviation from evaluating an LM on `FlashHolmes` compared to `Holmes`. For example, a rank resolution of one means that an LM evaluated on `FlashHolmes` and `Holmes` has the same rank or switch place with its neighbors with a probability of 95%. Figure 12 shows the resulting rank resolution when training only on a fraction of the instances, from $1/2$ to $1/512$. Solely focusing on efficiency ($1/512$) still provides a decent rank resolution of ~2.7. In contrast, considering $1/2$ of the training data results in the best reliability of ~1.0. To balance benchmark reliability and efficiency, we compose `FlashHolmes` using $1/32$ of the training instances. Precisely, it reduces the computation expenses of evaluating LMs to ~3% of what `Holmes` would have required while pre-

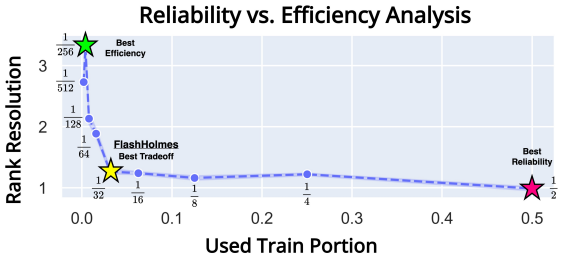

Figure 12: Analysis of the reliability vs. efficiency trade-off when reducing the number of training data.

serving a high rank-correlation of ~1.3.

7 Related Work

Benchmarking LMs Benchmarks approximate LMs abilities like general language understanding (Wang et al., 2019b,a), out-of-distribution generalization (Yang et al., 2023; Waldis et al., 2024b), adversarial scenarios (Nie et al., 2020; Wang et al., 2021), or retrieval like *BEIR* (Thakur et al., 2021) or *MTEB* (Muennighoff et al., 2023). With the advent of larger LMs, the methodological focus shifted to prompting-based evaluations which evaluate the LMs' response to provided instructions (Brown et al., 2020; Hendrycks et al., 2021; Srivastava et al., 2022) covering application-oriented tasks (Liang et al., 2023), or mathematical reasoning (e.g., *GSM8K* (Cobbe et al., 2021)).

Assessing the Linguistic Competence of LMs The analysis of LMs' linguistic competence ranges from analyzing static word vectors (Köhn, 2015), sentence embeddings (Conneau et al., 2018; Adi et al., 2017), the internals of translation models (Shi et al., 2016; Bau et al., 2019), or contextualized LMs (Tenney et al., 2019b,a; Hewitt and Manning, 2019). Other work addressed methodological aspects, such as using control tasks (Hewitt and Liang, 2019), assessing LMs from an information theory perspective (Voita and Titov, 2020; Pimentel et al., 2020), or evaluating causal effects in LMs (Elazar et al., 2021). Finally, another line of work focuses on whether LMs follow human understanding of linguistic competence when solving downstream tasks (Belinkov, 2022; Aw et al., 2023; Mahowald et al., 2024). However, Mosbach et al. (2020b) and Waldis et al. (2024a) found fine-tuning for downstream tasks actually hurting the understanding of linguistic phenomena.

While prior studies assessing the linguistic competence of LMs tend to focus on a limited set of linguistic phenomena or models, `Holmes` provides extensive coverage of both phenomena and evaluated LMs. Unlike recent evaluations based on prompting methods (Blevins et al., 2023; Liang et al., 2023; Amouyal et al., 2024), `Holmes` assesses the internal representations of LMs directly. This approach allows for detailed analysis of specific model characteristics, such as architecture, and helps separate the linguistic competence from other cognitive abilities. Thereby, we respond to recent calls for a thorough and explicit evaluation of linguistic phenomena (Hu and Levy, 2023; Lu et al., 2023; Mahowald et al., 2024).

8 Conclusion

`Holmes` marks the most up-to-date and extensive consolidation of existing resources addressing the need to assess the linguistic competence of LMs in isolation. Our experiments demonstrate that LMs' linguistic competence is pronounced regarding formal phenomena but lacks functional ones when information about broader textual contexts, such as rhetorical structure, is required. Further, size, architecture, and instruction tuning crucially account for differences among LMs. As LM and resources in the landscape of linguistics continue to grow, we will actively extend `Holmes` with further probing datasets, evaluate upcoming LMs, and plan to incorporate multilingualism.

Ethical Considerations and Limitations

Language `Holmes` as well as `FlashHolmes` solely assess linguistic phenomena for the English language. As we plan to expand the benchmark and scope of multilingual data, we focus momentarily on English because of the widespread availability of resources, including curated corpora and the diversity of available LMs.

Phenomena and LM Coverage We agree with Liang et al. (2023) and see one fundamental aspect in composing a benchmark in acknowledging its incompleteness. Both linguistic phenomena and LMs are a subset of the variety of available resources. We consolidated them carefully to provide a comprehensive scope of the linguistic competence and various LMs. However, as benchmarks evolve as tools to assess LMs, we will further expand `Holmes` both with the existing and upcoming LMs and data resources.

Data Availability Linguistic annotations, in particular more complex ones targeting phenomena

like *discourse*, are money and time-wise expensive. Out of 208 datasets included in `Holmes`, 18 probing datasets are based on licensed resources and are not freely available. However, with `FlashHolmes`, we provide an effective and efficient alternative based on open-access resources. Furthermore, upon confirming the granted access, we are happy to share our probing datasets, including those based on the licensed resources.

Bias As `Holmes` relies on existing resources, it inherits the bias embodied in this data. Examples of such bias are gender equality or gender fairness, like the use of neo pronouns such as *em* in Lauscher et al. (2023).

References

Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2017. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Ehsan Aghazadeh, Mohsen Fayyaz, and Yadollah Yaghoobzadeh. 2022. Metaphors in pre-trained language models: Probing and generalization across datasets and languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2037–2050, Dublin, Ireland. Association for Computational Linguistics.

Samuel Amouyal, Aya Meltzer-Asscher, and Jonathan Berant. 2024. Large language models for psycholinguistic plausibility pretesting. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 166–181, St. Julian's, Malta. Association for Computational Linguistics.

Khai Loong Aw, Syrielle Montariol, Badr AlKhamissi, Martin Schrimpf, and Antoine Bosselut. 2023. Instruction-tuning aligns llms to the human brain. *CoRR*, abs/2312.00575.

