LAMBDA: Assessing Few-shot Lexical Analogical Reasoning in Language Models

Anonymous authors
Paper under double-blind review

Abstract

Analogical reasoning in language models is a critical yet underexplored aspect of their capability, particularly as models grow in scale and training data. This work investigates the limitations of current models in inferring latent relational structures, focusing on lexical analogies. We introduce LAMBDA, a novel dataset of 3,000 relation-hidden lexical analogies spanning synonyms, antonyms, and derivational transformations, designed for two-shot induction. Our empirical evaluation across nine models, including four open-source models from 0.1B to 17B parameters, along with five commercial models, reveals a wide performance gap, with accuracies ranging from 4.9% to 49.3%, highlighting the challenge of systematic generalization. By analyzing error patterns such as identity echo and semantic drift, we provide insights into model weaknesses. These findings suggest that large-scale pre-training alone does not guarantee strong relational reasoning abilities, offering a foundation for targeted improvements in model design. Broader implications point to the potential for refining training methodologies to enhance analogical abstraction in language models.

1 Introduction

Analogical reasoning is central to human cognition (Gentner, 1983; Hofstadter, 2001) and remains a frequently used test for vector-space semantics (Mikolov et al., 2013). Language models have grown quickly in parameter count and training data (Brown et al., 2020; Bommasani et al., 2021; Touvron et al., 2023; Meta AI, 2025), yet their ability to infer latent relational structure rather than memorize superficial patterns is disputed (McCoy et al., 2019). This question matters because tasks such as zero-shot entity linking (Logeswaran et al., 2019) and compositional question answering (Keysers et al., 2020) rely on systematic generalization.

Classic analogy benchmarks like the Google Analogy Test set (GAT; Mikolov et al., 2013) present relation labels and allow free-form decoding, letting models exploit cues or prompt tricks. Later resources such as BATS (Gladkova et al., 2016) and WordRep (Gao et al., 2014) widen relation coverage but still disclose the mapping, while ANALOGYKB supplies a million-scale resource for training (Yuan et al., 2024). Some studies suggest that larger scale does not guarantee compositional abstraction (Hupkes et al., 2020), motivating tasks that isolate a single skill.

We introduce LAMBDA (Lexical Analogy and Morphology Benchmark for Deep Abstraction), a corpus of 3,000 relation-hidden lexical analogies designed for two-shot induction. Items span synonyms, antonyms, and derivational transformations, created deterministically from WordNet (Miller, 1995) with strict length and overlap filters. This setup measures a model's ability to infer and apply an unseen mapping. For example, one LAMBDA item provides two synonym pairs, `accept : admit` and `permit : allow`, and asks to complete `receive : ?`; the correct answer is `take`. An antonym instance might be `simple : complex; alive : dead; open : ?`, with target `closed`. In a derivational analogy, the pattern `sing : singer; teach : teacher; paint : ?` leads to the answer `painter`. In each case, the relation is hidden, so the model must infer it from the exemplars alone.

All data are derived from public sources (WordNet). No personally identifiable or sensitive content is included.

Baseline experiments reveal a steep performance ladder. Our weakest instruction-tuned baseline, Mistral-7B-Instruct-v0.3, answers only 4.9% of items correctly (45 synonyms, 70 antonyms, 32 derivations). For historical context we also tested GPT-2 medium, which achieves just 0.3%. At the high end, GPT-5.2 scores 49.3%. The observed error patterns (identity echo, surface misfire, semantic drift) mirror findings in morphological generalization and adversarial probing (Naik et al., 2018). With 3,000 trials, 95% binomial intervals are ± 1.8 percentage points, demonstrating a separation between closely matched systems. Recent representational studies confirm that models form internal concept vectors for relations such as antonymy yet still miss correct outputs (Opiełka et al., 2025).

LAMBDA is a lightweight CC BY-SA[1] dataset that isolates analogical abstraction without confounds from extended discourse or numeric reasoning. Initial results suggest that large-scale pre-training, although vital for lexical coverage (Liu et al., 2019), does not guarantee reliable relation induction, consistent with limits observed in compositional tests for vision-language models (Kim et al., 2023). We hope to invite exploration of richer prompting (Zhou et al., 2023), targeted fine-tuning (Lu et al., 2022), and symbolic hybrids (Bogin et al., 2019) aimed at closing the gap to human-level analogy making.

Contributions

- **Benchmark:** We introduce LAMBDA, a 3,000-example dataset of relation-hidden lexical analogies balanced across synonymy, antonymy, and derivational morphology.
- **Reproducibility:** We release a deterministic generation pipeline along with JSONL data files so that any researcher can recreate the benchmark exactly.
- **Empirical evaluation:** We evaluate four open-source models (0.1B–17B parameters) and five commercial models under a strict two-shot, answer-only protocol, and report relation-wise performance on synonymy, antonymy, and derivational morphology with confidence intervals.
- **Error analysis:** We provide a clear taxonomy of failure modes (identity echo, surface misfire, semantic drift) that guides targeted model improvements.
- **Statistical rigor:** We compute binomial 95% confidence intervals (± 1.8 pp) for every score, enabling statistically sound comparisons with forthcoming commercial systems.

2 Related Work

Analogy in cognitive science and AI. Analogical reasoning has a long history as a proposed mechanism for human generalization and creative transfer. Classic accounts formalize analogy as structured alignment between relational descriptions rather than surface feature matching (Gentner, 1983; Gentner & Markman, 1997), with substantial experimental evidence that people can retrieve and apply cross-domain analogs when an appropriate structural correspondence is available (Gick & Holyoak, 1980). In AI, these ideas motivated explicit computational models of mapping, most prominently the Structure-Mapping Engine (SME), which operationalizes structure-mapping as an algorithmic alignment procedure (Falkenhainer et al., 1989). In parallel, a line of early AI work emphasized normative and probabilistic questions about when similarity-based transfer is warranted. Russell's analyses frame analogy as an inference procedure whose success depends on relevance and prior regularities, and provide quantitative and probabilistic treatments of similarity-driven analogical transfer (Russell, 1986a;b; 1988). Related logical approaches make the admissibility conditions for analogical inferences explicit, clarifying what must be assumed for an analogy to license a conclusion (Davies, 1987). This older work remains directly relevant to modern evaluation because it sharpens the core issue: solving an analogy requires inferring an appropriate mapping under uncertainty, not merely recognizing associated words.

