Elements of World Knowledge (EWoK): A cognition-inspired framework for evaluating basic world knowledge in language models

Anna A. Ivanova[*1] **Aalok Sathe**[*2] **Benjamin Lipkin**[*2]

Unnathi Kumar[1] **Setayesh Radkani**[2] **Thomas H. Clark**[2] **Carina Kauf**[2]

Jennifer Hu[3] **R. T. Pramod**[2] **Gabriel Grand**[2] **Vivian Paulun**[2]

Maria Ryskina[2] **Ekin Akyürek**[2] **Ethan Wilcox**[4] **Nafisa Rashid**[5] **Leshem Choshen**[2,6]

Roger Levy[2] **Evelina Fedorenko**[2] **Joshua Tenenbaum**[2] **Jacob Andreas**[2]

[1]Georgia Tech [2]MIT [3]Harvard University
[4]ETH Zürich [5]Wellesley College [6]MIT-IBM Watson AI

a.ivanova@gatech.edu, {asathe,lipkinb}@mit.edu [*]authors contributed equally

Abstract

The ability to build and leverage world models is essential for a general-purpose AI agent. Testing such capabilities is hard, in part because the building blocks of world models are ill-defined. We present Elements of World Knowledge (EWoK), a framework for evaluating world modeling in language models by testing their ability to use knowledge of a concept to match a target text with a plausible/implausible context. EWoK targets specific concepts from multiple knowledge domains known to be vital for world modeling in humans. Domains range from social interactions (*help/hinder*) to spatial relations (*left/right*). Both, contexts and targets are minimal pairs. Objects, agents, and locations in the items can be flexibly filled in enabling easy generation of multiple controlled datasets. We then introduce EWoK-CORE-1.0, a dataset of 4,374 items covering 11 world knowledge domains. We evaluate 20 open-weights large language models (1.3B–70B parameters) across a battery of evaluation paradigms along with a human norming study comprising 12,480 measurements. The overall performance of all tested models is worse than human performance, with results varying drastically across domains. These data highlight simple cases where even large models fail and present rich avenues for targeted research on LLM world modeling capabilities.

1 Introduction

Large language models (LLMs) acquire a substantial amount of knowledge from their training data, both through direct memorization and through learning co-occurrence-based text patterns (Bender and Koller, 2020; Grand et al., 2022; Pavlick, 2022). This knowledge comprises both *knowledge about language* (e.g. word meanings and rules of syntax) and *knowledge about the world* (e.g. social conventions and physical properties of objects). Contemporary LLMs' knowledge of language is very robust (Mahowald et al., 2024), as evidenced by the fluency of the text that they generate. But how robust is their understanding of the basic social, physical, and relational concepts that are foundational to our everyday experience?

In this paper, we present Elements of World Knowledge (EWoK)[1], a flexible framework for evaluating world modeling in LLMs (see Figure 1 for an overview). The EWoK framework consists of: **(a)** several *domains* that constitute the foundation for basic human world knowledge and are processed by dedicated cognitive and neural systems; **(b)** specific *concepts* within each domain; **(c)** a set of *item templates* (and modular components to procedurally generate these templates) that test knowledge of these

[1]Data and associated code are available at: http://ewok-core.github.io

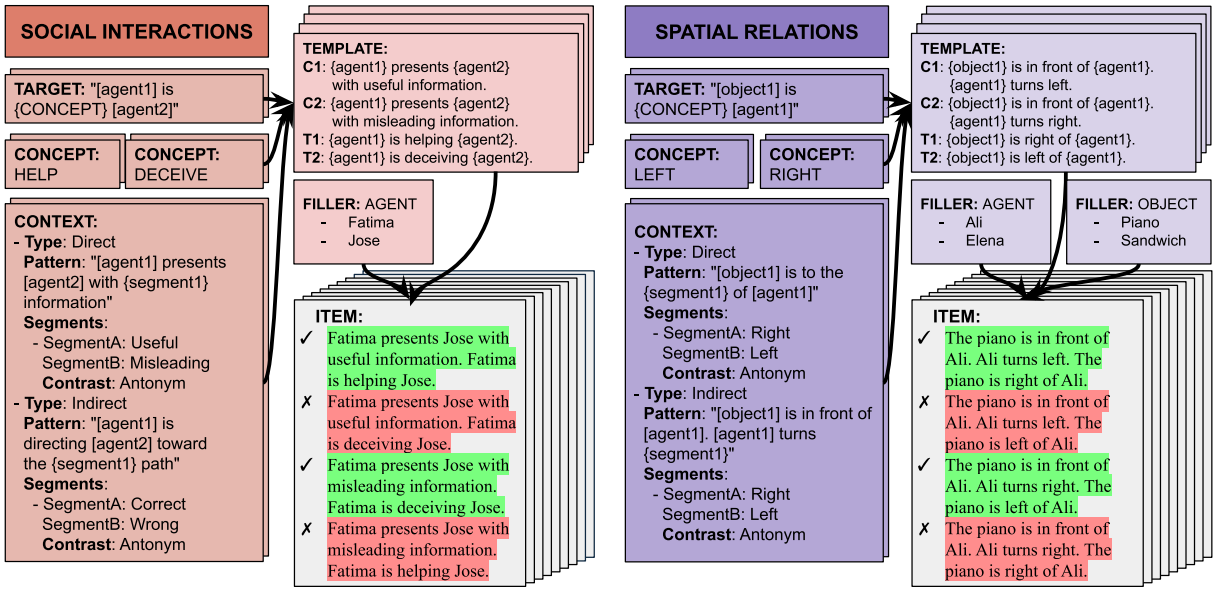

Figure 1: Dataset design. Here we present examples from *social interactions* & *spatial relations*, which LLMs do best and worst at, respectively. All domains can be found in Figure 3. Each *domain* contains a set of *concepts*, *contexts*, and *targets*. These combine to form many *templates*, which specify minimal pairs of contexts (C) and targets (T), such that T_1 matches C_1 but not C_2, and T_2 matches C_2 but not C_1. Each template can be combined with *fillers* to generate an even larger collection of *items*.

concepts by contrasting plausible and implausible context–target combinations; **(d)** a set of *fillers* to populate the templates, such that each template can be used multiple times; **(e)** a pipeline to generate a specific set of *items* (a dataset) based on these source materials.

Why elements? Our framework targets specific cognitive concepts (or concept pairs, such as *friend/enemy*). Concept knowledge is not limited to definitions, but can be used productively across a wide range of scenarios. Thus concepts leveraged in context are the first-class object of the EWoK framework, as opposed to individual sentences or facts. This approach stands in contrast to many NLP benchmarks, which often aim to evaluate knowledge based on individual items, or to develop sophisticated tasks that require drawing on many concepts simultaneously. Although real-world tasks can be more useful for comparing model performance, individual item complexity makes it hard to assess why a model fails. For instance, LLMs exhibit mixed performance on theory of mind tasks, which has led to disagreements about whether LLMs fail at inferring mental states or auxiliary cognitive abilities, such as knowing what it means for a box to be "transparent" (Kosinski, 2023; Ullman, 2023). Our framework mitigates this problem by explicitly linking the items with the concepts that they test.

Why cognition-inspired? World knowledge is a notoriously fuzzy capability; which knowledge domains should one focus on? Here, we select a range of domains that have been shown to recruit dedicated cognitive and/or neural machinery in humans, such as knowledge of intuitive physics (McCloskey, 1983; Battaglia et al., 2013), knowledge of physical and spatial relations (Hafri and Firestone, 2021), intuitive number sense (Dehaene, 2011), social reasoning (Carlson et al., 2013), and reasoning about agents that involves both physical and social knowledge (Liu et al., 2024). These knowledge domains are not specific to language (Jackendoff, 2002); in fact, many are present in preverbal infants (Spelke and Kinzler, 2007). However, language contains a rich amount of information that reflects grounded world knowledge (Roads and Love, 2020; Abdou et al., 2021; Patel and Pavlick, 2021), such that LLMs might in principle acquire such domain-specific knowledge from text alone.

Why plausibility? To evaluate basic world modeling, we use combinations of plausible vs. implausible context–target pairs. Plausibility here serves as a proxy for factual accuracy: instead of deciding whether a factual statement is true or

false, a model needs to determine whether a given scenario makes sense (is plausible). Having an accurate world model is necessary for consistently distinguishing plausible and implausible scenarios no matter how they are worded.