Simone Balloccu, Patrícia Schmidtová, Mateusz Lango, and Ondrej Dusek. 2024. Leak, cheat, repeat: Data contamination and evaluation malpractices in closed-source LLMs. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 67–93, St. Julian's, Malta. Association for Computational Linguistics.

Anthony Bau, Yonatan Belinkov, Hassan Sajjad, Nadir Durrani, Fahim Dalvi, and James R. Glass. 2019. Identifying and controlling important neurons in neural machine translation. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Edward Beeching, Clémentine Fourrier, Nathan Habib, Sheon Han, Nathan Lambert, Nazneen Rajani, Omar Sanseviero, Lewis Tunstall, and Thomas Wolf. 2023. Open llm leaderboard. `https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard`.

Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219.

Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. 2017. What do neural machine translation models learn about morphology? In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 861–872, Vancouver, Canada. Association for Computational Linguistics.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 2397–2430. PMLR.

Julia Birke and Anoop Sarkar. 2006. A clustering approach for nearly unsupervised recognition of nonliteral language. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 329–336, Trento, Italy. Association for Computational Linguistics.

Terra Blevins, Hila Gonen, and Luke Zettlemoyer. 2023. Prompting language models for linguistic structure. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6649–6663, Toronto, Canada. Association for Computational Linguistics.

Nikolay Bogoychev and Adam Lopez. 2016. N-gram language models for massively parallel devices. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1944–1953, Berlin, Germany. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33:*

Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurovsky. 2001. Building a discourse-tagged corpus in the framework of rhetorical structure theory. In *Proceedings of the SIGDIAL 2001 Workshop, The 2nd Annual Meeting of the Special Interest Group on Discourse and Dialogue, Saturday, September 1, 2001 to Sunday, September 2, 2001, Aalborg, Denmark*. The Association for Computer Linguistics.

Noam Chomsky. 1965. *Aspects of the Theory of Syntax*. The MIT Press, Cambridge.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Y. Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *CoRR*, abs/2210.11416.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: pre-training text encoders as discriminators rather than generators. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *CoRR*, abs/2110.14168.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia. Association for Computational Linguistics.

Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. 2023. Free dolly: Introducing the world's first truly open instruction-tuned llm.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, K. Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. 2021. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175.

Rudolf Franz Flesch. 1948. A new readability yardstick. *The Journal of applied psychology*, 32(3):221–233.

William Gantt, Lelia Glass, and Aaron Steven White. 2022. Decomposing and recomposing event structure. *Transactions of the Association for Computational Linguistics*, 10:17–34.

Vagrant Gautam, Eileen Bingert, D. Zhu, Anne Lauscher, and Dietrich Klakow. 2024. Robust pronoun use fidelity with english llms: Are they reasoning, repeating, or just biased? *CoRR*, abs/2404.03134.

Mario Giulianelli, 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, Brussels, Belgium. Association for Computational Linguistics.

Venkata Govindarajan, Benjamin Van Durme, and Aaron Steven White. 2019. Decomposing generalization: Models of generic, habitual, and episodic statements. *Transactions of the Association for Computational Linguistics*, 7:501–517.

Zellig S Harris. 1954. Distributional structure. *Word*, 10(2-3):146–162.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: decoding-enhanced bert with disentangled attention. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. 2010. SemEval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the 5th International*

Workshop on Semantic Evaluation, pages 33–38, Uppsala, Sweden. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Moshik Hershcovitch, Leshem Choshen, Andrew Wood, Ilias Enmouri, Peter Chin, Swaminathan Sundararaman, and Danny Harnik. 2024. Lossless and near-lossless compression for foundation models. *CoRR*, abs/2404.15198.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. 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 2733–2743, Hong Kong, China. Association for Computational Linguistics.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. 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 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Yufang Hou. 2018. Enhanced word representations for bridging anaphora resolution. 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 1–7, New Orleans, Louisiana. Association for Computational Linguistics.

Yufang Hou. 2020. Bridging anaphora resolution as question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1428–1438, Online. Association for Computational Linguistics.

Jennifer Hu and Roger Levy. 2023. Prompting is not a substitute for probability measurements in large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5040–5060, Singapore. Association for Computational Linguistics.

Philip A. Huebner, Elior Sulem, Fisher Cynthia, and Dan Roth. 2021. BabyBERTa: Learning more grammar with small-scale child-directed language. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 624–646, Online. Association for Computational Linguistics.

Alexander Immer, Lucas Torroba Hennigen, Vincent Fortuin, and Ryan Cotterell. 2022. Probing as quantifying inductive bias. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1839–1851, Dublin, Ireland. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *CoRR*, abs/2310.06825.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts. *CoRR*, abs/2401.04088.

Josef Klafka and Allyson Ettinger. 2020. Spying on your neighbors: Fine-grained probing of contextual embeddings for information about surrounding words. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4801–4811, Online. Association for Computational Linguistics.

Arne Köhn. 2015. What's in an embedding? analyzing word embeddings through multilingual evaluation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2067–2073, Lisbon, Portugal. Association for Computational Linguistics.

Natalia Konstantinova, Sheila C.M. de Sousa, Noa P. Cruz, Manuel J. Maña, Maite Taboada, and Ruslan Mitkov. 2012. A review corpus annotated for negation, speculation and their scope. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3190–3195, Istanbul, Turkey. European Language Resources Association (ELRA).

Fajri Koto, Jey Han Lau, and Timothy Baldwin. 2021. Discourse probing of pretrained language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3849–3864, Online. Association for Computational Linguistics.

Katarzyna Krasnowska-Kieraś and Alina Wróblewska. 2019. Empirical linguistic study of sentence embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5729–5739, Florence, Italy. Association for Computational Linguistics.

Murathan Kurfalı and Robert Östling. 2021. Probing multilingual language models for discourse. In

Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021), pages 8–19, Online. Association for Computational Linguistics.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Anne Lauscher, Debora Nozza, Ehm Miltersen, Archie Crowley, and Dirk Hovy. 2023. What about "em"? how commercial machine translation fails to handle (neo-)pronouns. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 377–392, Toronto, Canada. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, 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*. Featured Certification, Expert Certification.