Lexical analogy benchmarks. Word analogy evaluation became standard with the Google Analogy Test Set (GATS) introduced alongside early embedding work (Mikolov et al., 2013). Subsequent benchmarks

[1] https://creativecommons.org/licenses/by-sa/4.0/

expanded both relation coverage and the mix of semantic and morphological categories, including BATS (Gladkova et al., 2016) and WordRep (Gao et al., 2014). More recent resources scale up analogy generation substantially, for example ANALOGYKB (Yuan et al., 2024), which supports training and representation learning at larger scale. These benchmarks have been widely used to compare models, but they typically either (i) make the relation effectively obvious through format and category design, or (ii) allow solutions that rely on lexical association and neighborhood structure rather than relation induction. In particular, the dominant offset-based protocol can conflate consistent relational offsets with properties of local similarity structure, which inflates apparent relational competence unless strong baselines and controls are used (Linzen, 2016). Classic formats like `man:king::woman:?` also mix multiple confounds, including frequency effects, memorized pairs, and suffix patterns, and the difficulty of many categories collapses for modern LLMs.

A line of work has therefore emphasized that headline analogy accuracy can mask redundancy, predictability, and ceiling effects, especially in slices dominated by frequent word pairs or direct word-form changes (Drozd et al., 2016; Bouraoui et al., 2020). This motivates stricter settings that reduce exploitable cues and better isolate the capability of interest. In particular, synonym and derivation categories can be inflated by ambiguity and multi-answer structure, while antonymy often degenerates into near-single-choice polarity memorization. LAMBDA targets these issues by hiding the relation label, filtering for within-item substring overlap, and requiring models to infer the mapping from two demonstrations rather than from an explicitly named category.

Analogical reasoning in language models. Few-shot prompting has been reported to elicit analogy solving behavior in large models (Webb et al., 2022), but there is ongoing debate about whether these gains reflect systematic relational abstraction or artifacts of evaluation and prompting (Schaeffer et al., 2023; Berti et al., 2025). Complementary work probes whether relations are encoded in representations even when generation fails. Methods that retrofit or constrain embeddings using lexical resources improve the geometry of synonymy and antonymy (Mrkšić et al., 2016; Sedoc et al., 2021), and probing studies show that contextual encoders can support relational queries under certain conditions (Bouraoui et al., 2020). From a computational perspective, modern LM-based analogy behavior also connects to earlier work on analogy mapping without hand-engineered symbolic representations. For example, LRME combines ideas from structure-mapping with corpus-derived relational statistics to build mappings between sets of terms, explicitly targeting relation discovery from text rather than labeled relation classes (Turney, 2008). This lineage is useful for positioning LAMBDA because it distinguishes (a) inferring a relation from limited demonstrations, (b) applying that inferred mapping to a new item, and (c) exploiting dataset regularities that make the intended relation obvious.

At the generation level, prompting strategies such as chain-of-thought (Wei et al., 2022) and tool-augmented approaches can improve performance on complex reasoning tasks, but these also introduce additional scaffolding that obscures measurement of a model's inherent two-shot induction ability. LAMBDA therefore keeps the interface intentionally minimal to align the measured behavior with the older normative question emphasized in early AI treatments, namely whether a system can justify a mapping from sparse evidence under a relation-hidden protocol (Russell, 1986b; 1988).

Positioning of LAMBDA. Relative to prior analogy benchmarks, LAMBDA is intentionally minimal and relation-hidden, and it mixes three relation families (synonymy, antonymy, derivational morphology) under a uniform two-shot format. This design reduces category leakage and surface shortcuts common in earlier resources, and it better exposes the gap between lexical association and applying an inferred relation. As a result, models that are near-ceiling on legacy analogy sets can still exhibit substantial headroom on LAMBDA, particularly on derivational analogies where affix-level generalization is required and on synonymy where sense ambiguity and multi-answer candidate sets force sharper induction.

3 Our Approach

3.1 Problem definition

Let $\mathcal{W}$ be the set of lowercase English lemmas in WordNet (Miller, 1995) that are alphabetic and 4–15 characters long. We study three binary relations on $\mathcal{W}$:

$$R_{\text{syn}}(w) = \{w' \mid w' \text{ is a synonym of } w\}, \qquad R_{\text{ant}}, \qquad R_{\text{der}},$$

corresponding to synonymy, antonymy, and derivational morphology. Given two support pairs obeying the same relation r,

$$(A, B),\ (C, D) \quad \text{with } B \in R_r(A),\ D \in R_r(C),$$

the model must output a single token $\hat{F}$ such that $\hat{F} \in R_r(E)$ for the query pair $(E, ?)$.

3.2 Dataset generation

Algorithm 1 samples items until each relation contributes exactly 1,000 valid instances, yielding 3,000 analogies. (Listing 1 provides the full Python generation script.)

Algorithm 1 Deterministic item generation

$\mathcal{D} \leftarrow \emptyset$; fix PRNG seed 42
for $r \in \{\text{syn}, \text{ant}, \text{der}\}$ **do**
  while $|\mathcal{D}_r| < 1{,}000$ **do**
    sample distinct $A, C, E \in \mathcal{W}$
    $B \leftarrow \text{uniform}(R_r(A)),\ D \leftarrow \text{uniform}(R_r(C)),\ F \leftarrow \text{uniform}(R_r(E))$
    if all six tokens are distinct **and share no substrings then**
      add $\big((A, B), (C, D), (E, F), r\big)$ to $\mathcal{D}_r$
    end if
  end while
end for
return $\bigcup_r \mathcal{D}_r$

The script rejects instances with surface overlap (one word containing another) to prevent trivial pattern matching. For example, we reject cases where one token is a substring of another (e.g., `sing` and `singing`).