Why minimal pairs (of pairs)? Both contexts and targets in EWoK have a minimal-pairs design, such that a specific targeted change to a sentence (e.g., *left* → *right*) results in an opposite result (plausible/implausible combination). This approach can help identify specific manipulations that LLMs are and are not sensitive to, with the goal of targeted diagnostics. It can also be leveraged in the future for mechanistic interpretability research.

Why context–target combinations? LLMs have a remarkable capacity for memorization, such that many plausible and implausible sentences can be distinguished solely based on their presence in the training data (e.g., *The fox chased the rabbit* is more common than *The rabbit chased the fox*). In contrast, our framework tests LLMs' ability to evaluate contextual plausibility, such that the same exact target (*The piano is left of Ali.*) is either plausible or implausible depending on the context (see Figure 1 right).

Our paper is structured as follows. In Section 2, we review prior research related to this topic. In Section 3, we describe the core components of the EWoK framework. In Sections 4 and 5, we describe our evaluation strategy and the practical steps that we take to reduce the risk of dataset contamination. In Section 6, we show LLM performance on EWoK-CORE-1.0, a dataset generated via the EWoK framework. In Section 7, we discuss our results and future prospects for this line of work.

2 Related Work

The capabilities we evaluate are closely related to the notion of commonsense knowledge, an area that has numerous text-based benchmarks (e.g., Levesque et al., 2012; Sakaguchi et al., 2021; Zellers et al., 2019), including targeted evaluations of physical (Bisk et al., 2020) and social commonsense (Sap et al., 2019). LLMs often struggle on such commonsense benchmarks, likely due to the reporting bias in their training data (Shwartz and Choi, 2020): conversations and texts typically do not include commonly observed

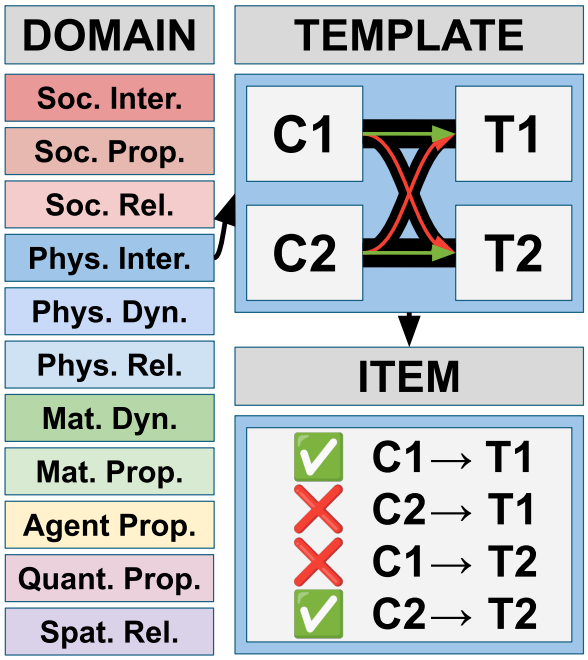

Figure 2: Item Setup. Each item is generated from a template from a domain. Items are designed such that pairs of C_i and T_j are matched only when $i = j$.

or obvious information (Gordon and Van Durme, 2013). Thus, although our items target basic world knowledge and are not designed to be challenging, LLMs might still struggle with them due to the reporting bias.

One specific version of reporting bias affects perceptually grounded knowledge. Co-occurrence information that is easily available through perception (e.g., the fact that bananas are typically yellow or wheels are typically round) is often underrepresented in language corpora. This bias has led an earlier generation of language models to underperform on physically and perceptually grounded world knowledge tasks (Lucy and Gauthier, 2017; Utsumi, 2020) and exhibit representational differences in physical features that are more/less talked about (Abdou et al., 2021; Lewis et al., 2019). That said, in spite of such biases, models trained on text do learn a substantial amount of distributional information from perceptual domains (Roads and Love, 2020; Abdou et al., 2021; Sorscher et al., 2022), meaning that much of the perceptual world knowledge can be acquired without grounding. Overall, we expect LLMs to perform above-chance on world knowledge domains that are perceptually grounded (such as physical relations or material properties), although they might be

DOMAIN	COUNT	SAMPLE TEMPLATE	SAMPLE TEMPLATE TYPE (context type, context contrast, target contrast)
SOCIAL INTERACTIONS	165	C: {agent1} is being [polite/rude] towards {agent2}. T: {agent1} is [respecting/insulting] {agent2}.	Direct, Antonym, Concept Swap
SOCIAL PROPERTIES	185	C: {agent1} [can/cannot] be depended upon. T: {agent1} is [trustworthy/untrustworthy].	Direct, Negation, Concept Swap
SOCIAL RELATIONS	785	C: {agent[1/2]} pays salary to {agent[2/1]}. T: {agent[1/2]} is agent[2/1]}'s boss.	Indirect, Variable Swap, Variable Swap
PHYSICAL INTERACTIONS	280	C: {agent1} [dropped/lifted] {object1}. T: {object1} is moving [toward/away from] the ground.	Indirect, Antonym, Concept Swap
PHYSICAL DYNAMICS	75	C: The speck on {object1:rollable=T} [is/is not] rotating. T: {object1:rollable=T} is [rolling/sliding].	Indirect, Negation, Concept Swap
PHYSICAL RELATIONS	435	C: {object[1/2]} occupies more space than {object[2/1]}. T: {object[1/2]} is bigger than {object[2/1]}.	Direct, Variable Swap, Variable Swap
MATERIAL DYNAMICS	780	C: {agent1}'s item is {item:material=[fabric/liquid]}. T: {agent1} can [fold/pour] it.	Indirect, Other, Concept Swap
MATERIAL PROPERTIES	125	C: {agent1} [can/cannot] see through {object1}. T: {object1} is [transparent/opaque].	Direct, Negation, Concept Swap
AGENT PROPERTIES	1130	C: {agent1} likes {object[1/2]} less than {object[2/1]}. T: {agent1} prefers {object[2/1]} to {object[1/2]}.	Direct, Variable Swap, Variable Swap
QUANTITATIVE PROPERTIES	195	C: {agent1} [needs/does not need] more {object1}. T: {agent1} has [not enough/enough] {object1}.	Indirect, Negation, Concept Swap
SPATIAL RELATIONS	245	C: {object1} is to the [left/right] of {object2}. T: {object2} is to the [right/left] of {object1}.	Direct, Antonym, Concept Swap

Figure 3: Domains. EWoK-CORE-1.0 includes 11 domains, each contributing between 75 and 1130 items to EWoK-CORE-1.0. Here we include a sample template (pair of context–target pairs) for each domain, and note their types. Templates may be *direct*, testing concepts explicitly, or *indirect*, testing concepts implicitly. Context and target *contrasts* reflect how concepts are tested. For example, looking at contexts, note how *antonym* contrasts opposing concepts, *negation* leverages "not", and *variable swap* exploits ordering. Such contrasts are elaborated upon in Section 3.

more challenging than, e.g., social domains.

Our evaluation is also related to works on natural language inference and entailment. The recognizing textual entailment (RTE) task (Dagan et al., 2010) poses two sentences (a text expression T and and hypothesis H) to a system and asks it to determine whether H follows from T. The natural language inference (NLI) task follows a similar challenge (Bowman et al., 2015; Williams et al., 2018; Conneau et al., 2018) and involves making a 3-way judgment about whether a *premise* entails, contradicts, or is neutral relative to a *hypothesis*. EWoK asks whether a target sentence T is plausible given a context C, which might—but does not have to—indicate an entailment relationship between the two. Though widely successful as a challenge, a large body of subsequent work has highlighted issues with RTE- and NLI-style evaluation: language models can often use heuristics (such as artifacts left behind by human annotators and lexical statistics) to "solve" the task without meaningful semantic understanding or reasoning (Poliak et al., 2018; Liu et al., 2020; McCoy et al., 2020; Gururangan et al., 2018). We address this limitation by (1) posing the task as a minimal pair where each of two targets must be preferred over its alternative given the right context, making reliance on target plausibility alone impossible; (2) annotating minimal pair contrast type to test whether which item design features drive model performance; and (3) testing the relationship between LLM performance and surface-level item properties, such as item length, average word frequency, and performance of a baseline bag-of-words embedding model.