Tal Linzen, 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.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Sheng Lu, Irina Bigoulaeva, Rachneet Sachdeva, Harish Tayyar Madabushi, and Iryna Gurevych. 2023. Are emergent abilities in large language models just in-context learning? *CoRR*, abs/2309.01809.

Kyle Mahowald, Anna A. Ivanova, Idan A. Blank, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. 2024. Dissociating language and thought in large language models. *Trends in Cognitive Sciences*.

George A. Miller. 1995. Wordnet: A lexical database for english. *Communications of the ACM*, 38(11):39–41.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Arindam Mitra, Luciano Del Corro, Shweti Mahajan, Andrés Codas, Clarisse Simões, Sahaj Agrawal, Xuxi Chen, Anastasia Razdaibiedina, Erik Jones, Kriti Aggarwal, Hamid Palangi, Guoqing Zheng, Corby Rosset, Hamed Khanpour, and Ahmed Awadallah. 2023. Orca 2: Teaching small language models how to reason. *CoRR*, abs/2311.11045.

Moran Mizrahi, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. 2024. State of what art? A call for multi-prompt LLM evaluation. *CoRR*, abs/2401.00595.

Michael Mohler, Mary Brunson, Bryan Rink, and Marc Tomlinson. 2016. Introducing the LCC metaphor datasets. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4221–4227, Portorož, Slovenia. European Language Resources Association (ELRA).

Roser Morante and Eduardo Blanco. 2012. *SEM 2012 shared task: Resolving the scope and focus of negation. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 265–274, Montréal, Canada. Association for Computational Linguistics.

Marius Mosbach, Anna Khokhlova, Michael A. Hedderich, and Dietrich Klakow. 2020a. On the interplay between fine-tuning and sentence-level probing for linguistic knowledge in pre-trained transformers. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 68–82, Online. Association for Computational Linguistics.

Marius Mosbach, Anna Khokhlova, Michael A. Hedderich, and Dietrich Klakow. 2020b. On the Interplay Between Fine-tuning and Sentence-level Probing for Linguistic Knowledge in Pre-trained Transformers. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2502–2516, Online. Association for Computational Linguistics.

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. MTEB: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037, Dubrovnik, Croatia. Association for Computational Linguistics.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

Allen Nie, Erin Bennett, and Noah Goodman. 2019. DisSent: Learning sentence representations from explicit discourse relations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4497–4510, Florence, Italy. Association for Computational Linguistics.

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

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. 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 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Gustavo Paetzold and Lucia Specia. 2016. SemEval 2016 task 11: Complex word identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 560–569, San Diego, California. Association for Computational Linguistics.

Onkar Pandit and Yufang Hou. 2021. Probing for bridging inference in transformer language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4153–4163, Online. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Yotam Perlitz, Elron Bandel, Ariel Gera, Ofir Arviv, Liat Ein-Dor, Eyal Shnarch, Noam Slonim, Michal Shmueli-Scheuer, and Leshem Choshen. 2023. Efficient benchmarking (of language models). *CoRR*, abs/2308.11696.

Fabio Petroni, Patrick S. H. Lewis, Aleksandra Piktus, Tim Rocktäschel, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. 2020. How context affects language models' factual predictions. In *Conference on Automated Knowledge Base Construction, AKBC 2020, Virtual, June 22-24, 2020*.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick S. H. Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander H. Miller. 2019. Language models as knowledge bases? 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 2019, Hong Kong, China, November 3-7, 2019*, pages 2463–2473. Association for Computational Linguistics.

Tiago Pimentel, Josef Valvoda, Rowan Hall Maudslay, Ran Zmigrod, Adina Williams, and Ryan Cotterell. 2020. Information-theoretic probing for linguistic structure. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4609–4622, Online. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Pedro Rodriguez, Joe Barrow, Alexander Miserlis Hoyle, John P. Lalor, Robin Jia, and Jordan Boyd-Graber. 2021. Evaluation examples are not equally informative: How should that change NLP leaderboards? 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 4486–4503, Online. Association for Computational Linguistics.

Rachel Rudinger, Adam Teichert, Ryan Culkin, Sheng Zhang, and Benjamin Van Durme. 2018a. Neural-Davidsonian semantic proto-role labeling. In *Pro-*

ceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 944–955, Brussels, Belgium. Association for Computational Linguistics.

Rachel Rudinger, Aaron Steven White, and Benjamin Van Durme. 2018b. Neural models of factuality. 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 731–744, New Orleans, Louisiana. Association for Computational Linguistics.

Naomi Shapiro, Amandalynne Paullada, and Shane Steinert-Threlkeld. 2021. A multilabel approach to morphosyntactic probing. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4486–4524, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Pratyusha Sharma, Jordan T. Ash, and Dipendra Misra. 2023. The truth is in there: Improving reasoning in language models with layer-selective rank reduction. *CoRR*, abs/2312.13558.

Xing Shi, Inkit Padhi, and Kevin Knight. 2016. Does string-based neural MT learn source syntax? In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1526–1534, Austin, Texas. Association for Computational Linguistics.

Natalia Silveira, Timothy Dozat, Marie-Catherine de Marneffe, Samuel Bowman, Miriam Connor, John Bauer, and Chris Manning. 2014. A gold standard dependency corpus for English. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2897–2904, Reykjavik, Iceland. European Language Resources Association (ELRA).

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ameet Rahane, Anantharaman S. Iyer, Anders Andreassen, Andrea Santilli, Andreas Stuhlmüller, Andrew M. Dai, Andrew La, Andrew K. Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakas, and et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *CoRR*, abs/2206.04615.

Gerard J Steen, Aletta G Dorst, J Berenike Herrmann, Anna A Kaal, Tina Krennmayr, Tryntje Pasma, et al. 2010. *A method for linguistic metaphor identification*. Converging evidence in language and communication research. John Benjamins Publishing Company Amsterdam.

Shivchander Sudalairaj, Abhishek Bhandwaldar, Aldo Pareja, Kai Xu, David D. Cox, and Akash Srivastava. 2024. LAB: large-scale alignment for chatbots. *CoRR*, abs/2403.01081.

György Szarvas, Veronika Vincze, Richárd Farkas, and János Csirik. 2008. The BioScope corpus: annotation for negation, uncertainty and their scope in biomedical texts. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, pages 38–45, Columbus, Ohio. Association for Computational Linguistics.