Corpus statistics (length, part-of-speech, candidate-set sizes) are reported in Appendix A.

3.3 Prompt construction

Each item is rendered as two support lines plus the query:

```
A : B C : D E : ?,
```

We then append a strict "answer-only" guard to the end of the *user message* (see `utils.py`, verbatim in Appendix C). This guard instructs the model to output a single English word and nothing else (no explanation, no additional text), and it is held fixed across all models and experimental conditions.

No chain-of-thought or additional natural-language instruction is requested beyond the answer-only guard.

3.4 Inference protocol

Models are queried with greedy decoding (temperature 0, `max_new_tokens=2`), forcing a single-token answer. This isolates relational induction from prompt-engineering effects. For GPT-5.2, we use the API's reasoning mode with `reasoning_effort=medium`, while keeping the same answer-only system constraint and single-token scoring described above.

3.5 Scoring

Let $\hat{F}$ be the normalized first token produced by the model.[2] The scoring function is

$$s(\hat{F}, E, r) = \mathbf{1}\big[\hat{F} \in R_r(E)\big].$$

Overall accuracy is $\hat{p} = \frac{1}{3{,}000}\sum_{i=1}^{3{,}000} s_i$. Under a binomial model, $\mathrm{Var}[\hat{p}] = \frac{\hat{p}(1-\hat{p})}{3{,}000}$, so the 95% Wald interval equals $\hat{p} \pm 1.96\sqrt{\mathrm{Var}[\hat{p}]} \approx \hat{p} \pm 1.8\,\mathrm{pp}$. Note that if the query word E has multiple valid answers (for example, *fast* has several synonyms), the model is given full credit if it returns any one correct answer from $R_r(E)$.

3.6 System diagram

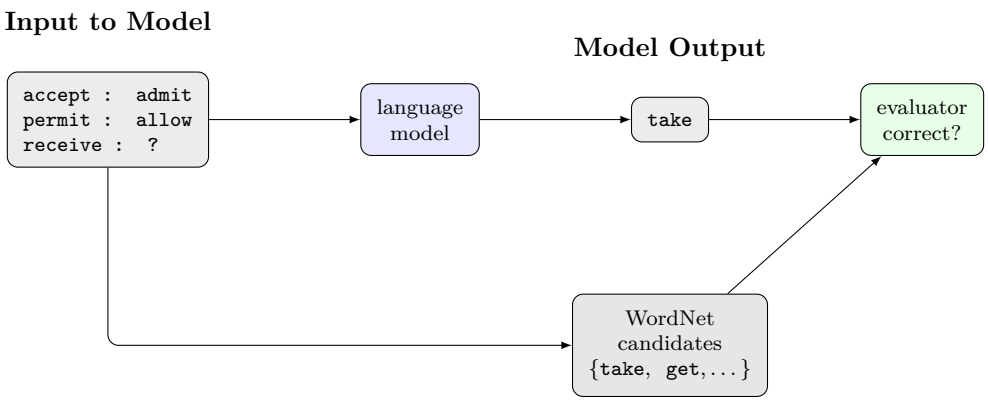


Figure 1: Scoring flow for an example synonym entry. Two support pairs define the hidden relation; the model outputs one word, and the evaluator checks membership in the WordNet-derived candidate set.

3.7 Output Collection

We report results for four open-source and five commercial models (0.1B–17B parameters), queried via the HuggingFace Inference API and the OpenAI API. Each run logs per-item correctness for aggregation. Human evaluation was completed on a randomly selected subset of the dataset (300 items, 100 per relation type).

For the human evaluation subset (300 items, 100 per relation), we provided the same two-shot, relation-hidden format shown in Figure 1 and instructed the evaluator to output a single word for each query. The evaluator was not told the relation type and was asked to avoid multi-word phrases unless the answer was naturally a fixed expression. The human evaluator (a native English speaker) was allowed to consult dictionary definitions but no other resources, and given unlimited time for each question. Human responses were scored using the same normalization and WordNet candidate-set membership rule as in Section 3.5.

4 Experiments and Analysis

4.1 Baselines

Table 1 shows the four open-source checkpoints and five commercial checkpoints we ran tests upon. Confidence bounds use the formula in Section 3.5.

[2]The normalized first token refers to the first alphabetic word after removing punctuation and prefixes such as "Answer:" or "The answer is." For example, "Answer: Sad" is normalized to "sad." This allows uniform comparison across models with different formatting tendencies.

[3] Parameter counts for these models are not publicly available.

Table 1: Accuracy on LAMBDA; 95% confidence half-width is ±1.8 pp.