Our approach to dataset design is similar in spirit to the bAbi framework (Weston et al., 2016), which used simple synthetic tasks probing world knowledge and reasoning; however, our items are both simpler in design (they target individual concepts and do not, as of yet, require multi-chain reasoning or transitive inference) and harder in practice (a minimal pair, context-dependent design greatly reduces the availability of response heuristics, a serious problem in bAbi; Kaushik and Lipton, 2018).

The minimal pair design is common in datasets inspired by psycholinguistics and cognitive science, such as SyntaxGym (Gauthier et al.,

2020), BLiMP (Warstadt et al., 2020), and COMPS (Misra et al., 2023). In particular, it has previously been used to evaluate the models' ability to distinguish plausible and implausible events (Pedinotti et al., 2021; Kauf et al., 2023), a task that draws heavily on commonsense knowledge. The popular Winograd Schema Challenge also had a minimal pairs setup (Levesque et al., 2012), even though that requirement was relaxed in later versions. We here extend this approach by employing a *minimal pairs-of-pairs* design, where both context and target sentences have a minimal pair counterpart.

Until 2023, the dominant approach of assessing language model performance on a minimal pair has been to calculate each item's (pseudo) log probability under the model. This method is effective at distinguishing grammatical and ungrammatical sentences (e.g., Warstadt et al., 2020), plausible and implausible events (Kauf et al., 2023), and relevant vs. irrelevant object properties (Misra et al., 2023), while often being calibrated to human sentence ratings and processing costs (Lipkin et al., 2023; Shain et al., 2024). Yet raw log probabilities reflect a number of surface-level properties of the input, such as word frequency (Kauf et al., 2023) and the number of possible paraphrases (Holtzman et al., 2021). An alternative approach, recently made possible with more powerful LLMs, is to prompt an LLM to to rate item plausibility, either absolute (on a Likert scale) or relative to the other item in the minimal pair. This approach can theoretically result in more task-specific estimates. However, for a range of linguistic and word prediction tasks, LLMs actually perform worse with direct prompting than via implicit log probability assessment (Hu and Levy, 2023), likely because of additional task demands imposed by the need to decipher instructions in the prompt (Hu and Frank, 2024). Thus, we report both log probability comparisons and explicit prompting results.

3 The Framework

We provide a flexible generative synthetic data pipeline (Figure 1), capable of producing many diverse datasets, each with unique specifications and statistics, while preserving metadata and decision traces. In Section 6, we use this framework to generate EWoK-CORE-1.0, a systematic, broad-coverage, context-sensitive world knowledge dataset containing 4,400 items.

Item format Effective use of world knowledge incorporates both access to a rich set of priors about the world's structure and the ability to integrate this information on-the-fly with surrounding context. Thus, our items challenge models to leverage concepts in context. Each item consists of two minimal pair contexts (e.g., C_1: *The piano is in front of Ali. Ali turns* ***left****.*, C_2: *The piano is in front of Ali. Ali turns* ***right***) and two minimal pair target sentences (e.g., T_1: *The piano is* ***right*** *of Ali.*, T_2: *The piano is* ***left*** *of Ali.*). The two target concepts are juxtaposed such that in any item, $P(T_1 \mid C_1) > P(T_1 \mid C_2)$ and $P(T_2 \mid C_1) < P(T_2 \mid C_2)$. Thus, base target probabilities $P(T_1)$ and $P(T_2)$ cannot serve as plausibility cues: a model has to rely on context to establish plausibility.

Domains & concepts EWoK is designed around domains of general world knowledge, with the current set shown in Figure 3. Each domain includes a set of concepts. For example, the domain *social relations* includes *friend*, *enemy*, *teacher*, *student*, *boss* and *subordinate*, and so on.

Dataset generation procedure Each concept is associated with several items that test knowledge of the concept (often, but not always, by contrasting it with another concept). Items are created in a flexible yet controlled manner using the EWoK dataset generation procedure. At the core of the framework, we have atomic units and combination rules that, subject to constraints, lead to the generation of templates which are then populated with fillers—this leads to generating many more carefully-controlled items than a single-step template-filling approach could generate.

Contexts & targets A *target* is a simple sentence that incorporates a concept; each domain has one or more associated targets depending on the lexical properties of the concepts in that domain. A contrasting target pair is generated using one of two mechanisms: *concept swap*, which contrasts the same target with different concepts filled in, and *variable swap*, which swaps two objects or agents mentioned in the target (only possible for certain concepts). For instance, `{agent1} is to the left of`

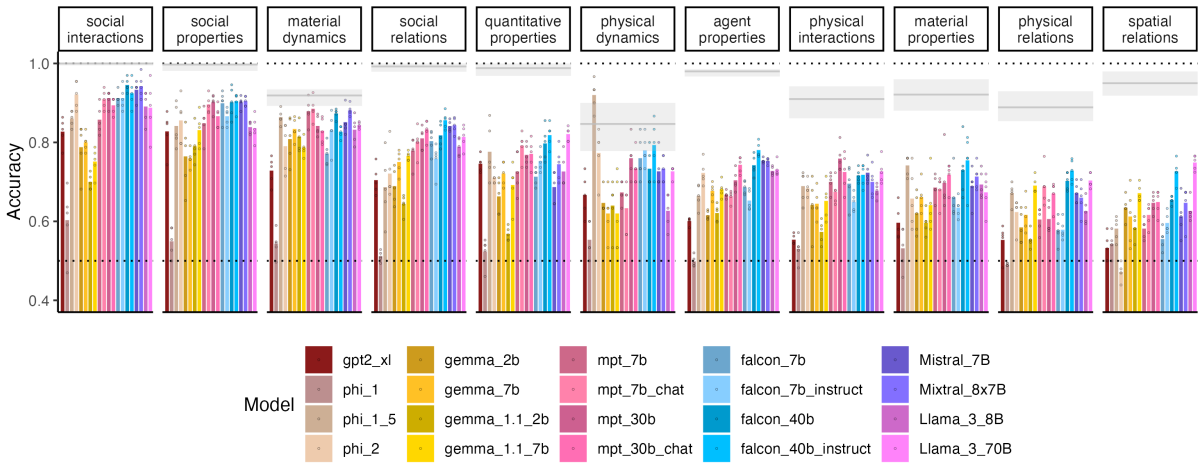

Figure 4: LLM performance across domains (evaluated with LOGPROBS). Here and elsewhere, the dotted line at 0.5 denotes chance accuracy. The light gray rectangle shows the range of human accuracy across 5 dataset versions, with the light gray line showing the mean. Each dot reflects LLM performance on a single version of EWOK-CORE-1.0, with the bar reflecting the mean across the 5 versions. LLM performance varies drastically by domain and is often substantially worse than human performance. In general, individual LLMs show similar performance patterns across domains, but these patterns are not always consistent with the human pattern.

`{agent2}` can be contrasted with `{agent1} is to the right of {agent2}` (concept swap) or with `{agent2} is to the left of {agent1}` (variable swap).

A *context pair* is one or more minimal pair(s) of sentences that can each be paired with a target pair such that C_1 but not C_2 matches T_1 and C_2 but not C_1 matches T_2. They are typically specific to an opposing concept pair (*left/right*) or to a single concept (*left*; only possible when variable swap is allowed). Similarly, a contrasting context pair is generated using one of two mechanisms, namely, a *filler swap* and a *variable swap*. A filler swap uses contrasting fillers to generate two versions of the item. A variable swap changes positions of two entities of the same kind to change the nature of the relationship or dynamic between them as long as the concept supports it.

Templates & fillers Each collection of concepts, contexts, and targets can then be compiled to a set of *templates*, partial items with typed variables describing the range of fillers for which the intended meaning holds. For example, in the template `{object2:can_bounce=True} bounced off {object1} from below`, any object may fill `{object1}`, e.g., `the desk` or `the crate`, whereas `{object2:can_bounce=True}` must be filled by an object marked with a flag for `can_bounce=True`, e.g., `the ball` or `the tire`. Across domains we list over 500 filler items across 13 classes with 28 type restrictions.

When populating the templates with fillers, users can specify various custom parameters, e.g.: (1) specify the number of items to generate from each template; each full set of items is referred to as a *version*; (2) select whether fillers should be held constant across all items in a version or allowed to vary; and (3) apply transformations to filler restrictions at compile-time, for instance restricting all agents to take non-Western names via `agent->agent:western=False`, or swapping all objects with nonce words via `object->nonword`. Such flexibility allows for controlled experimentation of the features modulating model performance. With these arguments and others easily accessible from the command line interface, users are supported to generate an enormous space of EWOK variations, and may easily add new domains, concepts, fillers, and more.