Alon Talmor, Yanai Elazar, Yoav Goldberg, and Jonathan Berant. 2020. oLMpics-On What Language Model Pre-training Captures. *Transactions of the Association for Computational Linguistics*, 8:743–758.

Yi Tay, Mostafa Dehghani, Vinh Q. Tran, Xavier Garcia, Jason Wei, Xuezhi Wang, Hyung Won Chung, Dara Bahri, Tal Schuster, Huaixiu Steven Zheng, Denny Zhou, Neil Houlsby, and Donald Metzler. 2023. UL2: unifying language learning paradigms. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019a. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R. Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R. Bowman, Dipanjan Das, and Ellie Pavlick. 2019b. What do you learn from context? probing for sentence structure in contextualized word representations. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*.

Lucas Torroba Hennigen, Adina Williams, and Ryan Cotterell. 2020. Intrinsic probing through dimension selection. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 197–216, Online. Association for Computational Linguistics.

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

Teemu Vahtola, Mathias Creutz, and Jörg Tiedemann. 2022. It is not easy to detect paraphrases: Analysing semantic similarity with antonyms and negation using the new SemAntoNeg benchmark. In *Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 249–262, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Siddharth Vashishtha, Benjamin Van Durme, and Aaron Steven White. 2019. Fine-grained temporal relation extraction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2906–2919, Florence, Italy. Association for Computational Linguistics.

Sara Veldhoen, Dieuwke Hupkes, and Willem H. Zuidema. 2016. Diagnostic classifiers revealing how neural networks process hierarchical structure. In *Proceedings of the Workshop on Cognitive Computation: Integrating neural and symbolic approaches 2016 co-located with the 30th Annual Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain, December 9, 2016*, volume 1773 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Elena Voita and Ivan Titov. 2020. Information-theoretic probing with minimum description length. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 183–196, Online. Association for Computational Linguistics.

Andreas Waldis, Yufang Hou, and Iryna Gurevych. 2024a. Dive into the chasm: Probing the gap between in- and cross-topic generalization. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 2197–2214, St. Julian's, Malta. Association for Computational Linguistics.

Andreas Waldis, Yufang Hou, and Iryna Gurevych. 2024b. How to handle different types of out-of-distribution scenarios in computational argumentation? a comprehensive and fine-grained field study. *CoRR*, abs/2309.08316.

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

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019b. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Boxin Wang, Chejian Xu, Shuohang Wang, Zhe Gan, Yu Cheng, Jianfeng Gao, Ahmed Hassan Awadallah, and Bo Li. 2021. Adversarial GLUE: A multi-task benchmark for robustness evaluation of language models. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*.

Yizhong Wang, Hamish Ivison, Pradeep Dasigi, Jack Hessel, Tushar Khot, Khyathi Chandu, David Wadden, Kelsey MacMillan, Noah A. Smith, Iz Beltagy, and Hannaneh Hajishirzi. 2023. How far can camels go? exploring the state of instruction tuning on open resources. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Kuntal Kumar Pal, Maitreya Patel, Mehrad Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Savan Doshi, Shailaja Keyur Sampat, Siddhartha Mishra, Sujan Reddy A, Sumanta Patro, Tanay Dixit, and Xudong Shen. 2022. Super-NaturalInstructions: Generalization via declarative instructions on 1600+ NLP tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages

5085–5109, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Bonnie Webber, Rashmi Prasad, Alan Lee, and Aravind Joshi. 2019. The penn discourse treebank 3.0 annotation manual. *Philadelphia, University of Pennsylvania*.

Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, et al. 2013. Ontonotes release 5.0. *Linguistic Data Consortium, Philadelphia, PA*, 23:170.

Aaron Steven White, Drew Reisinger, Keisuke Sakaguchi, Tim Vieira, Sheng Zhang, Rachel Rudinger, Kyle Rawlins, and Benjamin Van Durme. 2016. Universal decompositional semantics on Universal Dependencies. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1713–1723, Austin, Texas. Association for Computational Linguistics.

Zhiyong Wu, Yun Chen, Ben Kao, and Qun Liu. 2020. Perturbed masking: Parameter-free probing for analyzing and interpreting BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4166–4176, Online. Association for Computational Linguistics.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. 2023. Wizardlm: Empowering large language models to follow complex instructions. *CoRR*, abs/2304.12244.

Prateek Yadav, Leshem Choshen, Colin Raffel, and Mohit Bansal. 2023. Compeft: Compression for communicating parameter efficient updates via sparsification and quantization. *CoRR*, abs/2311.13171.

Linyi Yang, Shuibai Zhang, Libo Qin, Yafu Li, Yidong Wang, Hanmeng Liu, Jindong Wang, Xing Xie, and Yue Zhang. 2023. GLUE-X: Evaluating natural language understanding models from an out-of-distribution generalization perspective. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12731–12750, Toronto, Canada. Association for Computational Linguistics.

Amir Zeldes. 2017. The GUM corpus: Creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3):581–612.

Xikun Zhang, Deepak Ramachandran, Ian Tenney, Yanai Elazar, and Dan Roth. 2020. Do language embeddings capture scales? In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 292–299, Online. Association for Computational Linguistics.

Yian Zhang, 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, Online. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023. LIMA: less is more for alignment. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Zining Zhu, Soroosh Shahtalebi, and Frank Rudzicz. 2022a. Predicting fine-tuning performance with probing. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11534–11547, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Zining Zhu, Jixuan Wang, Bai Li, and Frank Rudzicz. 2022b. On the data requirements of probing. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 4132–4147, Dublin, Ireland. Association for Computational Linguistics.