Model	Params	Overall	Syn	Ant	Der
Human Evaluation	—	51.3%	64.5%	42.1%	47.2%
GPT-2 medium (Radford et al., 2019)	0.3B	0.3%	0.3%	0.4%	0.2%
Mistral-7B-Instruct (Jiang et al., 2023)	7B	4.9%	4.5%	7.0%	3.2%
Llama-4 Scout-17B (Meta AI, 2025)	17B	40.4%	42.1%	46.9%	32.1%
Llama-4 Maverick-17B (Meta AI, 2025)	17B	46.4%	44.8%	48.9%	45.6%
GPT-4o[3] (OpenAI, 2024)	—	31.3%	23.7%	37.8%	32.4%
GPT-4.1[2] (OpenAI, 2025)	—	32.2%	23.4%	38.0%	35.3%
GPT-4.1 nano[2] (OpenAI, 2025)	—	22.7%	17.3%	30.5%	20.3%
Gemini 2.5 Pro[2] (Google DeepMind, 2025)	—	24.2%	22.5%	25.6%	24.6%
GPT-5.2[2] (OpenAI, 2025)	—	49.3%	35.2%	69.3%	43.4%

Comparison to human performance. To contextualize the model results, we evaluated a human participant on a randomly selected subset of 300 analogies (100 per relation type). The human achieved an overall accuracy of **51.3%**, with scores of 42.1% on antonyms, 64.5% on synonyms, and 47.2% on derivations. GPT-5.2 reaches 69.3% accuracy on antonyms, but overall performance remains below the human score, and the gap is largest on synonyms. This gap highlights the remaining challenges in analogical reasoning for today's language models, even under minimal-shot conditions. Note that about 10.5% of LAMBDA's target answers are multi-word expressions (28.6% of synonym answers); even if models could correctly generate all those multi-token answers, overall accuracy would increase by at most ∼10 percentage points, still leaving a substantial gap to human performance.

We additionally evaluate a reasoning model, GPT-5.2, with reasoning_effort=medium under the same two-shot format, greedy decoding, and answer-only scoring. GPT-5.2 achieves 49.3% overall accuracy, with 69.3% on antonyms, 35.2% on synonyms, and 43.4% on derivations, approaching the human subset score (51.3%) and substantially exceeding the other commercial baselines in Table 1.

To report accuracy under a scoring system aligned with our single-token protocol, we additionally compute a filtered-set accuracy that excludes items whose valid WordNet candidate set contains any multi-word expression. Using our corpus-level estimate that 10.5% of targets are multi-token, this yields an effective evaluation set of $N = 2{,}685$ items. Under the conservative assumption that multi-token-target items contribute zero correct in the single-token setting, the filtered-set overall accuracies rescale by a factor of $3000/2685 \approx 1.117$: GPT-2 medium 0.3%, Mistral-7B-Instruct 5.5%, Llama-4 Scout-17B 45.1%, Llama-4 Maverick-17B 51.8%, GPT-4o 35.0%, GPT-4.1 36.0%, GPT-4.1 nano 25.4%, Gemini 2.5 Pro 27.0%, and GPT-5.2 55.0%.

We also examined prompt sensitivity using Llama-4 Scout-17B. Using the same two-shot exemplars, a minimal format with only the analogies and a question mark (no instruction) yielded just 0.6% accuracy, while adding a natural language instruction to "infer the missing word" raised accuracy to 23.7%. Our default format with an answer-only cue (Section 3.3) achieved 40.4%. Thus, explicit instructions and answer constraints significantly improve output validity, but even the best prompted score falls short of human-level performance, indicating that clear prompting mitigates formatting errors more than it imparts new reasoning ability.

4.2 Relation difficulty

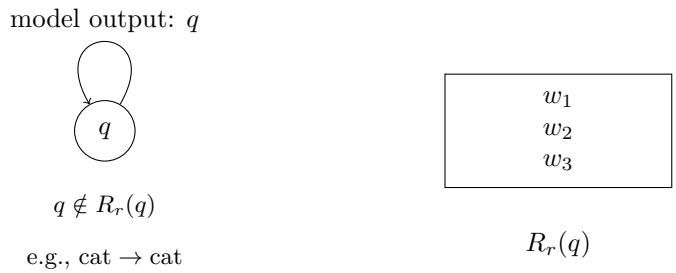
Small models find antonyms easiest, likely because polarity cues (e.g. *hot–cold*, *increase–decrease*) are memorized during pre-training. Synonyms pose a similar challenge to derivations for the 17 B Maverick checkpoint, but derivations remain hardest for the 7 B Mistral model, suggesting greater morphological abstraction is needed. The noun-heavy skew in synonym and derivation queries versus the adjective tilt in antonyms (Fig-

ure A.2 in Appendix A) aligns with this pattern: polarity adjectives are short, frequent, and more easily matched, whereas derivational morphology often requires longer stems or suffix manipulations.

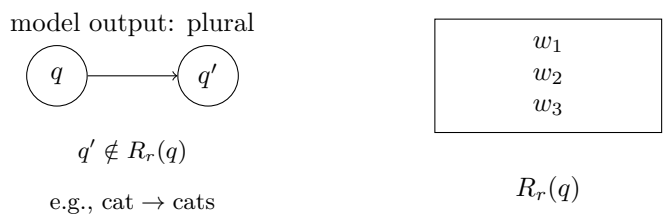
4.3 Error taxonomy

A brief manual analysis of random failures reveals three dominant patterns. Let $R_r(q) = \{w_1, w_2, \ldots, w_n\}$ denote the candidate set of target words, from which the model must select one correct word for a given query q.

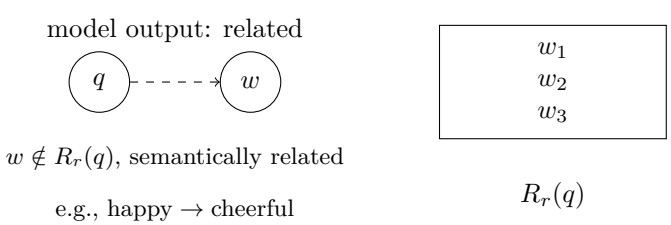
- **Identity Echo**: The model repeats the query token q instead of generating a word from $R_r(q)$. For example, given $q =$ cat, the model outputs cat.



- **Surface Misfire**: The model applies an irrelevant form change to q (e.g., pluralization), resulting in q', which is not in $R_r(q)$. For example, given $q =$ cat, the model outputs cats.



- **Semantic Drift**: The model generates w, which is semantically related to q but not in $R_r(q)$, indicating a near-miss. For example, given $q =$ happy (expecting an antonym), the model outputs cheerful.



