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**[*3]   **Unnathi U. Kumar**[1]
Setayesh Radkani[3]   **Thomas H. Clark**[3]   **Carina Kauf**[3]   **Jennifer Hu**[4]   **R. T. Pramod**[3]
Gabriel Grand[3]   **Vivian C. Paulun**[3]   **Maria Ryskina**[5]   **Ekin Akyürek**[3]
Ethan G. Wilcox[6]   **Nafisa Rashid**[7]   **Leshem Choshen**[3,8]   **Roger Levy**[3]
Evelina Fedorenko[3]   **Joshua Tenenbaum**[3]   **Jacob Andreas**[3]

[1]Georgia Tech, USA   [2]Brown University, USA   [3]MIT, USA
[4]Johns Hopkins University, USA   [5]Vector Institute, Canada   [6]Georgetown University, USA
[7]UC Berkeley, USA   [8]MIT-IBM Watson AI, USA
`a.ivanova@gatech.edu`, `aalok@brown.edu`, `lipkinb@mit.edu`

Abstract

The ability to build and reason about models of the world is essential for situated language understanding. But evaluating world modeling capabilities in modern AI systems—especially those based on language models—has proven challenging, in large part because of the difficulty of disentangling *conceptual* knowledge about the world from knowledge of surface co-occurrence statistics. This paper presents Elements of World Knowledge (EWoK), a framework for evaluating language models' understanding of the conceptual knowledge underlying world modeling. EWoK targets specific concepts from multiple knowledge domains known to be important for world modeling in humans, from social interactions (*help, deceive*) to spatial relations (*left, right*). 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) and compare them with human performance. All tested models perform worse than humans, with results varying drastically across domains. Performance on social interactions and social properties was highest and performance on physical relations and spatial relations was lowest. Overall, this dataset highlights simple cases where even large models struggle and presents 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 (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 demonstrate substantial knowledge of language (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 (Figure 1). 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)** a set of *concepts* within each domain; **(c)** a set of *templates* (and modular components to procedurally generate these templates) that test knowledge of specific concepts by contrasting plausible and implausible context–target combinations; **(d)** a set of *fillers* to populate the templates, such that each

---
*authors contributed equally.

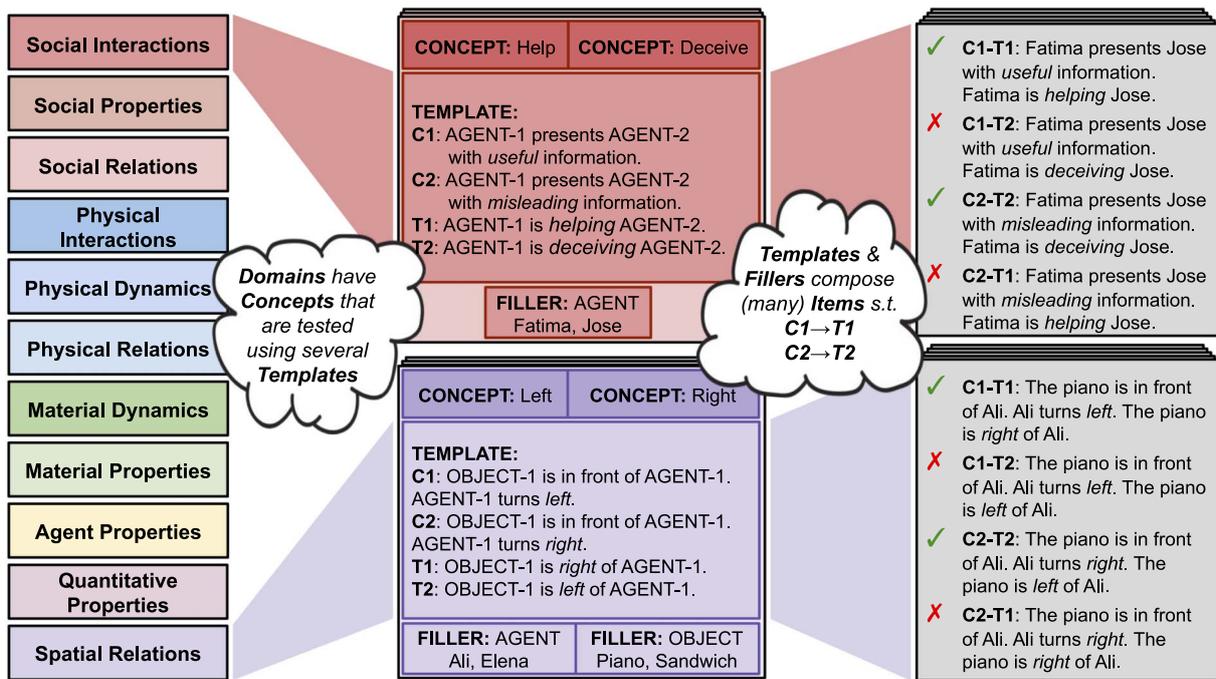[1]Data and associated code are available at: `https://ewok-core.github.io/`.

Figure 1: EWoK design, illustrated with examples from *social interactions* & *spatial relations*. 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*.

template can be used multiple times; **(e)** a pipeline to generate a specific set of *items* (a dataset) based on these source materials; and **(f)** an *evaluation procedure* that measures language models' ability to distinguish plausible and implausible items.

Why Elements? Our framework targets specific cognitive concepts (or concept pairs, such as *left/right*). Concept knowledge is not limited to memorized definitions, but can be used productively across a wide range of scenarios. Thus concepts (not isolated sentences or facts) are the primary focus of the EWoK framework. This approach stands in contrast to many NLP datasets, which often include naturalistic, lengthy stimuli that draw on multiple unspecified concepts. Although naturalistic tasks are useful for evaluating model performance in real-life settings, individual item complexity and lack of controls can make it difficult 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, 2024; Ullman, 2023). Our framework mitigates

this problem by explicitly linking individual items with the specific concepts that they test.

Why Cognition-inspired? World knowledge is a notoriously fuzzy capability; which knowledge domains should one focus on? Dataset design in NLP is often centered around data availability, with examples sourced from video captions, Wikipedia articles, or social media logs. In contrast, here, we systematically select a range of knowledge 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; Thomas, 2024), 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). At the same time, traces of these signals are deeply embedded in our daily communicated language. Text itself thus contains rich information that reflects grounded world knowledge (Roads and Love, 2020; Abdou

et al., 2021; Patel and Pavlick, 2021), and it may be reasonable to predict that LLMs would acquire this domain-specific knowledge from text alone.