A Additional Details of Holmes

A.1 Additional Details on the Evolution of Probing Literature

We analyze publication trends by year and venue as shown in Table 2. Less work was published between 2015-2018 (*earlier*) focusing on LSTM-based (Linzen et al., 2016; Conneau et al., 2018) and static LMs (Köhn, 2015; Linzen et al., 2016; Belinkov et al., 2017; Conneau et al., 2018). With the release of BERT (Devlin et al., 2019) in 2019, we note increasing attention to analyzing linguistic abilities within LMs, with a peak of 90 papers in 2022.[5] Considering the venue, more than half of the relevant work (149 papers) was published at major conferences (ACL and EMNLP), and 68 papers were published at AACL, EACL, NAACL, and COLING.[6] In addition, we observe a constant contribution of TACL, various workshops, such as Analyzing and Interpreting Neural Networks for NLP or Representation Learning for NLP.

	earlier	*2019*	*2020*	*2021*	*2022*	*2023*	Total
ACL	2	10	12	9	34	25	92
AACL	-	-	-	-	1	-	1
COLING	-	-	10	-	9	-	19
EACL	-	-	-	7	-	15	22
EMNLP	2	4	13	17	21	-	57
NAACL	-	3	-	9	14	-	26
TACL	1	1	2	3	3	1	11
Workshops	4	4	10	10	7	1	36
Other	1	2	1	1	1	4	10
Probing	10	24	48	56	90	46	274
All Papers	8,056	3,111	3,822	4,294	5,133	3,647	28,063

Table 2: Evolution of probing studies. Note that EMNLP-23 and AACL-23 proceedings were not published when conducting this meta-study.

A.2 Experimental Details

Probing Hyperparameters Following previous work (Hewitt and Liang, 2019; Voita and Titov, 2020), we use fixed hyperparameters for training the probes: 20 epochs, where we find the best one using dev instances; AdamW (Loshchilov and Hutter, 2019) as optimizer; a batch size of 64; a learning rate of 0.0005; a dropout rate of 0.2; a warmup rate of 10% of the steps; random seeds: $[0, 1, 2, 3, 4]$

Hardware We run all of our experiments using 12 Nvidia RTX A6000 GPUs. Every GPU provides 48GB of memory and 10752 CUDA Cores.

Considered LMs Table 8 outlines the details of the LMs we evaluate on Holmes in this work.

A.3 Linguistic Task Categorization

We show in Table 3, Table 4, Table 7, Table 5, and Table 6 which resources Holmes use to cover *morphology*, *syntax*, *semantics*, *reasoning*, and *discourse* phenomena. This includes 33 works providing the data, the specific linguistic phenomena, or both. For example, for *readability* we use the data of Weischedel et al. (2013) and calculated the flesch score (Flesch, 1948).

Morphology First, we feature 19 tasks verifying *morphology* phenomena: *Anaphor agreement*, *determiner noun agreement*, *subject-verb agreement* and *irregular forms* (Warstadt et al., 2020; Huebner et al., 2021).

Syntax The second group of 75 tasks verifies the following *syntax* phenomena: *Part-of-speech tagging* and *constituent labeling* (Weischedel et al., 2013); *dependency labeling* (Silveira et al., 2014); *bigram-shift*, *tree-depth*, *top-constituent-task*, and *sentence-length* (Conneau et al., 2018); *subject-* & *object-number*, and *deoncausative-inchoative alternation* based on Klafka and Ettinger (2020); *binding*, *control/raising*, *negative polarity item licensing*, *island-effects*, *argument-structure*, *ellipsis*, and *filler-gap* (Warstadt et al., 2020; Huebner et al., 2021).

Semantics Third, consider 67 tasks covering *semantics* phenomena: *Named-entity labeling* and *semantic-role labeling* (Weischedel et al., 2013); *subject-* and *object-number*, *tense*, *semantic odd man out*, *word content*, and *coordination inversion* (Conneau et al., 2018); *semantic relation classification* (Hendrickx et al., 2010); *semantic proto-roles* (Rudinger et al., 2018a); *factuality* (Rudinger et al., 2018b); *genericity* (Govindarajan et al., 2019); *event structure* (Gantt et al., 2022); *time* (Vashishtha et al., 2019); *word sense* (White et al., 2016); *sentiment analysis* (Socher et al., 2013); *object-* and *subject-animacy*, *object-* and *subject-gender*, *verb-tense*, and *verb-dynamic* Klafka and Ettinger (2020); *metaphor* (Mohler et al., 2016; Birke and Sarkar, 2006; Steen et al., 2010); *complex word identification* (Paetzold and Specia, 2016); and *passive* (Krasnowska-Kieraś and Wróblewska, 2019). In addition, we derive

[5] Note that EMNLP-23 and AACL-23 proceedings were not published when conducting this meta-analysis.

[6] Note that EMNLP-23 and AACL-23 proceedings were not published when conducting this meta-study.

synonym-/antonym-detection task using WordNet (Miller, 1995) and the texts from OntoNotes v5 Weischedel et al. (2013).

Reasoning Forth, 19 tasks cover *reasoning* phenomena: *Paraphrasticity* with negation and antonyms (Vahtola et al., 2022); *negation detection* (Szarvas et al., 2008; Konstantinova et al., 2012; Morante and Blanco, 2012); *negation-span classification* (Szarvas et al., 2008; Konstantinova et al., 2012); *negation-correspondence* (Szarvas et al., 2008; Konstantinova et al., 2012); *speculation detection*, *speculation-span classification*, and *speculation-correspondence* (Szarvas et al., 2008); and *always-never*, *age comparison*, *objects comparison*, *antonym negation*, *property conjunction*, *taxonomy connection*, *encyclopedic composition*, and *multi-hop composition* (Talmor et al., 2020).

Discourse Finally, `Holmes` embodies 28 task addressing *discourse* phenomena: *Co-reference resolution* Weischedel et al. (2013); *bridging* (Hou, 2018, 2020; Pandit and Hou, 2021); *discourse connective* (Nie et al., 2019); *sentence order* and *next-sentence prediction* (Narayan et al., 2018); *discourse correspondence*, *discourse order*, *discourse relation*, *discourse distance*, *discourse explicit classes*, *discourse implicit classes* (Webber et al., 2019; Kurfalı and Östling, 2021); and *rst-count/-depth/-distance/-relation/-relation-group/-successively/-type* (Carlson et al., 2001; Koto et al., 2021; Kurfalı and Östling, 2021; Zeldes, 2017).