4 Evaluation

Using the EWOK framework, we compiled the EWOK-CORE-1.0 dataset by generating 5 unique fixed substitutions of filler items across 880 templates from 11 domains. In this section, we describe the evaluation of EWOK-CORE-1.0 via three different evaluation paradigms: traditional

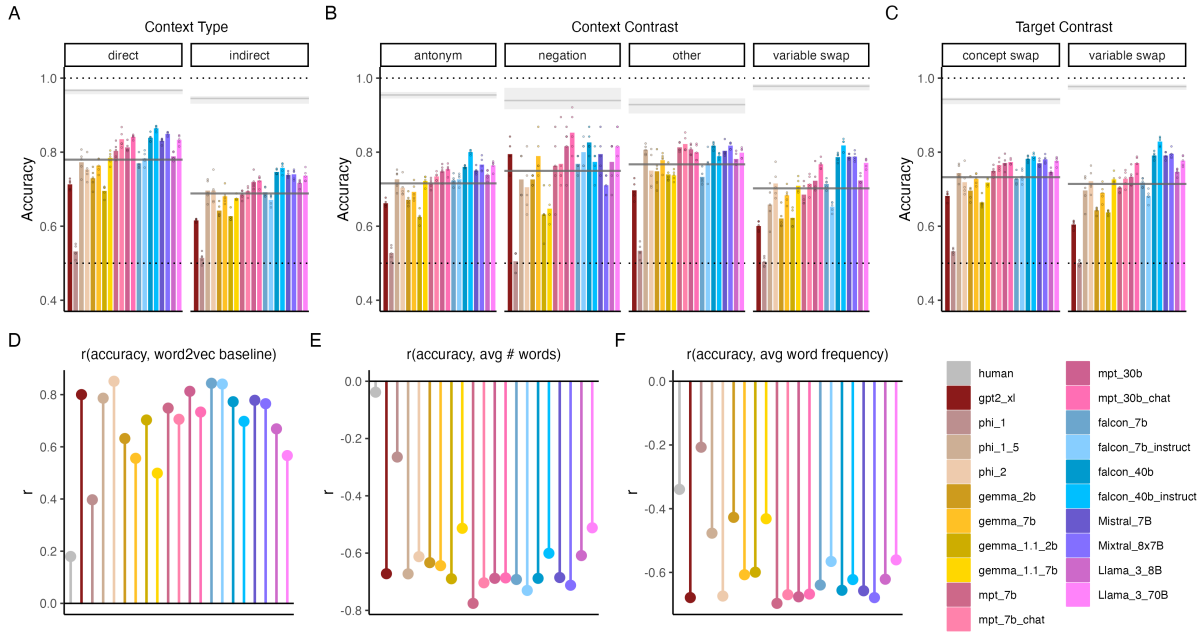

Figure 5: *Top:* LLM performance across context types (A) and minimal pair contrast for contexts (B) and for targets (C), evaluated with LogProbs. For examples of manipulations in A-C, see Figure 3. The light gray rectangle shows the range of human accuracies across dataset versions, with the light gray line showing the mean. Dark gray line shows average model performance. *Bottom:* correlation between LLM accuracy and surface-level item features: (D) word2vec bag-of-words baseline, (E) average item length, and (F) average word frequency in the item. Humans are not sensitive or only weakly sensitive to these features, whereas LLM performance strongly correlates with them. The (counterintuitive) negative relationship between accuracy and word frequency is driven by the fact that hard domains happen to have high word frequency and is reversed once domain is controlled for (Table S3).

plausibility estimates via querying the probability of sentences under the model, LogProbs, as well as two prompt-based strategies, Likert and Choice. The majority of the results reported use the LogProbs evaluation method, which allows comparing a wide range of base and finetuned models. As we show, LogProbs outperforms direct prompting even for large and/or instruction-tuned models.

For the prompt-based evaluations, we collected data from both LLMs as well as human participants using paired identical prompts. Drawing inspiration from comparative psychology, such an approach of matched evaluation has been proposed as a way to support increasingly *fair* evaluations of LLMs, allowing for more direct comparison of performance, consistency, and context-sensitivity with human participants (Lampinen, 2022).

4.1 Scoring metrics

For LogProbs evaluation, we use token-level LLM probabilities to calculate $\log P_\theta(T \mid C)$ as a sum of conditional log probabilities of each token: $\sum_{k=1}^{n} \log P_\theta(\mathbf{t}_k \mid C, \mathbf{t}_{<k})$, where $\mathbf{t}$ is the vector of tokens composing the target T. For Likert, participants (humans and models) are prompted to rate the plausibility of the concatenation of each C_i and T_j pair on a 1–5 scale. For Choice, participants (humans and models) are presented with C_1 and C_2, followed by a single target (T_1 or T_2), and then prompted to select the context (1 or 2) that better matches the target. For Likert and Choice, details about text generation hyperparameters can be found in A.3 and exact prompt templates can be found in Appendix A.4. The metric for correctness of a given item is the recovery of the designed item structure that $\texttt{score}(T_1 \mid C_1) > \texttt{score}(T_1 \mid C_2)$ and $\texttt{score}(T_2 \mid C_1) < \texttt{score}(T_2 \mid C_2)$, where $\texttt{score}$ reflects P_θ for LogProbs, the integer rating for Likert, and correct context index selection for Choice. In all cases, participants must correctly identify both C, T matches to get the full score (1.0 point). Identification of only one match receives 0.5 points. In the case of

the LIKERT task, if a model returns the same rating for both pairs, the model receives 0.5 points. Such a paradigm supports a trivial 50% baseline for all scenarios. Either a random coin flip or a deterministic model generating the same response for each query independent of context will trivially achieve this baseline.

4.2 Models

We evaluated N=20 transformer language models, selected to span a few points in the model design space. Models primarily vary in size (# of parameters; ranging from 1.3B–70B) and pre-training diet (both # of tokens and source of training corpora). While most evaluated LLMs are dense pre-trained transformers (N=13), there are a few one-off comparisons supported including the presence of supervised fine-tuning for instructions (N=4) or chat (N=2), and mixture-of-experts (MoE) ensembling (N=1). We do not intend to draw conclusions about any of these design decisions, but rather to expose variation. Aside from these considerations in exploring variation, the selection of fine-tuned models was filtered to those that did not require specific formatting via the use of a prompt template. Since we evaluate LLMs and humans on identical prompts, it was critical to have complete flexibility in formatting. The full set of evaluated models are listed in Fig. 4 as well as in Appendix A.2.

4.3 Surface-level item properties

We additionally tested a baseline bag-of-words model based on word2vec embeddings (Mikolov et al., 2013). Embeddings for each word in a context or a target were summed together to derive one vector per context/target. The context/target match was determined by a cosine similarity metric, such that an item is scored correctly if $\cos(C_1, T_1) > \cos(C_2, T_1)$ (and vice versa for T_2). We then examined the extent to which LLM and human performance correlates with the word2vec baseline. We also tested whether LLM performance correlates with the number of words in an item, as well as with average word frequency in an item using unigram counts from the Google Ngrams (Michel et al., 2011) 2012 American English corpus.

4.4 Human data

For independent norming of the items in EWOK-CORE-1.0, we collected data from human participants. We recruited a total of N=1,262 participants (591 female, 579 male, 27 other; median age 36; all US-residents with English as their first language) via Prolific, an online study platform. Raters with poor agreement with others ($R < 0.3$) were excluded. The task participants performed was nearly identical to the LIKERT version of the task for LLMs (in a pilot study, we determined that human results for LIKERT and CHOICE show $R = 0.96$ correlation; see Appendix A.1 & B.1 for more details).