Why Plausibility? Concept understanding is often fuzzy. Thus, deciding whether a factual statement is true or false in isolation is often ill-defined. Instead, we use combinations of plausible vs. implausible context–target pairs. In EWoK, a model needs to determine which scenarios make more sense (more plausible). Having an accurate world model would enable a model to consistently distinguish 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*) yields an opposite effect (e.g., *plausible → implausible*). This approach can help identify specific manipulations that LLMs are and are not sensitive to, with the goal of targeted diagnostics. Such controlled manipulations are also particularly well-suited for mechanistic interpretability research, which may seek to explain the circuits through which models deploy world knowledge in context.

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). Establishing this control is critical for isolating contextually-sensitive world knowledge from heuristic knowledge associated with surface forms.

Our paper is structured as follows. In Section 2, we review prior research related to the topics explored in this work. In Section 3, we describe the core components of the EWoK framework. In Section 4, we describe our evaluation strategy. In Section 5, we show LLM performance on EWoK-CORE-1.0, a dataset generated via the EWoK framework. In Section 6, we discuss our results and future prospects for this line of work. Finally, in Section 7, we discuss dataset release considerations.

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 or obvious information (Gordon and Van Durme, 2013). Thus, although our items target basic world knowledge and are not designed to be challenging, it is possible that LLMs could 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 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 to a system (a text expression T and hypothesis H) and asks that it 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 is held static while paired contexts are used to modulate preferences, rendering it impossible to rely solely on target plausibility; (2) annotating minimal pair contrast type to test whether any item design feature drives 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 require multi-chain reasoning) and are 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 test 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.

Before prompt-based evaluation became commonplace, the dominant approach for 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 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, capable of producing many diverse datasets, each with unique specifications and statistics, while preserving metadata and decision traces. Here, we discuss the design principles that make it possible to reuse the EWoK framework for systematically generating various world knowledge datasets. In Section 5, we use this framework to generate EWoK-CORE-1.0, a systematic, broad-coverage, context-sensitive world knowledge dataset containing 4,374 items that probe 192 concepts (for a full concept list, see Table A3).

Items Test Concept Knowledge in Context. 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 knowledge on-the-fly with information about the current environment. Thus, our items challenge models to leverage concepts in context. Each item consists of two minimal pair contexts (e.g.,

Domain	Templates	Concepts	Example
Social Interactions	165	16	C: AGENT-1 presents AGENT-2 with [useful/misleading] information. T: AGENT-1 is **[helping/deceiving]** AGENT-2.
Social Properties	185	16	C: AGENT-1 [can/cannot] be depended upon. T: AGENT-1 is **[trustworthy/untrustworthy]**.
Social Relations	785	15	C: AGENT-[1/2] lectures AGENT-[2/1]. T: AGENT-1 is AGENT-2's **[teacher/student]**.
Physical Interactions	280	20	C: OBJECT-1 is moving [toward/away from] the ground. T: AGENT-1 **[dropped/lifted]** OBJECT-1.
Physical Dynamics	75	12	C: The speck on OBJECT-1 [is/is not] rotating. T: OBJECT-1 is **[rolling/sliding]**.
Physical Relations	435	20	C: OBJECT-1 occupies [more/less] space than OBJECT-2. T: OBJECT-[1/2] is **bigger** than OBJECT-[2/1].
Material Dynamics	780	21	C: AGENT-1 sees something that is [fabric/liquid]. T: AGENT-1 can **[fold/pour]** it.
Material Properties	125	18	C: AGENT-1 [can/cannot] see through OBJECT-1. T: OBJECT-1 is **[transparent/opaque]**.
Agent Properties	1130	22	C: AGENT-1 likes OBJECT-1 [more/less] than OBJECT-2. T: AGENT-1 **prefers** OBJECT-[2/1] to OBJECT-[1/2].
Quantitative Properties	195	18	C: AGENT-1 [needs/does not need] more OBJECT-1. T: AGENT-1 has **[not enough/enough]** OBJECT-1.
Spatial Relations	245	14	C: OBJECT-1 is in front of AGENT-1. AGENT-1 turns [left/right]. T: OBJECT-1 is to the **[right/left]** of AGENT-2.

Table 1: EWoK-core-1.0 contains 11 domains, each contributing between 75 and 1130 templates testing between 12 and 22 concepts. Here we include sample templates (pairs of context–target pairs) for each domain. Each target includes an explicit mention of a concept from that domain (highlighted in bold).

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.

Concepts Comprise Domains of World Knowledge. EWoK is designed around domains of general world knowledge, with the current set of 11 domains shown in Table 1. We selected a range of domains that have been shown to recruit dedicated cognitive and/or neural machinery in humans, backed by extensive literature in cognitive science (introduced in Section 1, ''Why cognition inspired?''). Domains were contributed

based on past literature during the development of EWoK by a team of experts in the field (authors of this paper: professors, post-docs, and grad students in cognitive science and neuroscience across a few institutions). Each domain includes a set of concepts, also curated by the team. For example, the domain *social relations* includes *friend*, *enemy*, *teacher*, *student*, *boss*, *subordinate*, and others. The number of concepts in each domain is mentioned in Table 1.

Items are Generated from Concept-specific Templates. 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 generated from templates, which contain placeholders for specific objects, agents, and locations. The placeholder value should not affect the resulting plausibility judgments; thus, a template can be populated with an
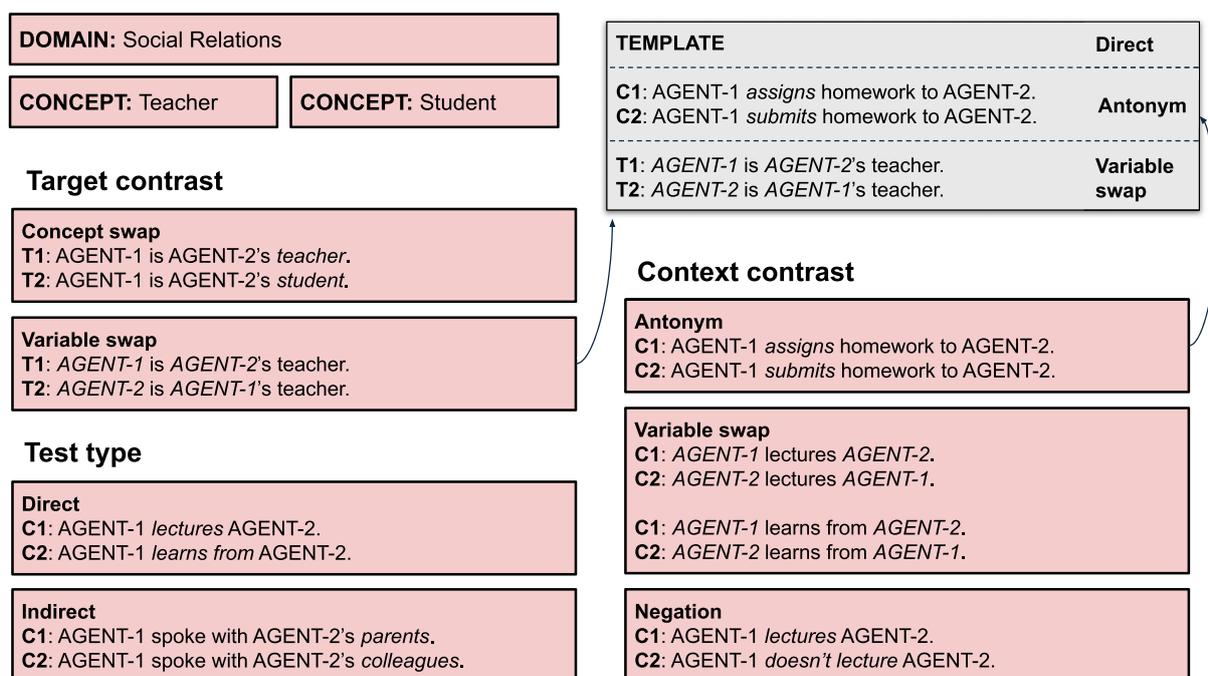
Figure 2: Types of minimal pair contrasts in Context and Target pairs. Examples shown here are for one domain and concept pair. Templates may be *direct*, testing concepts explicitly, or *indirect*, testing concepts implicitly using likely scenarios (e.g., a student is more likely to talk to a teacher's colleagues than parents are). Context and target *contrasts* reflect how concepts are tested. For instance, *antonym* contrasts words with opposing meanings, *negation* leverages ''not'', and *variable swap* exploits relative ordering of entities.