A.4 Details of Probing Dataset Composition

Whenever possible, we rely on established probing datasets and transform instances into a unified format: **1)** an input x which is either one or a pair of span(s) or sentence(s), including the string and an optional starting and ending index in the context c when task type is either a span or span-pair classification; **2)** an optional textual context c to encode x, for example the sentence in which a span occurs; and **3)** a corresponding label y. If given, we use the original train/dev/test splits. However, if this division does not exist, we use a 70/10/20 ratio to form these splits. Furthermore, we adapt the design of some tasks to map to our task format. Exemplary, for the oLMmpics (Talmor et al., 2020) dataset, we transform the mask-filling tasks into a binary classification where the *correct* label corresponds to a sentence with a correctly filled mask and *incorrect* to a sentence where the mask was filled wrongly.

OnToNotes Following Tenney et al. (2019b,a), we use the *OntoNotes* (Weischedel et al., 2013) dataset to derive *part-of-speech tagging*, *constituent labeling*, *named-entity labeling*, *semantic role*, and *co-reference resolution* probing datasets. Further, we consider with *constituent maximum depth* and *constituent node length* further properties of the constituent tree this dataset *OntoNotes*.

Dependency Corpus As in Tenney et al. (2019b,a), we use Universal Dependencies annotations of the English Web Treebank to form a *dependency labeling* datasets.

Context Probes Presented in Klafka and Ettinger (2020), we compose nine datasets to verify information about context words.

BLiMP Dataset Using the data presented in the BLiMP benchmark (Warstadt et al., 2020), we derive 67 probing datasets verifying specific phenomena, like *island effect*, covering *morphology*, *syntax*, and *semantics*. Unlike the original version, we compose a binary classification task for every phenomenon. Precisely, whether to accept or reject a given sentence, where rejecting means that the given linguistic phenomena is violated.

Zorro Dataset As for the BLiMP tasks, we convert the 21 distinct Zorro tasks into a binary classification task on whether a sentence accepts or rejects the given linguistic phenomena is violated.

SemEval-2010 Task 8 For *semantic relation classification* we rely on the dataset of Hendrickx et al. (2010).

Decompositional Semantics Initiative The *Decompositional Semantics Initiative*[7] provides a large number of datasets to verify semantic phenomena. Apart of the common use *semantic proto-roles* (Rudinger et al., 2018a), we use their collection of works to compose probing datasets for *factuality* (Rudinger et al., 2018b), genericity (Govindarajan et al., 2019), event structure (Vashishtha et al., 2019), time (Vashishtha et al., 2019), and word sense (White et al., 2016).

Sentiment Analysis We use the commonly used work of Socher et al. (2013) and form a probing dataset targeting sentiment.

[7] https://decomp.io/

Metaphor As in Aghazadeh et al. (2022), we use the data from Mohler et al. (2016); Birke and Sarkar (2006); Steen et al. (2010) to form three metaphor datasets.

Complex Word Identification We consider word complexity for the first time and use the data presented in Paetzold and Specia (2016). It provides annotations for different complexity levels of words.

Passive We use data from Krasnowska-Kieraś and Wróblewska (2019) to form a probing dataset assessing knowledge about passive language.

Synonym / Antonym Replacement Using the text of the *OntoNotes* (Weischedel et al., 2013) and Wordnet (Miller, 1995), we form a probing dataset to detect synonym and antonym replacement. Specifically, the binary classification task is: given two texts (the original and an updated one), was the updated one changed by replacing a word with its synonym or antonym?

Negation With this work, we verify for the first time *negation* based on human annotated datasets (Vahtola et al., 2022; Szarvas et al., 2008; Konstantinova et al., 2012). Specifically, we form different probing datasets.

- Is a text negated or not?
- Given two text spans, does the negation within the first one correspond to the second one?
- Given a text span, is it the cue or the scope of the negation?

oLMmpics We form probing datasets addressing different lexical reasoning using the data presented in Talmor et al. (2020). As they provide multiple choices, we form *correct* instances by filling the gap with the correct option and *wrong* ones by filling in the other options. Specifically, we form dataset for *always-never*, *age comparison*, *objects comparison*, *antonym-negation*, *multi-hop composition property conjunction*, *taxonomy conjunction*, and *encyclopedic composition*.

Bridging We rely on the data presented in Pandit and Hou (2021) and form two probing datasets. One is to verify whether a text is linguistically applicable, considering bridging (antecedent matches anaphora). And a second one to verify whether an antecedent and anaphora match.

Discourse Connective Using data from Nie et al. (2019), we form a probing dataset to assess whether a given connective marker matches the discourse of the given text.

Sentence Order and Next Sentence Prediction Following Narayan et al. (2018), we form two datasets to verify the order of good or badness of a given sentence and whether two sentences occur after each other.

Discourse Representation Theory We use data from Webber et al. (2019) to compose eight probing datasets addressing *discourse representation theory*:

- Four probing dataset predicting the class of a given span. We distinguish between *implicit*, *explicit*, *implicit-coarse*, and *explicit-coarse*.
- The absolute distance, number of words, between two spans in the text.
- Whether the order of two spans is correct or not.
- Whether two spans have discourse relation or not.
- The specific discourse relation of two spans.

Rhetorical Structure Theory Using annotations from Carlson et al. (2001); Zeldes (2017), we compose 14 probing datasets addressing *rhetorical theory*. Specifically, we compose the following seven types of datasets for both works:

- The rhetorical type of a text span, either nucleus or satellite.
- The number of children of a text span within the rhetorical tree of the text.
- The depth of a text span within the rhetorical tree of the text.
- The number of edges between two text spans within the rhetorical tree.
- The specific rhetorical relation between two text spans like *conclusion*.
- The relation group of a specific rhetorical relation between two text spans like *evaluation* for the relation *conclusion*.
- Whether two text spans occur after each other in the rhetorical tree.

Phenomena	Text	Text-Pair	Span	Span-Pair	Warstadt et al. (2020)	Huebner et al. (2021)
anaphor agreement	3				✓	✓
determiner noun agreement	10				✓	✓
irregular forms	3				✓	✓
subject-verb agreement	10				✓	✓

Table 3: Overview of resources and linguistic phenomena mapping for *morphology*. It shows the number of datasets for the phenomena by dataset type.