5 Release Considerations

Our release-related goals are to (a) reduce the chances of accidental incorporation of EWOK items into LLMs' training data and (b) promote accountability and reporting when such incorporation is done intentionally. Thus:

- EWOK-CORE-1.0 is released on HuggingFace Datasets (Lhoest et al., 2021) with gated user access to prevent scrapers from accessing it automatically. Users will simply accept a CC-BY license and accompanying Terms of Use (TOU) wherein they will agree to explicitly report any instances when a language model was trained on the EWOK-CORE-1.0 items, and will be granted access automatically. Link: https://huggingface.co/datasets/ewok-core/EWoK-core-1.0

- The code for the EWOK item generation framework is shared in a separate repository on GitHub, with template files downloadable as password-protected archives to prevent automatic scraping. The repository is also protected with a TOU that requires anyone training or fine-tuning on any data generated using EWOK to report that fact. Link: https://github.com/ewok-core/ewok

- The code required to replicate the results in this paper, along with human study and model performance data, is shared as a separate GitHub repository following the same protections. Link: https://github.com/ewok-core/ewok-paper.

Although not flawless, this strategy allows us to minimize the chances of accidental incorporation

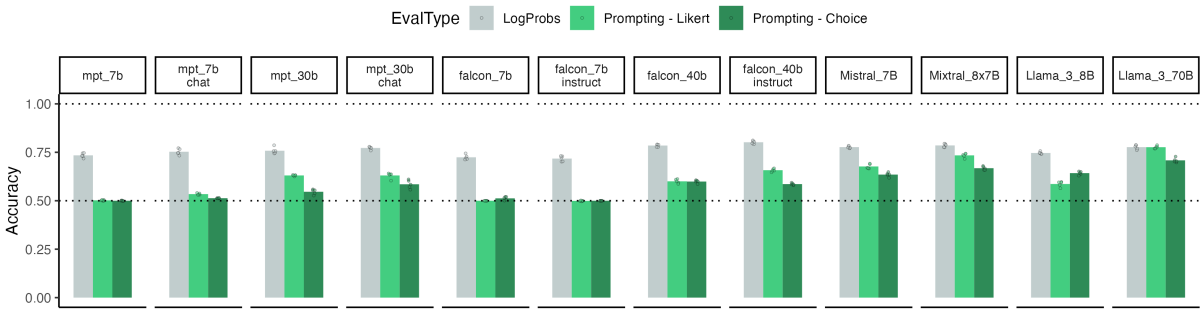

Figure 6: LLM performance assessed with LogProbs vs. two prompt-based tasks, Likert and Choice. The prompting setup is 2-shot, and the outputs are constrained to the set of allowed values ($1-5$ for Likert, 1 or 2 for Choice); this setup was chosen to maximize model performance. LogProbs is a better strategy in nearly all cases. Within the two prompting tasks, it was common for models to always generate the same value, e.g., "1" in response to any item. Our metric was designed such that even in this scenario, the 50% baseline would remain intact (see Section 3). Looking at Likert, without this safeguard and requiring strict inequality, the top performing Meta-LLama-3-70B model drops to 59%. MPT-7B, which performs comparably with LogProbs, drops to 2% (See Table 4). Thus, these data are far from "solved" with prompting in the general case.

of our data into training sets, as well as to send a clear signal to model developers about the importance of accountable reporting of evaluation dataset use during model training.

6 Experiments: EWoK-core-1.0

EWoK-core-1.0 is challenging for LLMs Despite the fact that EWoK-core-1.0 was not designed to be a hard dataset, we found that even the larger models generally perform much below humans: the best model tested, falcon-40b-instruct, has an accuracy of 0.80 whereas humans perform at 0.95 (Table A1). As expected, larger models tend to do better, although the performance gains are modest. Instruction tuning does not consistently increase or decrease LLM performance under the LogProbs metric.

Performance varies drastically by domain Figure 4 shows model results by domain, from easiest to hardest (see also Table A2). *Social interactions* is the easiest for both LLMs (mean=0.86, best=0.95) and humans (1.0); *spatial interactions* is both hardest for LLMs (mean=0.62, best=0.75) and shows the largest gap with human performance (who score highly at 0.96). Overall, domain difficulty is consistent across LLMs, such that hard domains are hard for all and vice versa. One notable exception is the heterogeneous performance of the phi models, with phi-1 consistently among the worst models, phi-1.5 outperforming all models and even humans on physical dynamics, and phi-2 ranging from on par with the largest models on some domains to worse than gpt2-xl on *spatial relations*. These results can perhaps be attributed to their unique training procedure, which focuses prominently on LLM-generated synthetic data.

LLMs show heterogeneous performance across dataset versions Our generation framework is designed to allow easy substitution of different values for names, objects, and locations, among other variables. In principle, these values should not affect the results (T_1 still best matches C_1 and not C_2). However, models show somewhat different performance on the 5 different versions of the dataset, with phi-2 and phi-1.5 showing the largest performance range (0.07 for both). This heterogeneity indicates that arbitrary item choices can substantially affect model performance, such that, in our dataset and in others, "1%" improvements might not be truly meaningful because they would not generalize. Humans show somewhat heterogeneous performance too (range 0.02), although that heterogeneity is driven only by a subset of the domains (Figure 4).

Domain content, item design features, and surface-level item features all affect LLM performance It is critically important to test whether cross-domain performance differences are indeed driven by domain content or are rather attributable to other factors. We consider item design features (direct/indirect context; contrast type for context sentences; and contrast type

for target sentences) and surface-level features (average word frequency and sentence length). Figure 5 shows that all these factors affect model performance, often in different ways than they affect humans. Performance of a bag-of-words word2vec-based baseline on different dataset domains and versions is predictive of LLM but not human performance (although baseline performance overall is only .39). The number of words in an item negatively affects LLM but not model performance. Unexpectedly, we found a negative relationship between word frequency and both LLM and human performance; follow-up examination of the data showed that this effect is driven by the fact that *physical-relations* and *spatial-relations*, the two hardest domains for LLMs, have the highest word frequency. To evaluate relative contributions of these factors to model performance, we jointly modeled all features using mixed effects regression. Design features and surface-level features all contributed to model performance (Table A3), with word frequency having a significant positive effect and the number of words having a significant negative effect, as expected. Importantly, domain remained a significant predictor of performance even when accounting for other factors.

LogProbs yield higher accuracy than prompting Evaluating LLMs with LogProbs resulted in above-chance accuracy for almost all models (Figure 6). As expected, the gap between LogProbs and prompting was larger for smaller models (Hu and Frank, 2024). We used the same prompting setup across all models to assess them in the same way as humans (Lampinen, 2022) (although note that models are also aided via 2-shot examples). It is possible that targeted prompt-engineering or alternative evaluation strategies will result in above-LogProbs performance, a direction we leave to future work.

Human ratings are often, but not always accurate Finally, we examined discrepancies between human ratings and experimental labels. Sometimes, the discrepancy resulted from specific fillers changing the plausibility of a C, T pair; for instance *The cooler is inside the car. Chao cannot see the cooler.* is implausible because *the cooler* is large and *the car* has windows, although, for smaller objects and containers without windows, the scenario is plausible. Sometimes, humans made mistakes. One such example is cardinal directions from the *spatial relations* domain. The scenario *The bakery is north of Chao. Chao turns around. The bakery is south of Chao.* is implausible because cardinal directions do not depend on the agent's orientation, and yet our participants often marked it as plausible. Overall, human data collection is a valuable source of information on our dataset but it does not replace ground truth labels.

7 Discussion

We present a systematic, flexible framework that can be used to test basic world knowledge in language models. Our goal was to develop a dataset that: (1) uses a uniform item format to probe diverse domains of physical and social knowledge, (2) contains items that probe specific concepts ("elements of world knowledge") within these domains, (3) requires integrating information across sentences, such that the same target sentence is plausible given one context and implausible given another, and (4) consists of generic templates that can be used to generate a wide variety of items.

We then presented evaluation results for a set of openly available models on EWoK-core-1.0, a dataset generated using the EWoK framework. This dataset is moderately challenging for LLMs, with performance varying substantially across domains. We also show that LogProbs contain enough information for most LLMs to perform above chance and that evaluation via prompting underperforms relative to LogProbs even for better-performing models, in agreement with prior results (Hu and Frank, 2024; Hu and Levy, 2023; Hu et al., 2024; Kauf et al., 2024).

The EWoK framework opens up multiple avenues for future work:

Targeted experiments The flexibility of our framework allows conducting specific experiments using customized sets of fillers. For instance, one might investigate whether LLMs perform differently on items that include western vs. nonwestern names, items that refer to people by names vs. longer descriptors ("the man in the black hat"), or even items featuring nonwords (like "*florp*") instead of real object names.