arbitrarily large number of fillers, which enables generating many carefully controlled items.

Templates Consist of Contrasting Context and Target Pairs. A template includes 2 targets and 2 contexts. A *target* is a simple sentence that incorporates a concept. A *context* is a short sequence of words that can be paired with a target to yield either a plausible or an implausible combination. Context and target pairs are designed in such a way that C_1 but not C_2 matches T_1 and C_2 but not C_1 matches T_2. The differences between the two targets or the two contexts result in specific types of contrasts, shown in Figure 2.

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 (only possible for certain targets). For instance, `AGENT-1 is AGENT-2's teacher` can be contrasted with `AGENT-1 is AGENT-2's student` (concept swap) or with `AGENT-2 is AGENT-1's teacher` (variable swap).

A contrasting context pair is generated using *antonyms*, *negation*, or *variable swap*.

Finally, the templates themselves test concept knowledge in either *direct* or *indirect* way. Direct tests rely on immediate interpretation of a concept, whereas indirect tests leverage the probabilistic relationship between concept and target (T_1 doesn't have to be true given C_1 but it's more likely than T_1 given C_2).

Templates Enable Flexible Yet Controlled Generation of Items Using Typed Fillers. Placeholders in templates (e.g., OBJECT or AGENT) can be restricted to only allow fillers of specific types. For example, in the template `{object2:can_bounce=True} bounced off {object1} from below`, `{object2:can_bounce=True}` must be filled by an object marked with a flag for `can_bounce=True`, e.g., `the ball`. These type restrictions prevent generation of semantically anomalous or incomprehensible sentences (e.g., `The blanket bounced off the table from below`). Type restrictions were specified by the author team during the template development stage and later validated via human behavioral ratings (Section 4.4 and Appendix A.2).

To create an initial set of object and location fillers, we generated a set of arbitrary fillers based on constraints specified in templates (e.g., objects that can bounce, that have a smell) that was filtered down to ensure the 'naturalness' of the items that include these fillers. The agent fillers were selected from the list of culturally diverse high frequency male, female and non-gendered baby names. Across domains we list over 500 filler items across 13 classes with 28 type restrictions. The full set of fillers is provided in the paper's GitHub repository.

When populating the templates with fillers, users wanting to use the EWoK framework to construct their own instance of an EWoK-CORE-1.0-like dataset can specify several parameters, e.g.: (1) the number of items to generate from each template; each full set of items is referred to as a *version* (corresponding to a random seed); (2) whether fillers should be held constant across all items in a version or allowed to vary; (3) whether to 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 many EWoK variations, and can additionally extend EWoK to add new domains, concepts, and fillers.

4 Evaluation

Using the EWoK framework, we compiled the EWoK-CORE-1.0 dataset. To do so, our team (which includes experts in specific domains of cognitive science) curated a list of 880 templates that test knowledge of 192 concepts from 11 domains. These templates were then populated with 5 fillers that were randomly sampled from our initial filler set (described in Section 3), resulting in 5 dataset versions. This strategy enables us to explicitly measure the variability associated with the random sampling of filler items (which, in principle, should not affect the results). For EWoK-CORE-1.0, we chose to hold the set of fillers constant across all items within each version, although that constraint can be relaxed if needed (see Appendix A.1).

We evaluated LLM performance on EWoK-CORE-1.0 using three different approaches: traditional 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, 2024).

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.[2] This is functionally similar to the LOGPROBS condition, but explicitly based on the model's generative behavior as opposed to intrinsic to its scoring. 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. This is another prompting approach that uses the language model discriminatively with access to both items. For LIKERT and CHOICE, details about text generation hyperparameters can be found in A.4 and exact prompt templates can be found in Appendix A.5. The metric for correctness of a given item is the recovery of the designed item structure that $\text{score}(T_1 \mid C_1) > \text{score}(T_1 \mid C_2)$ and $\text{score}(T_2 \mid C_1) < \text{score}(T_2 \mid C_2)$, where score reflects P_θ for LOGPROBS, the integer rating