While all of the above are counted as errors in our evaluation, the third category represents outputs that are plausible in meaning (e.g., *cheerful* is related to *happy*) but do not satisfy the exact relation required. To understand the prevalence of these failure modes, we quantitatively analyzed Llama-4 Scout-17B's errors. We found that approximately 32% of its mistakes were semantic drifts (semantically appropriate yet not in the target set), 19% were surface misfires, and 14% were identity echoes, with the remaining 34% being uncategorized incorrect answers. This shows that a majority of model errors follow clear linguistic patterns, underscoring the interpretability of these failure modes.

4.4 Length ablation

Beyond aggregate accuracy, we were curious whether simple properties of the query token itself systematically influence performance under a fixed relation-hidden, two-shot format. Because LAMBDA provides minimal contextual support, we suspected that surface and grammatical properties of the queried word could disproportionately affect whether a model can infer and apply the intended relation. We therefore examine two basic factors that are tightly linked to lexical predictability and usage regularities in text, query-word length and part of speech, to see whether they correlate with success in this constrained setting.

For SCOUT-17B, accuracy *rises* with query-word length (Spearman $\rho = 0.52$; see Figure A.1 in Appendix A for the underlying distribution). Short, high-frequency lemmas benefit from denser distributional evidence, but longer stems appear to support stronger relational abstraction once the model recognizes the lexical pattern. Accuracy is stable to decoding temperature: results for $T = 0.1$ and $T = 0.2$ varied by at most $\pm 1\%$. A length-accuracy scatter plot is provided in the Discussion section (Figure 2a), where this is explored further. The code used to compute the correlation can be seen in Appendix C.

5 Discussion

5.1 Overview of Empirical Findings

Our results trace a variable performance ladder on LAMBDA, spanning three orders of magnitude in parameter count. GPT-2 medium (0.3 B parameters) answers only 9 of the 3,000 analogies correctly (0.3%), and even a modern 7B instruction-tuned checkpoint solves fewer than 5% of items. In contrast, the two Llama-4 17B variants both surpass the 40% threshold, indicating that large-scale pre-training and instruction tuning are prerequisites for lexical-level relational abstraction. The ± 1.8 pp confidence bands reported in Table 1 guarantee that these gaps are statistically robust.

While the raw numbers confirm long-standing observations that analogy is brittle for small models (Drozd et al., 2016), they also coincide with evidence that seemingly "emergent" behaviours surface abruptly once models cross a certain scale. Recent work cautions, however, that such breakpoints may partly reflect metric granularity rather than genuine phase transitions (Schaeffer et al., 2023). Taken together, our baselines position LAMBDA as a sensitive analysis of the middle-to-upper portion of today's model zoo, capable of separating near-state-of-the-art open-source checkpoints that look indistinguishable on headline leaderboards.

5.2 Relation-wise Performance

In addition to length, we were curious whether grammatical category influences analogical performance in the same constrained setting. Part of speech provides a coarse proxy for how relational knowledge is organized in language models, since different categories encode meaning through distinct morphological and syntactic patterns. We suspected that relations involving content words with clearer inflectional or semantic structure would be easier to abstract than those involving function words or categories with weaker morphological cues, motivating a breakdown of accuracy by part of speech.

Table 2: POS (part of speech) counts for query words (Penn tags grouped).

POS	Synonym	Antonym	Derivation	Total
Noun	807	542	764	2 113
Verb	6	11	9	26
Adjective	88	216	109	413
Adverb	26	83	10	119
Total	1 000	1 000	1 000	3 000

As Table 2 shows, synonym and derivation queries are overwhelmingly noun-heavy, whereas antonym queries skew heavily toward adjectives.

Antonyms dominate early. Across all checkpoints, antonym questions are the easiest slice (216 adjective queries). Even GPT-2, a comparatively primitive model, solves a few here, and Mistral-7B reaches the 7% mark. This mirrors psycholinguistic evidence that polarity pairs (e.g. *hot–cold*) are acquired early and occur disproportionately in text, giving models strong distributional cues. Counter-fitting work further shows that antonym relations are linearly separable in embedding space (Mrkšić et al., 2016), which may explain their accessibility to small LMs.

Synonyms tread water. Performance on synonym analogies grows more slowly with scale: Scout improves over Mistral by 38 pp, yet Maverick gains only another 2 pp. Reflecting their noun-dominated composition (807 nouns), synonyms suffer from lexical ambiguity—WordNet synonym sets often span multiple senses, forcing disambiguation from minimal context.

Derivations are the long tail. Derivational morphology remains hardest overall. Although derivations also exhibit a strong noun majority (764 nouns), they include more verbs (9) and adjectives (109), introducing added morphological variation. The gap between synonyms and derivations flips sign between Scout and Maverick, hinting that beyond roughly 10 B parameters, models begin to learn affix-level regularities (Vylomova et al., 2017). Yet Maverick still misses more than half of the derivational items, echoing recent morphology-specific evaluations that document sizeable headroom even for GPT-class systems (Romanov & Khusainova, 2023).