Interpretability research Knowledge editing research (e.g., Meng et al., 2022, 2023) has

often focused on encyclopedic knowledge; but what about knowledge of basic physical and social concepts? Our controlled minimal pair stimuli can allow researchers to identify and manipulate model circuits that might be selectively responsible for knowledge of specific concepts from an array of domains.

From elements to world models For a model to function as a flexible and robust general purpose AI system, it needs to be able to construct, maintain, and update internal world models (Ha and Schmidhuber, 2018; LeCun, 2022) (in cognitive science, variants of such world models are also known as mental models or situation models). The extent to which LLMs possess and use internal world models is subject to ongoing investigation (Hao et al., 2023; Yildirim and Paul, 2024; Wong et al., 2023). The EWoK framework offers an opportunity to combine individual elements of world knowledge to construct multi-step scenarios for evaluating world modeling capabilities in LLMs, within and across physical and social knowledge domains.

Limitations Our dataset is written in English; LLM performance might be lower on other languages, especially under-resourced ones. Adapting the EWoK framework to other languages might require redesigning the set of concepts and materials we use, which are currently grounded in the English lexicon. Thus, a multilingual framework can help more cleanly dissociate linguistic and conceptual effects on model performance. Another limitation is that we use the same prompting setup for all models. With tailored prompt engineering, alternative generation methods, or chain-of-thought, LLM performance could improve. Regardless, we expect model generation to conform with general world knowledge under a wide range of evaluation scenarios—the fact that they do not is informative. Finally, due to the synthetic nature of our dataset, some items are semantically weird. Although not necessarily a problem, an alternative is to use LLMs to populate templates. We leave this approach to future work noting that it can introduce confounds that inflate model performance (e.g., Panickssery et al., 2024).

8 Conclusion

To evaluate the ability of LLMs to construct robust world models, we need to test their ability to reason about the fundamental elements of world knowledge. The EWoK framework provides a way to systematically evaluate such knowledge, highlights that LLMs continue to fall short on simple scenarios requiring physical, spatial, or social knowledge, and offers opportunities for further targeted evaluations of LLMs.

Acknowledgements

We thank everyone who contributed ideas on dataset design, especially Hayley Ross and Yuhan Zhang. We thank members of LINGO Lab at MIT for providing feedback on earlier versions of the work. This work was supported by the Language Mission of the MIT Quest for Intelligence.

References

Mostafa Abdou, Artur Kulmizev, Daniel Hershcovich, Stella Frank, Ellie Pavlick, and Anders Søgaard. 2021. Can language models encode perceptual structure without grounding? a case study in color. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 109–132, Online. Association for Computational Linguistics.

AI@Meta. 2024. Llama 3 model card. *GitHub*.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Merouane Debbah, Etienne Goffinet, Daniel Heslow, Julien Launay, Quentin Malartic, et al. 2023. Falcon-40B: an open large language model with state-of-the-art performance. *Findings of the Association for Computational Linguistics: ACL*, 2023:10755–10773.

Peter W Battaglia, Jessica B Hamrick, and Joshua B Tenenbaum. 2013. Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, 110(45):18327–18332.

Emily M Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th annual meeting of the*

association for computational linguistics, pages 5185–5198.

Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. 2020. PIQA: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 7432–7439.

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

Stephanie M Carlson, Melissa A Koenig, and Madeline B Harms. 2013. Theory of mind. *Wiley Interdisciplinary Reviews: Cognitive Science*, 4(4):391–402.

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

Ido Dagan, Bill Dolan, Bernardo Magnini, and Dan Roth. 2010. Recognizing textual entailment: Rational, evaluation and approaches–erratum. *Natural Language Engineering*, 16(1):105–105.

Stanislas Dehaene. 2011. *The number sense: How the mind creates mathematics*. OUP USA.

Jon Gauthier, Jennifer Hu, Ethan Wilcox, Peng Qian, and Roger Levy. 2020. SyntaxGym: An online platform for targeted evaluation of language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 70–76.

Gemma, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on Gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Jonathan Gordon and Benjamin Van Durme. 2013. Reporting bias and knowledge acquisition. In *Proceedings of the 2013 workshop on Automated knowledge base construction*, pages 25–30.

Gabriel Grand, Idan Asher Blank, Francisco Pereira, and Evelina Fedorenko. 2022. Semantic projection recovers rich human knowledge of multiple object features from word embeddings. *Nature human behaviour*, 6(7):975–987.

Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, et al. 2023. Textbooks are all you need. *arXiv preprint arXiv:2306.11644*.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112.

David Ha and Jürgen Schmidhuber. 2018. World models. *arXiv preprint arXiv:1803.10122*.

Alon Hafri and Chaz Firestone. 2021. The perception of relations. *Trends in Cognitive Sciences*, 25(6):475–492.

Shibo Hao, Yi Gu, Haodi Ma, Joshua Hong, Zhen Wang, Daisy Wang, and Zhiting Hu. 2023. Reasoning with language model is planning with world model. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8154–8173.

Ari Holtzman, Peter West, Vered Shwartz, Yejin Choi, and Luke Zettlemoyer. 2021. Surface form competition: Why the highest probability answer isn't always right. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7038–7051, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jennifer Hu and Michael C Frank. 2024. Auxiliary task demands mask the capabilities

of smaller language models. *arXiv preprint arXiv:2404.02418*.

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.

Jennifer Hu, Kyle Mahowald, Gary Lupyan, Anna Ivanova, and Roger Levy. 2024. Language models align with human judgments on key grammatical constructions. *arXiv preprint arXiv:2402.01676*.

Anna A Ivanova, Aalok Sathe, Benjamin Lipkin, Evelina Fedorenko, and Jacob Andreas. 2024. Log probability scores provide a closer match to human plausibility judgments than prompt-based evaluations. In *South NLP Symposium*.

Ray Jackendoff. 2002. *Foundations of Language: Brain, Meaning, Grammar, Evolution*. Oxford University Press.

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

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

Carina Kauf, Emmanuele Chersoni, Alessandro Lenci, Evelina Fedorenko, and Anna A Ivanova. 2024. Comparing plausibility estimates in base and instruction-tuned large language models. *arXiv preprint arXiv:2403.14859*.

Carina Kauf, Anna A Ivanova, Giulia Rambelli, Emmanuele Chersoni, Jingyuan Selena She, Zawad Chowdhury, Evelina Fedorenko, and Alessandro Lenci. 2023. Event knowledge in large language models: The gap between the impossible and the unlikely. *Cognitive Science*, 47(11):e13386.

Divyansh Kaushik and Zachary C Lipton. 2018. How much reading does reading comprehension require? a critical investigation of popular benchmarks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5010–5015.

Michal Kosinski. 2023. Theory of mind may have spontaneously emerged in large language models. *arXiv preprint arXiv:2302.02083*, 4:169.

Andrew Kyle Lampinen. 2022. Can language models handle recursively nested grammatical structures? a case study on comparing models and humans. *arXiv preprint arXiv:2210.15303*.

Yann LeCun. 2022. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 62(1).

Hector Levesque, Ernest Davis, and Leora Morgenstern. 2012. The Winograd schema challenge. In *Thirteenth international conference on the principles of knowledge representation and reasoning*.

Molly Lewis, Martin Zettersten, and Gary Lupyan. 2019. Distributional semantics as a source of visual knowledge. *Proceedings of the National Academy of Sciences*, 116(39):19237–19238.

Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Šaško, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, Angelina McMillan-Major, Philipp Schmid, Sylvain Gugger, Clément Delangue, Théo Matussière, Lysandre Debut, Stas Bekman, Pierric Cistac, Thibault Goehringer, Victor Mustar, François Lagunas, Alexander Rush, and Thomas Wolf. 2021. Datasets: A community library for natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 175–184, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yuanzhi Li, Sébastien Bubeck, Ronen Eldan, Allie Del Giorno, Suriya Gunasekar, and Yin Tat

Lee. 2023. Textbooks are all you need ii: phi-1.5 technical report. *arXiv preprint arXiv:2309.05463*.

Benjamin Lipkin, Lionel Wong, Gabriel Grand, and Josh Tenenbaum. 2023. Evaluating statistical language models as pragmatic reasoners. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 45.

Shari Liu, Joseph Outa, and Seda Akbiyik. 2024. Naive psychology depends on naive physics.