---

[2] A 1–5 point scale was chosen to distinguish between degrees of plausibility as opposed to solely absolute direction. Several items in our dataset were distinguishable through this increased granularity, e.g., a 1 vs 2 or 4 vs 5 that would not have been distinguishable otherwise.
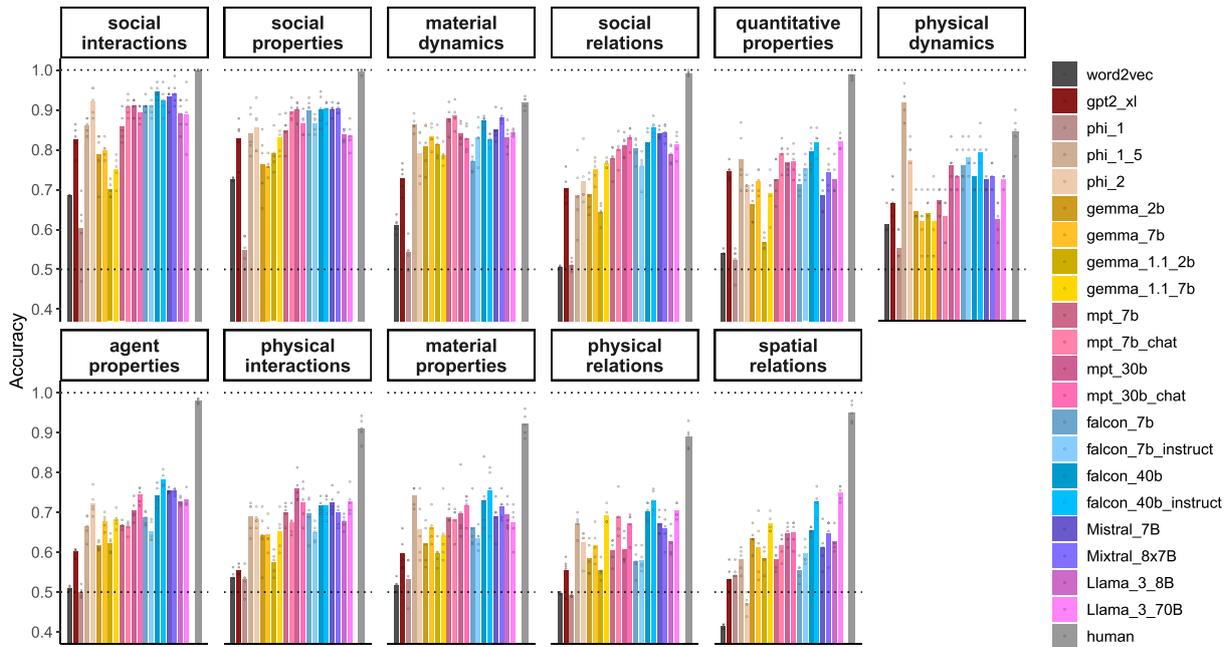
Figure 3: LLM performance across world knowledge domains (evaluated with LogProbs). Here and elsewhere, the dotted line at 0.5 denotes chance accuracy. Each dot reflects performance on a single *version* of EWoK-core-1.0 (see Section 3 ''Templates & Fillers'' and A.1), with the bar reflecting the mean 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.

for Likert, and correct context index selection for Choice. In all cases, models 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, which has a coarser integer scale, 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. Even a random coin flip or a deterministic model generating the same response for each query independent of context will trivially achieve this baseline.

For prompt-based evaluations (Likert and Choice), we used the same prompting setup across all models to assess them in the same way as humans; however, unlike humans, models were additionally provided with 2-shot examples. These examples did not come from any of our domains and were solely meant to outline the required response formatting including one positive and one negative example. See Appendix A.5 for more details. Our initial experiments revealed that supplying these examples substantially improves model performance, providing us with an opportunity to evaluate their ''best shot'' at EWoK-core-1.0.

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 LLMs are listed in Figure 3 as well as in Appendix A.3.

We additionally tested a baseline bag-of-words model based on word2vec embeddings (Mikolov et al., 2013). Embeddings for each word in a

1252

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). Word-embedding-based models provide useful baselines, and there is a long tradition of work in NLI and related tasks utilizing such easy-to-understand baselines to assess task difficulty and the presence of alternative solution strategies (Naik et al., 2018). Past NLI work has inadvertently fallen prey to design biases, making it easy for models to rely on heuristics, such as word overlap (McCoy et al., 2020). Additionally, word embeddings can contain rich semantic information (Grand et al., 2022) that may help solve the task by assessing semantic relatedness between words in the context and the target. Using bag-of-words baselines helps understand how robust the task is in the face of such heuristics.

4.3 Surface-level Item Properties

To determine the influence of surface-level item properties on model performance, we tested whether LLM performance correlates with the number of words in an item, as well as with average word frequency in an item. Word frequency was determined 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. The task human 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 Pearson's $R = 0.96$ correlation, and decided to drop the Choice condition for the full data collection; see Appendix A.2 (experimental design) & Appendix B.1 (results) for pilot details).

Data Collection To obtain reliable ratings across the full set of EWoK-core-1.0 items, we collected at least 5 responses per item from a total of $N = 1,262$ participants (591 female, 579 male, 27 other/unknown; median age 36 years; all US residents who reported English as their first language). Participants were recruited via Prolific

(https://prolific.co), an online study platform. To control for quality of data, we excluded 59 participants whose inter-subject Pearson correlation coefficient (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 filler-populated *versions*; see Section 3, ''Templates & Fillers'' and Section A.1) 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.5; 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 filler-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 rated an average of 57 items (all items came from a single domain). The resultant average inter-subject Pearson correlation across all items was 0.744, where the inter-subject correlation was calculated per-item based on all the participants that rated that item.

Aggregation and Evaluation We average ratings for a given C_i-T_j pair across all (typically, 5) participants who rated the pair. To obtain human norms, we compare the averaged Likert ratings of subparts, awarding 0.5 point for each half comparison, i.e., Likert$(C_1, T_1) >$ Likert(C_2, T_1) and Likert$(C_2, T_2) >$ Likert(C_1, T_2), similar to evaluating LLMs as described in Section 4.1. In case of a tie in scores for a half-comparison, we award 0 points.

5 Experiments: EWoK-core-1.0

EWoK-core-1.0 Can be Challenging for LLMs Although EWoK-core-1.0 was not designed to be a difficult dataset, we found that even large open models perform well below human baselines: The best model tested, falcon-40b-instruct, yields a mean accuracy of 0.80 whereas mean human accuracy is 0.95 (Table A4). As expected, larger models tend to do better, although it is not a sole predictor of performance. Instruction tuning does not consistently increase or decrease LLM performance under the LogProbs metric.

Model	Mean Accuracy
GPT-4 Turbo (2024-04-09)	0.912
Claude 3.5 Sonnet (2024-06-20)	0.911
Qwen2 Instruct (72B)	0.901
Claude 3 Opus (2024-02-29)	0.893
GPT-4o (2024-05-13)	0.886
Llama 3 Instruct (70B)	0.876
GPT-4o mini (2024-07-18)	0.864
Claude 3 Sonnet (2024-02-29)	0.848
Mixtral Instruct (8×22B)	0.842
Mixtral Instruct (8×7B)	0.837
Claude 3 Haiku (2024-03-07)	0.829
Claude 3.5 Haiku (2024-10-22)	0.810
GPT-3.5 Turbo (0125)	0.797
Mistral Instruct v0.3 (7B)	0.772
Llama 3 Instruct (8B)	0.171

Table 2: Many instruction-tuned frontier LLMs perform well (but not at ceiling) in a binary choice prompt-based evaluation setting. Prompting was 2-shot (see Appendix A.5). The values reported here are averages over LLM performance on each domain. Note that EWoK-core-1.0 was made publicly available in May 2024, so models updated/released after that time may have already been exposed to it.