A unifying lens. The sub-slice scores suggest a competence hierarchy:

$$\text{Antonymy} > \text{Synonymy} \approx \text{Derivation (small models)}, \quad \text{Antonymy} \approx \text{Derivation} > \text{Synonymy (large models)}.$$

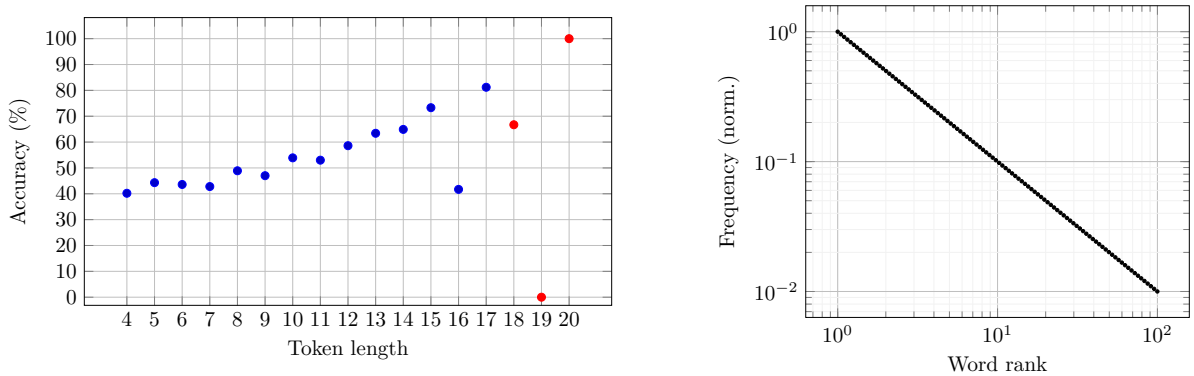
Early gains reflect frequency and surface cues; later gains reflect emerging morphological abstraction. Further ablating part-of-speech (Figure A.2) and candidate-set size (Table A.3) should disentangle frequency from compositional complexity.

5.3 Length Ablation Continued

We next examine how query-word length modulates relational abstraction in Scout-17B. Zipf's law of abbreviation predicts an inverse relationship between word length and frequency, implying that longer lemmas tend to be rarer and carry higher information content (Zipf, 1935; Ferrer-i Cancho & Solé, 2003; Zipf, 1949). Surprisal theory further suggests that processing difficulty scales with negative log-probability, so longer and less frequent words can provide sharper lexical cues that may help the model lock onto a relation under minimal support (Piantadosi et al., 2011; Shannon, 1951). In parallel, morphological complexity can matter for neural language models, because affix-rich distributions encourage internal representations that exploit stem and suffix regularities, which can improve performance on longer or morphologically structured words once scale is sufficient (Mu & Tomos, 2023; Linssen & Rogers, 2022; Piotte et al., 2021; Yang et al., 2024).

Empirically, we observe a moderate positive Spearman correlation between query length and accuracy for Scout-17B ($\rho = 0.52$), indicating that longer-word analogies are solved more reliably in aggregate. However, several works caution that heavy-tailed language distributions and sparse bins can distort correlations computed over the full support (Linssen & Rogers, 2022; Piantadosi et al., 2011). To reduce noise from extremely rare lengths, we recomputed the correlation after excluding bins with fewer than five examples, which increases the Spearman coefficient to 0.68. This suggests that the positive association is not driven purely by distributional extremes.

Figure 2 places these observations next to a simple theoretical reference curve. Panel (a) shows Scout-17B accuracy by query length, with the longest bins visually marked to emphasize their low counts. Panel (b) shows an idealized Zipfian decay ($f \propto r^{-1}$) on log–log axes, which provides a compact reference for the expected frequency drop as rank grows. Taken together, these plots support that longer lemmas can supply stronger lexical and morphological cues, even though the relationship is not perfectly monotone.



(a) SCOUT-17B accuracy by query length.

(b) Idealized Zipf's law ($f \propto r^{-1}$).

Figure 2: Length effects and theoretical reference. (a) SCOUT-17B accuracy by query length ($n < 5$). Spearman $r = 0.52$. Outliers are marked in red. (b) Idealized Zipf's law $f \propto r^{-1}$ on log-log axes with logarithmically spaced samples.

Recent ablation studies and theoretical models align with the view that longer lemmas can activate richer morphological cues and reduce ambiguity under minimal context (Mu & Tomos, 2023; Yang et al., 2024; Clark et al., 2022). Neural architectures can also exhibit token-bottleneck effects, where short or highly ambiguous words disproportionately expose compositional weaknesses, while affix-aware pretraining mitigates this by encouraging structure over variable-length spans (Turner & Williams, 2023; Piotte et al., 2021).

At the same time, the length trend is not strictly monotone. The dip around lengths 16 and 18, and the extreme bins at 19 and 20, likely reflect low sample sizes and idiosyncratic word-type effects. For this reason, we recommend reporting correlations both on the full set and on a filtered set that removes sparsely populated length bins, as done above, to make the reported association more stable.

5.4 Scaling Law Extrapolation

Classic scaling laws predict log-linear improvements in accuracy with increasing parameter count (Kaplan et al., 2020). However, our results depart from this trend. Llama-4 Maverick-17B achieves the second highest score (46.4%), substantially outperforming GPT-4o and GPT-4.1 (31.3% and 32.2%, respectively), even though the latter are widely presumed to be much larger and generally stronger models.

This discontinuity suggests that model size alone is a poor predictor of relational analogy induction; data curation, architecture, or specific training regimes may matter more than scale in this setting. Our findings echo recent work questioning the universality of scaling laws and the nature of "emergent" abilities in LLMs (Schaeffer et al., 2023). Overall, scaling up does not guarantee robust lexical relational abstraction, and highly tuned open-source models may surpass commercial systems on tasks outside the typical benchmark suite.

5.5 Broader Impact

Systematic analogical reasoning is a central component of human language understanding, with broad implications for tasks such as scientific discovery, education, and knowledge transfer. By developing LAMBDA, a minimal-shot lexical analogy benchmark, we provide researchers and practitioners with a targeted way to assess whether language models move beyond surface pattern matching toward genuine relational inference. While the current results reveal that even state-of-the-art models have not mastered these capabilities, a clearer diagnostic benchmark can help focus community efforts on genuine abstraction rather than superficial accuracy.

This work may also inform research in linguistic theory, cognitive modeling, and educational technology by clarifying which relational skills remain out of reach for current systems. At the same time, exposing specific

weaknesses in analogy-making helps mitigate the risk of overestimating language model generalization in downstream applications such as question answering, knowledge graph completion, or scientific information extraction. More rigorous evaluation of relational abstraction is a step toward safer and more interpretable AI, especially as models are deployed in high-stakes or high-impact language tasks.