Tianyu Liu, Zheng Xin, Baobao Chang, and Zhifang Sui. 2020. HypoNLI: Exploring the artificial patterns of hypothesis-only bias in natural language inference. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6852–6860, Marseille, France. European Language Resources Association.

Li Lucy and Jon Gauthier. 2017. Are distributional representations ready for the real world? Evaluating word vectors for grounded perceptual meaning. In *Proceedings of the First Workshop on Language Grounding for Robotics*, pages 76–85, Vancouver, Canada. Association for Computational Linguistics.

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*, 28(6):460–476.

Michael McCloskey. 1983. Intuitive physics. *Scientific american*, 248(4):122–131.

R Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2020. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *57th Annual Meeting of the Association for Computational Linguistics, ACL 2019*, pages 3428–3448. Association for Computational Linguistics (ACL).

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. *Advances in Neural Information Processing Systems*, 35:17359–17372.

Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. 2023. Mass editing memory in a transformer. *The Eleventh International Conference on Learning Representations (ICLR)*.

Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K Gray, Google Books Team, Joseph P Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, et al. 2011. Quantitative analysis of culture using millions of digitized books. *science*, 331(6014):176–182.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

Kanishka Misra, Julia Rayz, and Allyson Ettinger. 2023. COMPS: Conceptual minimal pair sentences for testing robust property knowledge and its inheritance in pre-trained language models. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2928–2949, Dubrovnik, Croatia. Association for Computational Linguistics.

MosaicML. 2023. Introducing MPT-30B: Raising the bar for open-source foundation models. *MosaicML*.

Arjun Panickssery, Samuel R Bowman, and Shi Feng. 2024. LLM evaluators recognize and favor their own generations. *arXiv preprint arXiv:2404.13076*.

Roma Patel and Ellie Pavlick. 2021. Mapping language models to grounded conceptual spaces. In *International conference on learning representations*.

Ellie Pavlick. 2022. Semantic structure in deep learning. *Annual Review of Linguistics*, 8:447–471.

Paolo Pedinotti, Giulia Rambelli, Emmanuele Chersoni, Enrico Santus, Alessandro Lenci, and Philippe Blache. 2021. Did the cat drink the coffee? Challenging transformers with generalized event knowledge. In *Proceedings of *SEM 2021: The Tenth Joint Conference on*

Lexical and Computational Semantics, pages 1–11, Online. Association for Computational Linguistics.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191, New Orleans, Louisiana. 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.

Brett D. Roads and Bradley C. Love. 2020. Learning as the unsupervised alignment of conceptual systems. *Nature Machine Intelligence*, 2(1):76–82.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106.

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

Cory Shain, Clara Meister, Tiago Pimentel, Ryan Cotterell, and Roger Levy. 2024. Large-scale evidence for logarithmic effects of word predictability on reading time. *Proceedings of the National Academy of Sciences*, 121(10):e2307876121.

Vered Shwartz and Yejin Choi. 2020. Do neural language models overcome reporting bias? In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6863–6870, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Ben Sorscher, Surya Ganguli, and Haim Sompolinsky. 2022. Neural representational geometry underlies few-shot concept learning. *Proceedings of the National Academy of Sciences*, 119(43):e2200800119.

Elizabeth S. Spelke and Katherine D. Kinzler. 2007. Core knowledge. *Developmental science*, 10(1):89–96.

Tomer Ullman. 2023. Large language models fail on trivial alterations to theory-of-mind tasks. *arXiv preprint arXiv:2302.08399*.

Akira Utsumi. 2020. Exploring what is encoded in distributional word vectors: A neurobiologically motivated analysis. *Cognitive Science*, 44(6):e12844.

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.

Jason Weston, Antoine Bordes, Sumit Chopra, and Tomás Mikolov. 2016. Towards AI-complete question answering: A set of prerequisite toy tasks. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.

Brandon T Willard and Rémi Louf. 2023. Efficient guided generation for large language models. *arXiv e-prints*, pages arXiv–2307.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Lionel Wong, Gabriel Grand, Alexander K Lew, Noah D Goodman, Vikash K Mansinghka, Jacob Andreas, and Joshua B Tenenbaum. 2023. From word models to world models: Translating from natural language to the probabilistic language of thought. *arXiv preprint arXiv:2306.12672*.

Ilker Yildirim and LA Paul. 2024. From task structures to world models: What do LLMs know? *Trends in Cognitive Sciences*.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

A Methods

A.1 Human data collection

We collected human data in two phases: a pilot study used to validate our task on a single domain and determine the measurement technique to use for data collection (CHOICE or LIKERT), and a main study where we repeated data collection for the full set of EWOK-CORE-1.0 items. In the pilot study we determined that human judgments are measurement-technique-invariant (Ivanova et al., 2024), so we did not collect CHOICE judgments on the full set of materials, instead relying on LIKERT judgments to score items. Results from both studies are discussed in Appendix B.1.

Pilot study The pilot study was done on materials from one of the EWOK-CORE-1.0 subdomains—*social relations*—using one set of variables to populate the fillers (not used in the main study). Participants from USA were recruited using Prolific, an online study platform, based on being self-reported native and fluent English speakers. We recruited a total of 30 participants across conditions. Of these, 18 reported identifying as 'female', 11 as 'male', and 1 preferred not to answer. Participants were assigned to either the LIKERT or the CHOICE condition, and saw items in only one of the two measurement techniques.

Each item in LIKERT was split into four sub-items: $(C_1, T_1), (C_1, T_2), (C_2, T_1), (C_2, T_2)$. Similarly, each CHOICE item was split into two sub-items: $(C_{\{1,2\}}, T_1), (C_{\{1,2\}}, T_2)$. A total of 16 participants provided LIKERT-scale judgments, whereas 14 provided CHOICE responses to the items. Most LIKERT sub-items (C_i, T_j) received at least 4-5 judgments, with all items receiving at least 3 judgments. The average no. of ratings per sub-item were 4. All CHOICE sub-items received 7 judgments per $(C_{\{1,2\}}, T_y)$ pair. Participants never saw more than one sub-item of the same item (i.e., participants couldn't rate both (C_1, T_1) and (C_2, T_1) in the LIKERT study).

Main study To obtain reliable ratings across the full set of EWOK-CORE-1.0 items, we collected at least 5 responses per item. To control for quality of data, we excluded 59 participants whose inter-subject correlation (compared with average ratings on items from other participant who rated the same items as this subject) was below 0.3. Each EWOK-CORE-1.0 generated item (corresponding to 5 variable-populated versions) was split into four subparts: $(C_1, T_1), (C_1, T_2), (C_2, T_1), (C_2, T_2)$ and presented in a LIKERT setting using the same prompt as that used for LLMs (Appendix A.4; adapted to presentation in a web browser with a free-form text-box for human input). Items were presented so that each participant only saw one of 5 variable-populated variants and one of the four possible sub-items (C_i, T_j). Therefore, participants provided sensibility judgments independently of any other subparts of the item, closely matching the conditions for LLM evaluation. Each participant saw a list of 57 items on average, depending on the domain (188 for *agent properties* and 27 for *material properties*).

A.2 Evaluated models

The full set of evaluated models are as follows: gpt2-xl (Radford et al., 2019), phi-1 (Gunasekar et al., 2023), phi-1.5, phi-2 (Li et al., 2023), gemma-2b, gemma-1.1-2b-it, gemma-7b, gemma-1.1-7b-it (Gemma et al., 2024), mpt-7b, mpt-7b-chat, mpt-30b, mpt-30b-chat (MosaicML, 2023), falcon-7b, falcon-7b-instruct, falcon-40b, falcon-40b-instruct (Almazrouei et al., 2023), mistral-7b-v0.1 (Jiang et al., 2023), mixtral-8x7b-v0.1 (Jiang et al., 2024), Meta-Llama-3-8B, and Meta-Llama-3-70B (AI@Meta, 2024). All models were accessed via HuggingFace transformers (Wolf et al., 2019), and all experiments were run on a 4x$A100$ 80GB GPU cluster.

A.3 Text completion experiments

For the prompting-based evaluations, CHOICE and LIKERT, we support two different generation options: *free* and *constrained*. For free generation, the LLM may greedily sample up to 20 tokens, and we match the first occurrence of a valid response (a numeral between 1–2 or 1–5) with a regular expression. Such a strategy avoids penalizing completions that begin with text or white-space, but doesn't guide the model to produce a valid response. For constrained generation, the LLM may greedily sample from a restricted set of tokens, either 1–2 or 1–5, constrained using logit masking (Willard and Louf, 2023). Such a strategy enforces well-structured responses, but requires a restricted response format. In addition to variation in generation options, we support both

zero- and few-shot prompting. In Figure 6, the prompting results we report use 2-shot constrained generation, as it yields the highest performance among our space of tested strategies.