Phenomena	Text	Text-Pair	Span	Span-Pair	Weischedel et al. (2013)	Silveira et al. (2014)	Conneau et al. (2018)	Flesch (1948)	Klafka and Ettinger (2020)	Warstadt et al. (2020)	Huebner et al. (2021)
argument-structure	20									✓	✓
bigram-shift	1						✓				
binding	8									✓	✓
case	1										✓
constituent parsing	2			1	✓						
control/raising	5									✓	✓
deoncausative-inchoative alternation	1								✓		
dependency parsing				1		✓					
ellipsis	3									✓	✓
filler-gap	9									✓	✓
island-effects	10									✓	✓
local attractor	1										✓
object-number	2						✓			✓	
part-of-speech			3		✓	✓					
readability	1				✓			✓			
sentence-length	1						✓				
subject-number	2						✓			✓	
top-constituent-task	1						✓				
tree-depth	1						✓				

Table 4: Overview of resources and linguistic phenomena mapping for *syntax*. It shows the number of datasets for the phenomena by dataset type.

Phenomena	Text	Text-Pair	Span	Span-Pair	Vahtola et al. (2022)	Szarvas et al. (2008)	Konstantinova et al. (2012)	Morante and Blanco (2012)	Talmor et al. (2020)
age comparison	1								✓
always-never	1								✓
antonym negation	1								✓
encyclopedic composition	1								✓
multi-hop composition	1								✓
negation	3	1	2	2	✓	✓	✓	✓	
objects comparison	1								✓
property conjunction	1								✓
speculation	1		1	1		✓			
taxonomy connection	1								✓

Table 5: Overview of resources and linguistic phenomena mapping for *reasoning*. It shows the number of datasets for the phenomena by dataset type.

Phenomena	Text	Text-Pair	Span	Span-Pair	Weischedel et al. (2013)	Pandit and Hou (2021)	Nie et al. (2019)	Narayan et al. (2018)	Webber et al. (2019)	Carlson et al. (2001)	Zeldes (2017)
bridging	1			1		✓					
co-reference resolution				1	✓						
discourse connective		1					✓				
discourse representation theory				8					✓		
next-sentence prediction		1						✓			
rethorical structure theory			6	8						✓	✓
sentence order		1						✓			

Table 6: Overview of resources and linguistic phenomena mapping for *discourse*. It shows the number of datasets for the phenomena by dataset type.

Phenomena	Text	Text-Pair	Span	Span-Pair	Weischedel et al. (2013)	Conneau et al. (2018)	Klafka and Ettinger (2020)	Warstadt et al. (2020)	Huebner et al. (2021)	Hendrickx et al. (2010)	Rudinger et al. (2018a)	Rudinger et al. (2018b)	Govindarajan et al. (2019)	Gantt et al. (2022)	Vashishtha et al. (2019)	White et al. (2016)	Socher et al. (2013)	Mohler et al. (2016)	Birke and Sarkar (2006)	Steen et al. (2010)	Paetzold and Specia (2016)	Krasnowska-Kieraś and Wróblewska (2019)	Miller (1995)
complex word identification			1																		✓		
coordination inversion	1					✓																	
event structure			4	2										✓									
factuality				1								✓											
genericity			6										✓										
metaphor			4															✓	✓	✓			
named-entity labeling			1		✓																		
negative polarity item licensing	4							✓	✓														
object-animacy	1						✓																
object-gender	1						✓																
passive	1																					✓	
quantifiers	6								✓														
semantic relation classification		1								✓													
semantic proto-roles				20							✓												
semantic odd man out	1					✓																	
semantic-role labeling				1	✓																		
sentiment analysis	1																✓						
subject-animacy	1						✓																
subject-gender	1						✓																
synonym-/antonym-detection	1																						✓
tense	2					✓	✓																
time			1												✓								
verb-dynamic	1						✓																
word content	1					✓																	
word sense			1													✓							

Table 7: Overview of resources and linguistic phenomena mapping for *semantics*. It shows the number of datasets for the phenomena by dataset type.