EWoK-core-1.0 was later evaluated on closed frontier models via HELM (Liang et al., 2023) in January of 2025. That evaluation was performed using Choice prompting given that LogProbs for closed models are not available. The highest performing closed models yield performance ~ 0.91, which is still below humans (~ 0.95) but not by much (see Table 2).

Performance Varies Drastically by Domain Figure 3 shows model results by domain, from easiest to hardest (see also Table A5). *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 performance gap with humans (who score highly at 0.96). We conclude that domains that tap into social knowledge are easier for LLMs whereas domains that require relational knowledge about the physical world are hardest.

Overall, domain difficulty is consistent across LLMs, such that the same domains are predominantly hard or easy for all LLMs. 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 (Table A4), 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 3). One of the strengths of EWoK is that it allows us to explicitly explore this variation in details of the dataset construction process. In most benchmarks, these types of details in item-level construction variation are not measurable, even though these effects likely exist.

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 4 shows that all these factors affect model performance, often in different ways than they affect humans. The number of words in an item negatively affects LLM but not human 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
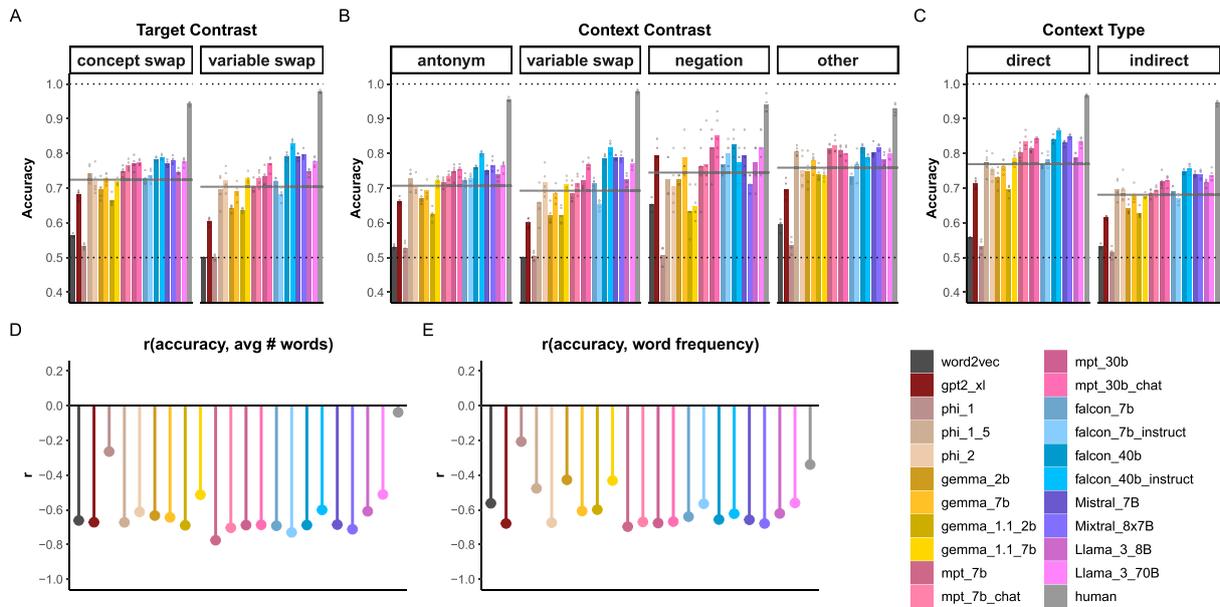
Figure 4: *Top:* LLM and human performance across target contrast (A) context contrast (B) and context type (C), evaluated with LogProbs. For examples of manipulations in A–C, see Figure 2. Dark gray line shows average model performance. *Bottom:* correlation between LLM accuracy and surface-level item features: (D) average item length and (E) average word frequency in the item. Humans are not sensitive or only weakly sensitive to these features, whereas model 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 S6).

two hardest domains for LLMs, have the highest word frequency. To evaluate the 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 A6), 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. These results indicate that surface-level factors, such as sentence length and word frequency, contribute but do not fully explain model performance.

Finally, the baseline bag-of-words model, which selects the context with the highest word2vec cosine similarity to the target, performed much worse than LLMs, indicating that simple embedding similarity between words in contexts and targets is insufficient to explain LLM performance. Of note, however, is that the phi-1 model does poorly on all domains and even worse than the baseline bag-of-words model. We believe the overly specific nature of phi-1's training data that consists of textbook-like demonstrations of

coding problems and solutions likely makes it unable to succeed on EWoK-core-1.0. This helps illustrate that without exposure to a naturalistic text corpus which would contain information about the world, even a model that is fluent in following English-language instructions is not able to flexibly reason about the world.

LogProbs Yields Higher Accuracy Than Prompting Evaluating LLMs with LogProbs resulted in above-chance accuracy for almost all open models (Figure 5). As expected, the gap between LogProbs and prompting was larger for smaller models (Hu and Frank, 2024). While we observed this general pattern of LogProbs as the dominant strategy, it is possible that highly targeted prompt engineering may result in above-LogProbs performance.

Human Ratings Are Usually, But Not Always, Accurate Finally, we examined discrepancies between human ratings and experimenter-designed ''ground truth'' 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*
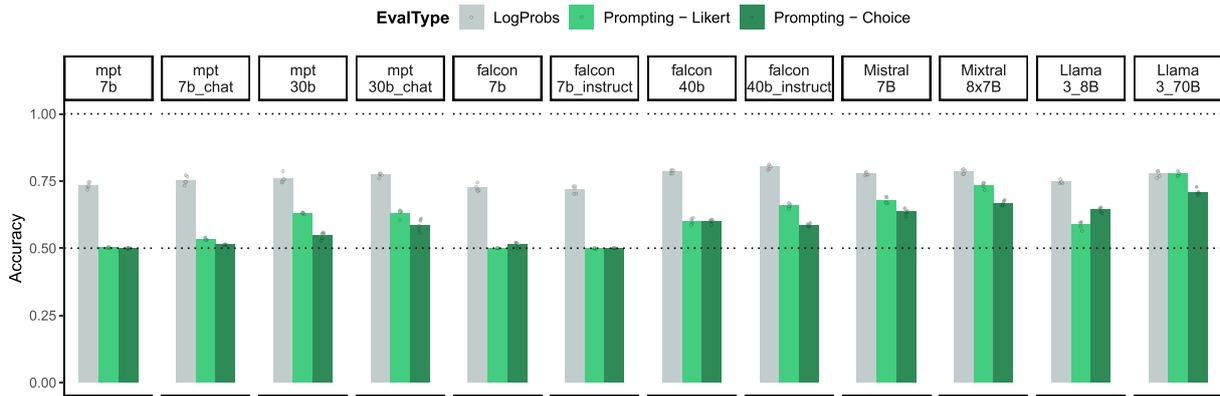
Figure 5: LLM performance assessed with LogProbs vs. two prompt-based tasks, Likert and Choice. Prompting is 2-shot, and the outputs are constrained to the set of allowed values ($1-5$ for Likert, 1 or 2 for Choice); this setup was chosen to maximize model performance. Still, LogProbs is a better strategy in nearly all cases. In prompting tasks, it was common for models to 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 7).