A.4 Prompts

For the two prompt-based evaluations, CHOICE and LIKERT, we include below our exact prompt templates. The LIKERT prompt was additionally used verbatim for human data evaluation.

CHOICE Template:

```
# INSTRUCTIONS

In this study, you will see
   ↪ multiple examples. In each
   ↪ example, you will be given
   ↪ two contexts and a scenario
   ↪ . Your task is to read the
   ↪ two contexts and the
   ↪ subsequent scenario, and
   ↪ pick the context that makes
   ↪  more sense considering the
   ↪  scenario that follows. The
   ↪  contexts will be numbered
   ↪ "1" or "2". You must answer
   ↪  using "1" or "2" in your
   ↪ response.

# TEST EXAMPLE

## Contexts
1. "{context1}"
2. "{context2}"

## Scenario
"{target}"

## Task
Which context makes more sense
   ↪ given the scenario? Please
   ↪ answer using either "1" or
   ↪ "2".

## Response
```

LIKERT Template:

```
# INSTRUCTIONS

In this study, you will see
   ↪ multiple examples. In each
   ↪ example, you will be given
   ↪ a scenario. Your task will
   ↪ be to read the scenario and
   ↪  answer how much it makes
   ↪ sense. Your response must
   ↪ be on a scale from 1 to 5,
   ↪ with 1 meaning "makes no
   ↪ sense", and 5 meaning "
   ↪ makes perfect sense".

# TEST EXAMPLE

## Scenario
"{context} {target}"

## Task
How much does this scenario make
   ↪ sense? Please answer using
   ↪ a number from 1 to 5, with
   ↪ 1 meaning "makes no sense",
   ↪  and 5 meaning "makes
   ↪ perfect sense".

## Response
```

A.5 Mixed effects modeling

To evaluate joint effects of domains, item design factors, and item surface features on LLM performance, we entered those predictors into a mixed effects logistic regression model implemented in R using *lme4*. The model formula is: *Accuracy* $\sim$ *0 + Domain + ContextContrast + TargetContrast + ContextType + Frequency + NumWords + (1|Model) + (1|Item)*

See Section 3 for possible values for *Domain*, *ContextType*, *ContextContrast*, and *targetContrast*. Domain effects were estimated relative to a 0 intercept. *ContextType*, *ContextContrast* and *TargetContrast* had deviation contrast coding. Relative word frequency and number of words per item were computed as $(C_1 + C_2)/2 + T$ (for either T_1 or T_2); these values were z-scored before being entered as regressors. *Model* refers to an LLM being used, and *Item* refers to each individual item, i.e. a minimal pair of pairs. The model was fit on item-level binary accuracy data, with 1 row per target sentence. The results we report are from model performance using LOGPROBS evaluation type. See Table 3 for results.

B Results

B.1 Human study

Pilot study: Human judgments are invariant to LIKERT or CHOICE measurement We determined humans are closely aligned when providing LIKERT-scale or CHOICE judgments (Fig. 7). LIKERT-scale judgments are less dependent on the specific set up (making a CHOICEjudgment requires a specific framing eliciting a comparison of two contexts given a target). In order to have data allowing more flexible comparisons we decided to stick to LIKERT-scale judgments for the full data collection in the main study.

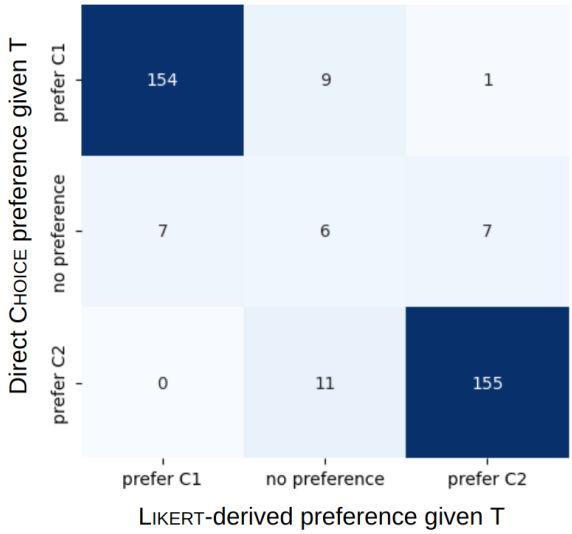

Figure 7: Confusion matrix showing agreement between a direct choice of $\{C_1, C_2\}|T_i$ and an indirect choice based on individual LIKERT ratings $\{C_i|T_i\} > \{C_j|T_i\}$. In this comparison we marked items as "no preference" if (1) the Likert scores weren't at least 1 apart, or (2) the average choice preference wasn't at least 0.2 away from the midpoint between two alternatives.

B.2 Model evaluation

Table 1: Performance on EWoK-CORE-1.0. Range reported across 5 dataset versions.

Model	Mean LogProbs Accuracy	Range
human	**0.951**	**0.942-0.957**
gpt2_xl	0.655	0.645-0.662
phi_1	0.522	0.517-0.530
phi_1_5	0.727	0.686-0.756
phi_2	0.718	0.696-0.771
gemma_2b	0.678	0.657-0.705
gemma_7b	0.714	0.697-0.734
gemma_1.1_2b	0.654	0.643-0.673
gemma_1.1_7b	0.720	0.708-0.736
mpt_7b	0.733	0.716-0.745
mpt_7b_chat	0.751	0.73-0.769
mpt_30b	0.757	0.743-0.786
mpt_30b_chat	0.771	0.758-0.777
falcon_7b	0.723	0.713-0.744
falcon_7b_instruct	0.717	0.701-0.730
falcon_40b	0.783	0.775-0.789
falcon_40b_instruct	**0.801**	**0.790-0.810**
Mistral_7B	0.775	0.768-0.783
Mixtral_8x7B	0.784	0.774-0.794
Llama_3_8B	0.746	0.740-0.756
Llama_3_70B	0.775	0.759-0.787

Table 2: LLM and human performance by domain.

Domain	LLM (average)	LLM (best)	Human
social interactions	0.859	0.945	1.000
social properties	0.839	0.905	0.997
material dynamics	0.816	0.885	0.911
social relations	0.761	0.856	0.992
quantitative properties	0.725	0.823	0.986
physical dynamics	0.706	0.920	0.833
agent properties	0.683	0.778	0.975
physical interactions	0.672	0.759	0.910
material properties	0.669	0.755	0.921
physical relations	0.627	0.723	0.886
spatial relations	0.615	0.749	0.958

Table 3: Domain, design, and surface level features jointly contribute to LLM performance. $*p < .05$; $**p < .01$; $***p < .001$

Predictor Type	Predictor	Effect
domain	social interactions	1.91 ***
	social properties	1.79 ***
	material dynamics	2.23 ***
	social relations	1.27 ***
	quantitative properties	1.09 ***
	physical dynamics	0.88 **
	agent properties	0.58 **
	physical interactions	0.83 ***
	material properties	0.75 **
	physical relations	0.38
	spatial relations	0.41
context contrast	antonym vs. rest	0.09 ***
	negation vs. rest	0.1 **
	variable swap vs. rest	0.0
target contrast	variable vs. concept swap	0.0
context type	direct vs. indirect	0.2 ***
surface features	word frequency	0.07 ***
	number of words	-0.04 **

Table 4: LLM LIKERT accuracy on EWOK-CORE-1.0 with stricter inequality metric.

Model	Mean Likert Acc	Range
mpt_7b	0.021	0.018-0.025
mpt_7b_chat	0.100	0.089-0.113
mpt_30b	0.310	0.307-0.316
mpt_30b_chat	0.307	0.25-0.332
falcon_7b	0.003	0.003-0.003
falcon_7b_instruct	0.005	0.004-0.008
falcon_40b	0.216	0.185-0.24
falcon_40b_instruct	0.353	0.337-0.368
Mistral_7B	0.396	0.375-0.429
Mixtral_8x7B	0.511	0.471-0.533
Llama_3_8B	0.195	0.145-0.222
Llama_3_70B	**.588**	**.576-.603**