Limitations and Future Directions

The present version of LAMBDA provides a targeted yet necessarily partial lens on analogical reasoning in language models. The design choices—English-only vocabulary, WordNet-derived relations, strict single-token scoring, and uniform two-shot prompts—were made to ensure interpretability, reproducibility, and statistical rigor. At the same time, these constraints define the edges of what the benchmark currently measures and suggest a broad pathway for future enhancements.

One central limitation is linguistic scope. LAMBDA is English-only and restricts its queries to lemmas covered by WordNet, a resource that, while comprehensive, does not encompass the full diversity of natural language morphology or evolving vocabulary. This leaves open questions about model performance in morphologically rich languages, low-resource settings, or informal registers. Future extensions of LAMBDA could include multilingual or code-switching analogies to examine the universality of analogical reasoning capabilities across languages.

The composition of the candidate sets reflects a further bias: WordNet favors formal, established vocabulary and under-represents colloquial, idiomatic, or domain-specific terms. As a result, models may struggle with analogy problems involving emerging or specialized terminology, and our observed difficulty levels may partly reflect this sampling bias. Addressing lexical coverage could involve augmenting LAMBDA with dynamic corpora or alternative lexical resources, ensuring analogies span both standard and non-standard language. Prior exposure to WordNet-derived content is additionally a background risk for many NLP benchmarks, but our use of a deterministic generation pipeline makes the construction fully auditable and exactly reproducible, which strengthens transparency and supports careful inspection of any potential leakage.

Another notable constraint lies in the evaluation protocol. Future work might allow free-form answers with embedding-based or human-in-the-loop evaluation to acknowledge semantically correct analogies outside the strict WordNet set, rather than our current single-token approach. Similarly, expanding beyond the lexical level to phrase-level or sentence-level analogies, or multimodal analogies (combining text and images), would push models closer to the complexity of human analogy-making.

Finally, while we correlate accuracy with model scale and training regimen, the present analysis does not isolate which aspects of pre-training data or architecture drive analogical competence. Controlled ablations—varying one factor at a time—could reveal, for instance, whether explicit morphological pretraining, relation-focused fine-tuning, or architectural tweaks (e.g., character-level representations) lead to improved systematicity. Addressing these questions requires new model variants and targeted training interventions beyond the scope of this evaluation.

In summary, LAMBDA is a step toward rigorous evaluation of analogical abstraction in language models, but it captures only a slice of this broad ability. Expanding the benchmark's language coverage, answer flexibility, prompt diversity, analogy types, and tying evaluations to specific model design choices will all further illuminate how, when, and why language models develop analogical reasoning. We hope this work lays a foundation for such investigations, moving the field closer to models that can truly reason by analogy in the rich way humans do.

6 Conclusion

This study introduces LAMBDA, a benchmark designed to evaluate few-shot lexical analogical reasoning in language models. We constructed a dataset of 3,000 relation-hidden analogies spanning synonyms, antonyms, and derivational morphology, and used it to systematically assess both open and closed models across a range of scales. Our experiments reveal a steep performance gradient: no model surpassed 50% accuracy overall or within any relation category. We analyzed the influence of query length and identified recurring error

patterns (identity echo, surface misfire, semantic drift), confirming that major challenges persist, particularly for synonyms and derivations, and that systematic generalization remains elusive even in today's strongest models.

Although language models have advanced rapidly, there is still a substantial gap between surface pattern recognition and reliable analogical reasoning. LAMBDA enables fine-grained comparison between models that otherwise appear similar on standard benchmarks, and it motivates the development of more specialized evaluation protocols and training strategies. The limitations we discussed point to concrete directions for future work, including multilingual extensions, flexible answer evaluation, richer prompt formats, and targeted ablations to understand the mechanisms behind relational inference.

Overall, our results emphasize the need to go beyond aggregate performance metrics toward more rigorous evaluations of abstract reasoning. By releasing LAMBDA, we aim to catalyze research that addresses these gaps and brings language models closer to more systematic language understanding. Our findings highlight a persistent discrepancy between human-level and model-level analogical reasoning, underscoring the value of targeted benchmarks like LAMBDA in driving progress toward more general and reliable AI language capabilities.

References

Andrea Berti, Thomas Scalf, and Hongjing Lu. A survey on emergent abilities in large language models. *arXiv preprint arXiv:2503.05788*, 2025. URL https://arxiv.org/abs/2503.05788.

Ben Bogin, Matt Gardner, and Jonathan Berant. Representing schema structure with graph neural networks for text-to-sql parsing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4560–4565, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1448. URL https://aclanthology.org/P19-1448.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021. URL https://arxiv.org/abs/2108.07258.

Zied Bouraoui, José Camacho-Collados, and Steven Schockaert. Inducing relational knowledge from BERT. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 7456–7463, 2020. doi: 10.1609/aaai.v34i05.6236. URL https://ojs.aaai.org/index.php/AAAI/article/view/6236.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language

models are few-shot learners. In *Advances in Neural Information Processing Systems*, 2020. URL https://arxiv.org/abs/2005.14165.

Jonathan H. Clark, Dan Garrette, Iulia Turc, and John Wieting. CANINE: Pre-training an efficient tokenization-free encoder for language representation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3566–3583, Seattle, United States, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.262. URL https://aclanthology.org/2022.naacl-main.262.

Todd R. Davies. A logical approach to reasoning by analogy. Technical Report 575, SRI International, 1987. URL https://www.sri.com/wp-content/uploads/2021/12/575.pdf.

Aleksandr Drozd, Anna Gladkova, and Satoshi Matsuoka. Word embeddings, analogies, and machine learning: Beyond king - man + woman = queen. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pp. 3519–3530, Osaka, Japan, December 2016. The COLING 2016 Organizing Committee. URL https://aclanthology.org/C16-1332/.