cooler.'' may be implausible because *the cooler* is a large object and *the car* has windows, although, for smaller objects and containers without windows, the scenario is more 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. Spatial cognition is known to be variable across individuals and cultures (Majid et al., 2004; Pitt et al., 2022), including preferential reliance on relative ('left'/'right') over absolute ('north'/'south') reference frames in Western cultures, which might explain why our (primarily Western) participants performed worse on items with absolute reference frames. Overall, human data collection is a valuable source of information on our dataset but it does not replace ground truth labels.

6   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) leverages a uniform item format to probe diverse domains of physical and social knowledge, (2) presents items that employ spe-

cific 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 large 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 (with social knowledge being the easiest and physical and spatial knowledge being the hardest).

We show that LLM-generated relative item probabilities estimated with LogProbs, allow distinguishing plausible and implausible scenarios above chance for most models and domains 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). Thus, we argue that a natural way to compare models is through the probabilities they output. This is especially the case for models before post-training/instruction-tuning, weak models, or checkpoints, all of which may not have the capability of following complex instructions (e.g., multiple choice) but have substantial knowledge that can be revealed with LogProbs.

We discovered that our dataset has high *domain distinguishability*—some world knowledge domains are much easier for LLMs than others—but low *model distinguishability*—many LLMs perform comparably across domains. This pattern indicates that the specifics of the model architecture and training data are less important than the semantic content being evaluated. Results from all models consistently show that social knowledge is easier to learn from text alone than knowledge about the physical world.

We also see that closed frontier models achieve high accuracy on EWoK-CORE-1.0. This is not surprising, as this dataset was not designed as a challenge benchmark but rather as a comprehensive, broad-coverage dataset of core world knowledge in humans. Thus, the broader framework will have value even when this specific version of the dataset gets saturated.

The EWoK framework opens up multiple avenues for future work:

Targeted Experiments The flexibility of our framework allows for 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 these specific concepts across several 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 reasoning, LLM performance could improve.

Finally, due to the synthetic nature of our dataset, some items may be atypical. We avoid blatant semantic violations by imposing type restrictions on template variables and validating our gold labels against human ratings, so the finalized set of items in EWoK-CORE-1.0 are semantically valid, just perhaps uncommon. There may be interest in using the EWoK framework to test semantically likely, not just semantically possible items (e.g., Hu et al., 2025). One way to achieve this would be to use LLMs to populate templates. We leave this approach to future work, noting that LLM-based item generation should be used with care to avoid confounds that inflate model performance (e.g., Panickssery et al., 2024a).

Another consequence of template-based generation is that items thus generated may not reflect the distributional statistics of natural language, and thus, log probability scores may not serve as an ideal linking function between the knowledge language models possess and the scenarios we test them on. Methods that explicitly identify ways to tap into LLMs' reasoning about scenario likelihood may be necessary (e.g., identifying relevant directions in activation-space: Panickssery et al., 2024b). Although we can only explore a limited set of evaluations in a single paper, the EWoK framework is evaluation-method-agnostic, and promises to continue being a useful resource allowing researchers to test for world knowledge in LLMs using novel methods that may be developed in the future.

7 Release Considerations

Our release-related goals are to (a) reduce the chances of accidental incorporation of EWoK items into LLM 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. `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. `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. `https://github.com/ewok-core/ewok-paper`.

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.

Acknowledgments

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. `https://doi.org/10.18653/v1/2021.conll-1.9`

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

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, Daniele Mazzotta, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. The Falcon series of open language models. *arXiv preprint arXiv:2311.16867. Version 2.*

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. `https://doi.org/10.1073/pnas.1306572110`, PubMed: 24145417

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. `https://doi.org/10.18653/v1/2020.acl-main.463`

Yonatan Bisk, Rowan Zellers, Jianfeng Gao, and Yejin Choi. 2020. PIQA: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7432–7439. https://doi.org/10.1609/aaai.v34i05.6239

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. https://doi.org/10.18653/v1/D15-1075

Stephanie M. Carlson, Melissa A. Koenig, and Madeline B. Harms. 2013. Theory of mind. *Wiley Interdisciplinary Reviews: Cognitive Science*, 4(4):391–402. https://doi.org/10.1002/wcs.1232, PubMed: 26304226

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. https://doi.org/10.18653/v1/D18-1269

Ido Dagan, Bill Dolan, Bernardo Magnini, and Dan Roth. 2010. Recognizing textual entailment: Rational, evaluation and approaches. *Journal of Natural Language Engineering*, 4. https://doi.org/10.1017/S1351324909990234

Stanislas Dehaene. 2011. *The Number Sense: How the Mind Creates Mathematics*. Oxford University Press 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. https://doi.org/10.18653/v1/2020.acl-demos.10

Team Gemma, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikula, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L. Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vlad Feinberg, Wojciech Stokowiec, Yu hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Warkentin, Ludovic Peran, Minh Giang, Clément Farabet, Oriol Vinyals, Jeff Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295. Version 4*.

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. https://doi.org/10.1145/2509558.2509563

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.

https://doi.org/10.1038/s41562-022-01316-8, PubMed: 35422527

Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Harkirat Singh Behl, Xin Wang, Sébastien Bubeck, Ronen Eldan, Adam Tauman Kalai, Yin Tat Lee, and Yuanzhi Li. 2023. Textbooks are all you need. *arXiv preprint arXiv:2306.11644. Version 2.*

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. https://doi.org/10.18653/v1/N18-2017

David Ha and Jürgen Schmidhuber. 2018. Recurrent world models facilitate policy evolution. In *Advances in Neural Information Processing Systems*, volume 31.

Alon Hafri and Chaz Firestone. 2021. The perception of relations. *Trends in Cognitive Sciences*, 25(6):475–492. https://doi.org/10.1016/j.tics.2021.01.006, PubMed: 33812770

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. https://doi.org/10.18653/v1/2023.emnlp-main.507

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. https://doi.org/10.18653/v1/2021.emnlp-main.564

Jennifer Hu and Michael Frank. 2024. Auxiliary task demands mask the capabilities of smaller language models. In *First Conference on Language Modeling*.