Model	Citation	Size	Pre-Training Objective	Pre-Training Data	Huggingface Tag
			Encoder-Only Language Models		
ALBERT	Lan et al. (2020)	10 million	MLM+SOP	16GB	albert-base-v2
BERT	Tenney et al. (2019a)	110 million	MLM+NSP	16GB	bert-base-uncased
DeBERTa	He et al. (2021)	100 million	MLM	80GB	microsoft/deberta-base
DeBERTa-v3	He et al. (2023)	86 million	MLM+DISC	160GB	microsoft/deberta-v3-base
ELECTRA	Clark et al. (2020)	110 million	MLM	16GB	google/electra-base-discriminator
RoBERTa	Liu et al. (2019)	110 million	MLM+DISC	160GB	roberta-base
			Decoder-Only Language Models		
GPT2	Radford et al. (2019)	117 million	LM	40GB	gpt2
Pythia-70m	Biderman et al. (2023)	70 million	LM	300 billion tokens	EleutherAI/pythia-70m
Pythia-160m	Biderman et al. (2023)	160 million	LM	300 billion tokens	EleutherAI/pythia-160m
Pythia-410m	Biderman et al. (2023)	410 million	LM	300 billion tokens	EleutherAI/pythia-410m
Pythia-1b	Biderman et al. (2023)	1 billion	LM	300 billion tokens	EleutherAI/pythia-1B
Pythia-1.4b	Biderman et al. (2023)	1.4 billion	LM	300 billion tokens	EleutherAI/pythia-1.4B
Pythia-2.8b	Biderman et al. (2023)	2.8 billion	LM	300 billion tokens	EleutherAI/pythia-2.8B
Pythia-6.9b	Biderman et al. (2023)	6.9 billion	LM	300 billion tokens	EleutherAI/pythia-6.9B
Pythia-12b	Biderman et al. (2023)	12 billion	LM	300 billion tokens	EleutherAI/pythia-12B
Pythia-70m-dedup	Biderman et al. (2023)	70 million	LM	207 billion tokens	EleutherAI/pythia-70m-deduped
Pythia-160m-dedup	Biderman et al. (2023)	160 million	LM	207 billion tokens	EleutherAI/pythia-160m-deduped
Pythia-410m-dedup	Biderman et al. (2023)	410 million	LM	207 billion tokens	EleutherAI/pythia-410m-deduped
Pythia-1b-dedup	Biderman et al. (2023)	1 billion	LM	207 billion tokens	EleutherAI/pythia-1B-deduped
Pythia-1.4b-dedup	Biderman et al. (2023)	1.4 billion	LM	207 billion tokens	EleutherAI/pythia-1.4B-deduped
Pythia-2.8b-dedup	Biderman et al. (2023)	2.8 billion	LM	207 billion tokens	EleutherAI/pythia-2.8B-deduped
Pythia-6.9b-dedup	Biderman et al. (2023)	6.9 billion	LM	207 billion tokens	EleutherAI/pythia-6.9B-deduped
Pythia-12b-dedup	Biderman et al. (2023)	12 billion	LM	207 billion tokens	EleutherAI/pythia-12B-deduped
Dolly-v2	Conover et al. (2023)	12 billion	LM+IT	300 billion token + 15K instructions	databricks/dolly-v2-12b
Llama-2-7b	Touvron et al. (2023)	7 billion	LM	2.4 trillion tokens	meta-llama/Llama-2-7b-hf
Llama-2-13b	Touvron et al. (2023)	13 billion	LM	2.4 trillion tokens	meta-llama/Llama-2-13b-hf
Llama-2-70b	Touvron et al. (2023)	70 billion	LM	2.4 trillion tokens	meta-llama/Llama-2-70b-hf
Llama-2-7b-chat	Touvron et al. (2023)	7 billion	LM+IT	2.4 trillion tokens + 27,5K instructions	meta-llama/Llama-2-7b-chat-hf
Llama-2-13b-chat	Touvron et al. (2023)	13 billion	LM+IT	2.4 trillion tokens + 27,5K instructions	meta-llama/Llama-2-13b-chat-hf
Llama-2-70b-chat	Touvron et al. (2023)	70 billion	LM+IT	2.4 trillion tokens + 27,5K instructions	meta-llama/Llama-2-70b-chat-hf
IBM-Merlinite	Sudalairaj et al. (2024)	7 billion	LM+IT	2.4 trillion tokens + 1400k instructions	ibm/merlinite-7b
IBM-Labradorite	Sudalairaj et al. (2024)	13 billion	LM+IT	2.4 trillion tokens + 1400k instructions	ibm/labradorite-13b
Vicuna-13b-v1.5	Zheng et al. (2023)	13 billion	LM+IT	2.4 trillion tokens + 125k instructions	lmsys/vicuna-13b-v1.5
Orca-2-13b	Mitra et al. (2023)	13 billion	LM+IT	2.4 trillion tokens + 817K instructions	microsoft/Orca-2-13b
Wizard-13B-v1.2	Xu et al. (2023)	13 billion	LM	unknown	WizardLM/WizardLM-13B-V1.2
Tülu-2-13b	Wang et al. (2023)	13 billion	LM+IT	2.4 trillion tokens + 330k instructions	allenai/tulu-2-13b
Tülu-2-dpo-13b	Wang et al. (2023)	13 billion	LM+IT	2.4 trillion tokens + 330k instructions	tulu-2-dpo-13b
Tülu-2-70b	Wang et al. (2023)	70 billion	LM+IT	2.4 trillion tokens + 330k instructions	allenai/tulu-2-70b
Tülu-2-dpo-70b	Wang et al. (2023)	70 billion	LM+IT	2.4 trillion tokens + 330k instructions	tulu-2-dpo-70b
Mistral-7b	Jiang et al. (2023)	7 billion	LM	unknown	mistralai/Mistral-7B-v0.1
Mistral-7b-Inst	Jiang et al. (2023)	7 billion	LM	unknown	mistralai/Mistral-7B-Instruct-v0.1
Mixtral-8x7b	Jiang et al. (2024)	47 billion	LM	unknown	mistralai/Mixtral-8x7B-v0.1
Mixtral-8x7b-Inst	Jiang et al. (2024)	47 billion	LM	unknown	mistralai/Mistral-7B-v0.1
			Encoder-Decoder Language Models		
BART	Lewis et al. (2020)	121 million	DAE	160GB	google/facebook/bart-base
T5-small	Raffel et al. (2020)	60 million	DAE	800GB	google/t5-small-lm-adapt
T5-base	Raffel et al. (2020)	220 million	DAE	800GB	google/t5-base-lm-adapt
T5-large	Raffel et al. (2020)	770 million	DAE	800GB	google/t5-large-lm-adapt
T5-xl	Raffel et al. (2020)	3 billion	DAE	800GB	google/t5-xl-lm-adapt
T5-xxl	Raffel et al. (2020)	11 billion	DAE	800GB	google/t5-xxl-lm-adapt
FLAN-T5-small	Raffel et al. (2020)	60 million	DAE+IT	800GB + 1.8k tasks	google/t5-small-lm-adapt
FLAN-T5-base	Raffel et al. (2020)	220 million	DAE+IT	800GB + 1.8k tasks	google/t5-base-lm-adapt
FLAN-T5-large	Raffel et al. (2020)	770 million	DAE+IT	800GB + 1.8k tasks	google/t5-large-lm-adapt
FLAN-T5-xl	Raffel et al. (2020)	3 billion	DAE+IT	800GB + 1.8k tasks	google/t5-xl-lm-adapt
FLAN-T5-xxl	Raffel et al. (2020)	11 billion	DAE+IT	800GB + 1.8k tasks	google/t5-xxl-lm-adapt
TK-Instruct	Wang et al. (2022)	11 billion billion	DAE+IT	800GB + 1.6k tasks	allenai/tk-instruct-11b-def
UL2	Tay et al. (2023)	20 billion	DAE	800GB	google/ul2
FLAN-UL2	Tay et al. (2023)	20 billion	DAE+IT	800GB + 100k instructions	google/flan-ul2
			Static Language Models		
Glove-6B	Pennington et al. (2014)	-	WP	6 billion tokens	glove.6B.300d
Glove-840B	Pennington et al. (2014)	-	WP	840 billion tokens	glove.840B.300d

Table 8: Overview of the evaluated LMS covering the corresponding citation, model size, model architecture, pre-training objective & data, and the Huggingface model tag. Regarding the pre-training objective, we distinguish between masked language modeling (MLM), sentence order prediction (SOP), next sentence prediction (NSP), next word prediction (LM), instruction fine-tuning (IT), word denoising (DAE), and word probabilities from word co-occurrences (WP). For pre-training data, we report known numbers, either as the size of the corpora in gigabytes (GB), the number of pre-training tokens, the number of instructions for fine-tuning, or the number of tasks for instruction fine-tuning.