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. https://doi.org/10.18653/v1/2023.emnlp-main.306

Jennifer Hu, Kyle Mahowald, Gary Lupyan, Anna Ivanova, and Roger Levy. 2024. Language models align with human judgments on key grammatical constructions. *Proceedings of the National Academy of Sciences*, 121(36):e2400917121. https://doi.org/10.1073/pnas.2400917121, PubMed: 39186652

Jennifer Hu, Felix Sosa, and Tomer D. Ullman. 2025. Shades of zero: Distinguishing impossibility from inconceivability. *Journal of Memory and Language*, 143:104640. https://doi.org/10.1016/j.jml.2025.104640

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, USA. https://doi.org/10.1093/acprof:oso/9780198270126.001.0001

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. *arXiv preprint arXiv:2310.06825. Version 1.*

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. *arXiv preprint arXiv:2401.04088. Version 1.*

Carina Kauf, Emmanuele Chersoni, Alessandro Lenci, Evelina Fedorenko, and Anna A. Ivanova. 2024. Log probabilities are a reliable estimate of semantic plausibility in base and instruction-tuned language models. In *Proceedings of the 7th BlackBoxNLP Workshop @ EMNLP 2024.* `https://doi.org/10.18653/v1/2024.blackboxnlp-1.18`

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. `https://doi.org/10.1111/cogs.13386`, PubMed: 38009752

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. `https://doi.org/10.18653/v1/D18-1546`

Michal Kosinski. 2024. Evaluating large language models in theory of mind tasks. *Proceedings of the National Academy of Sciences*, 121(45):e2405460121. `https://doi.org/10.1073/pnas.2405460121`, PubMed: 39471222

Andrew Lampinen. 2024. Can language models handle recursively nested grammatical structures? A case study on comparing models and humans. *Computational Linguistics*, 50(4):1441–1476. `https://doi.org/10.1162/coli_a_00525`

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. `https://doi.org/10.1073/pnas.1910148116`, PubMed: 31488726

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. `https://doi.org/10.18653/v1/2021.emnlp-demo.21`

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. Version 1.*

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 Cosgrove, Christopher D. Manning, Christopher Re, Diana Acosta-Navas, Drew A. 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.*

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. *PsyArXiv: u6xdz.* https://doi.org/10.31234/osf.io/u6xdz

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.

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. https://doi.org/10.18653/v1/W17-2810

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. https://doi.org/10.1016/j.tics.2024.01.011, PubMed: 38508911

Asifa Majid, Melissa Bowerman, Sotaro Kita, Daniel B. M. Haun, and Stephen C. Levinson. 2004. Can language restructure cognition? The case for space. *Trends in Cognitive Sciences*, 8(3):108–114. https://doi.org/10.1016/j.tics.2004.01.003, PubMed: 15301750

Michael McCloskey. 1983. Intuitive physics. *Scientific American*, 248(4):122–131. https://doi.org/10.1038/scientificamerican0483-122

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. https://doi.org/10.18653/v1/P19-1334

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, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden. 2011. Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014):176–182. https://doi.org/10.1126/science.1199644, PubMed: 21163965

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. https://doi.org/10.18653/v1/2023.eacl-main.213

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

Aakanksha Naik, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, and Graham Neubig. 2018. Stress test evaluation for natural language inference. In *Proceedings of the 27th*

International Conference on Computational Linguistics, pages 2340–2353.

Arjun Panickssery, Samuel Bowman, and Shi Feng. 2024a. LLM evaluators recognize and favor their own generations. *Advances in Neural Information Processing Systems*, 37:68772–68802.

Nina Panickssery, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. 2024b. Steering Llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15504–15522. https://doi.org/10.18653/v1/2024.acl-long.828

Roma Patel and Ellie Pavlick. 2021. Mapping language models to grounded conceptual spaces. In *Proceedings of the International Conference on Learning Representations*.

Ellie Pavlick. 2022. Semantic structure in deep learning. *Annual Review of Linguistics*, 8:447–471. https://doi.org/10.1146/annurev-linguistics-031120-122924

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. https://doi.org/10.18653/v1/2021.starsem-1.1

Benjamin Pitt, Alexandra Carstensen, Isabelle Boni, Steven T. Piantadosi, and Edward Gibson. 2022. Different reference frames on different axes: Space and language in indigenous Amazonians. *Science Advances*, 8(47):eabp9814. https://doi.org/10.1126/sciadv.abp9814, PubMed: 36427312

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. https://doi.org/10.18653/v1/S18-2023

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. https://doi.org/10.1038/s42256-019-0132-2

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. https://doi.org/10.1145/3474381

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. https://doi.org/10.18653/v1/D19-1454

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. https://doi.org/10.1073/pnas.2307876121, PubMed: 38422017

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. https://doi.org/10.18653/v1/2020.coling-main.605

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. https://doi.org/10.1073/pnas.2200800119, PubMed: 36251997

Elizabeth S. Spelke and Katherine D. Kinzler. 2007. Core knowledge. *Developmental Science*, 10(1):89–96. https://doi.org/10

.1111/j.1467-7687.2007.00569.x, PubMed: 17181705

Ashley J. Thomas. 2024. Cognitive representations of social relationships and their developmental origins. *Behavioral and Brain Sciences*, pages 1–53. https://doi.org/10.1017/S0140525X24001328, https://doi.org/10.31234/osf.io/xhrfu

Tomer Ullman. 2023. Large language models fail on trivial alterations to theory-of-mind tasks. *arXiv preprint arXiv:2302.08399. Version 5.*

Akira Utsumi. 2020. Exploring what is encoded in distributional word vectors: A neurobiologically motivated analysis. *Cognitive Science*, 44(6):e12844. https://doi.org/10.1111/cogs.12844, PubMed: 32458523

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. https://doi.org/10.1162/tacl_a_00321

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 preprint arXiv:2307.09702. Version 4.*

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. https://doi.org/10.18653/v1/N18-1101

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. *EMNLP 2020*, page 38. https://doi.org/10.18653/v1/2020.emnlp-demos.6

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. Version 2.*

Ilker Yildirim and L. A. Paul. 2024. From task structures to world models: What do LLMs know? *Trends in Cognitive Sciences*. https://doi.org/10.1016/j.tics.2024.02.008, PubMed: 38443199

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. https://doi.org/10.18653/v1/P19-1472

A Methods

A.1 Dataset Generation

EWoK-core-1.0 is the result of 5 randomly sampled filler-substituted versions of the templates in EWoK. Each *version* in the script below also serves as a random seed for the dataset. Each dataset includes 1 copy of each template, as specified by the num_fillers parameter (i.e., one unique set of fillers will be sampled for this version). The parameter fix_fillers makes it so that fillers are shared across templates within this version (i.e., agent1 is always filled into by *Fatima*, if *Fatima* is the first agent name sampled). Regardless of all of these parameters, variable constraints are always respected and additional fillers are sampled to satisfy constraints (e.g., can_bounce=true).

```bash
#!/bin/bash


# first, compile individual
# concepts and corresponding
# context/target chunks into
# templates that can be
# filled into with fillers


python -m ewok.compile \
    --compile_templates


# next, fill into templates using
# custom parameters to create a
# combined dataset


for version in {0..4}; do
    python -m ewok.compile \
    --compile_dataset=true \
    --fix_fillers=true \
    --num_fillers=1 \
    --version="$version" \
    --custom_{i}d="ewok-core-1.0"
```

The full list of concepts used in EWoK-core-1.0 templates is provided in Table 3.

A.2 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 the 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 from a total of N=1,262 participants (591 female, 579 male, 27 other/unknown; median age 36; all US residents who reported English as their first language). Details of the main study are reported in Section 4.4, therefore not repeated here.

Data Exclusion Some online participants may respond randomly and therefore need to be excluded. To evaluate the quality of individual participants' responses, we computed Pearson R between that participant's responses and an average of other responses for that domain. Participants whose correlations were <0.3 were excluded.

Validating Gold Dataset Labels using Human Data We used human data to validate author-assigned gold labels for items in EWoK-core-1.0. First, we identified all items where the majority human label did not match the gold label (e.g., the plausibility of $C1$–$T2$ was higher than $C2$-$T2$). Then, we manually inspected those

Domain	Concepts
Social Interactions	help, deceive, hinder, cooperate, compete, evade, seek, chase, learn, teach, respect, insult, flirt, comfort, tease, coerce
Social Properties	friendly, hostile, trustworthy, untrustworthy, tolerant, bigoted, boastful, humble, dominant, submissive, shy, extroverted, introverted, confident, warm, cold
Social Relations	friend, stranger, enemy, boss, subordinate, colleague, romantic partner, teacher, student, landlord, roommate, tenant, parent, sibling, child
Physical Interactions	heat, cool, lift, drop, attract, repel, throw, catch, climb on, push, pull, break, fix, collide with, revolve around, approach, kick, touch, sit on, stand on
Physical Dynamics	roll, slide, fall, rise, sink, float, grow, shrink, oscillate, spin, accelerate, slow down
Physical Relations	bigger, smaller, occlude, occluded, contain, inside, outside, support, supported, attached, touching, connected, block, trail, hang, tied, on, under, surround, cover
Material Dynamics	fold, ripple, pour, stir, flap, splash, droop, drip, hang, pile, trickle, disperse, tap, compress, drape, squeeze, break, rip, wrinkle
Material Properties	transparent, opaque, smooth, rough, heavy, light, glossy, matte, fragile, sturdy, cold, warm, soft, hard, elastic, inelastic, bouncy, not bouncy
Agent Properties	see, not see, hear, not hear, smell, not smell, taste, not taste, feel, not feel, believe, doubt, imagine, prefer, have no preference, intend, not intend, make an effort, make no effort, make more effort, make less effort, choose
Quantitative Properties	more, less, fewer, the same, different, a lot of, a little, none, some, all, enough, not enough, many, few, no, the most, the least, the fewest
Spatial Relations	left, right, above, below, north, south, east, west, in front of, behind, close, far, toward, away

Table 3: All concepts tested in EWoK-core-1.0 grouped by domain.

items to determine whether we could trace the discrepancy to an author error or to an error in the item generation pipeline. When possible, the generation pipeline was adjusted to generate the correct version of the item; alternatively, the faulty item was excluded from the dataset.

A.3 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-8×7b-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., 2020), and all experiments were run on a $4 \times A100$ 80GB GPU cluster.

A.4 Text Completion Experiments

For the prompting-based evaluations, Choice and Likert, we support two different generation op-tions: *free* and *constrained*. For free genera-tion, 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 gen-eration, the LLM may greedily sample from a restricted set of tokens, either 1–2 or 1–5, con-strained using logit masking (Willard and Louf, 2023). Such a strategy enforces well-structured responses, but requires a restricted response for-mat. In addition to variation in generation options, we support both zero- and few-shot prompting. In Figure 5, the prompting results we report use 2-shot constrained generation, as it yields the highest performance among our space of tested strategies.

A.5 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.6 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 6 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 (Figure 6). LIKERT-scale judgments are less dependent on the specific set up (making a CHOICE judgment 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.
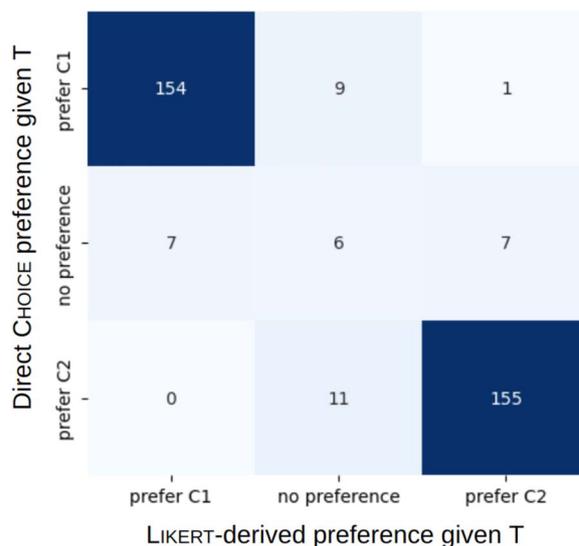

Figure 6: Confusion matrix showing agreement between a direct choice of $\{C_1, C_2\}|T_i$ and an indirect choice based on individual LIKERT ratings $\{C_i|T_i\} > \{C_j|T_i\}$. In this comparison we marked items as "no preference" if (1) the Likert scores weren't at least 1 apart, or (2) the average choice preference wasn't at least 0.2 away from the midpoint between two alternatives. However, we do not use a "no preference" bucket for the main study.

Full Study See the main text Sections 4.4 & 5 for results about the main study (number of participants, exclusions made, inter-subject correlation, and performance of humans on the task).

B.2 Model Evaluation

1268

Model	Mean LogProbs Accuracy	Range
human	**0.951**	**0.942–0.957**
word2vec	0.542	0.539–0.547
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_8×7B	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 4: Performance on EWoK-core-1.0. Range reported across 5 dataset versions.

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 5: LLM and human performance by domain.

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

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_8×7B	0.511	0.471–0.533
Llama_3_8B	0.195	0.145–0.222
Llama_3_70B	**0.588**	**0.576–0.603**

Table 7: LLM LIKERT accuracy on EWOK-CORE-1.0 with stricter inequality metric.