Brian Falkenhainer, Kenneth D. Forbus, and Dedre Gentner. The structure-mapping engine: Algorithm and examples. *Artificial Intelligence*, 41(1):1–63, 1989. doi: 10.1016/0004-3702(89)90081-5.

Ramon Ferrer-i Cancho and Ricard V. Solé. Least effort and the origins of scaling in human language. *Proceedings of the National Academy of Sciences*, 100(3):788–791, 2003. doi: 10.1073/pnas.0335980100. URL https://www.pnas.org/doi/10.1073/pnas.0335980100.

Bin Gao, Jiang Bian, and Tie-Yan Liu. Wordrep: A benchmark for research on learning word representations, 2014. URL https://arxiv.org/abs/1407.1640.

Dedre Gentner. Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, 7(2):155–170, 1983. doi: 10.1207/s15516709cog0702_3. URL https://doi.org/10.1207/s15516709cog0702_3.

Dedre Gentner and Arthur B. Markman. Structure mapping in analogy and similarity. *American Psychologist*, 52(1):45–56, 1997. doi: 10.1037/0003-066X.52.1.45.

Mary L. Gick and Keith J. Holyoak. Analogical problem solving. *Cognitive Psychology*, 12(3):306–355, 1980. doi: 10.1016/0010-0285(80)90013-4.

Anna Gladkova, Aleksandr Drozd, and Satoshi Matsuoka. Analogy-based detection of morphological and semantic relations with word embeddings: What works and what doesn't. In *Proceedings of the NAACL-HLT Student Research Workshop*, pp. 47–54, San Diego, California, 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-2002. URL https://aclanthology.org/N16-2002.

Google DeepMind. Gemini 2.5 pro: A multimodal frontier model, 2025. URL https://deepmind.google/gemini-25-pro.

Douglas R. Hofstadter. Analogy as the core of cognition. In Dedre Gentner, Keith J. Holyoak, and Boicho N. Kokinov (eds.), *The Analogical Mind: Perspectives from Cognitive Science*, pp. 499–538. MIT Press, 2001.

Dieuwke Hupkes, Verna Dankers, Mathijs Mul, and Elia Bruni. Compositionality decomposed: How do neural networks generalise? *Journal of Artificial Intelligence Research*, 67:757–795, 2020. doi: 10.1613/jair.1.11674. URL https://jair.org/index.php/jair/article/view/11674.

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. Mistral 7b. In *arXiv preprint arXiv:2310.06825*, 2023. URL https://arxiv.org/abs/2310.06825.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Daniel Keysers, Nathanael Schärli, Nathan Scales, Hylke Buisman, Daniel Furrer, Sergii Kashubin, Nikola Momchev, Danila Sinopalnikov, Lukasz Stafiniak, Tibor Tihon, Dmitry Tsarkov, Xiao Wang, Marc van Zee, and Olivier Bousquet. Measuring compositional generalization: A comprehensive method on realistic data. In *International Conference on Learning Representations*, 2020. URL https://arxiv.org/abs/1912.09713.

Minyoung Kim, Hyounghun Kim, Joonhyeok Shin, and Jinwoo Shin. VIOLIN: Virtual observatory for language model inference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11234–11243, Vancouver, Canada, 2023. IEEE. URL https://openaccess.thecvf.com/content/CVPR2023/html/Kim_VIOLIN_Virtual_Observatory_for_Language_Model_Inference_CVPR_2023_paper.html.

Thijs Linssen and Anna Rogers. Subword models struggle with word learning, but surprisal hides it. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 1987–1998, Dublin, Ireland, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.156. URL https://aclanthology.org/2022.findings-acl.156.

Tal Linzen. Issues in evaluating semantic spaces using word analogies. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pp. 13–18, Berlin, Germany, 2016. Association for Computational Linguistics. doi: 10.18653/v1/W16-2503. URL https://aclanthology.org/W16-2503/.

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

Lajanugen Logeswaran, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, Jacob Devlin, and Honglak Lee. Zero-shot entity linking by reading entity descriptions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3449–3460, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1335. URL https://aclanthology.org/P19-1335.

Yujia Lu, Yitao Duan, Zhengbao Jiang, Aaron Dant, Mukund Sundararajan, and Yuchen Li. Pretrained transformers for text attribution. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 2721–2734, Abu Dhabi, United Arab Emirates, 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.emnlp-main.178.

R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3428–3448, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1334. URL https://aclanthology.org/P19-1334.

Meta AI. The llama 4 herd: The beginning of a new era of natively multimodal intelligence, 2025. URL https://ai.meta.com/blog/llama-4-multimodal-intelligence.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 746–751, Atlanta, Georgia, 2013. Association for Computational Linguistics. URL https://aclanthology.org/N13-1090.

George A. Miller. Wordnet: A lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995. doi: 10.1145/219717.219748. URL https://dl.acm.org/doi/10.1145/219717.219748.

Nikola Mrkšić, Diarmuid Ó Séaghdha, Blaise Thomson, Milica Gašić, Lina Rojas-Barahona, Pei-Hao Su, David Vandyke, Tsung-Hsien Wen, and Steve Young. Counter-fitting word vectors to linguistic constraints. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 142–148, San Diego, California, 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-1018. URL https://aclanthology.org/N16-1018.

Jiaxin Mu and Zeming Tomos. CALM: Character-aware language models are more robust to tokenization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 5678–5690, Singapore, 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.emnlp-main.347.

Aakanksha Naik, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, and Graham Neubig. Stress test evaluation for natural language inference. In *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 2340–2353, Santa Fe, New Mexico, USA, 2018. Association for Computational Linguistics. URL https://aclanthology.org/C18-1198.

OpenAI. Gpt-4o technical report, 2024. URL https://openai.com/research/gpt-4o.

OpenAI. Introducing gpt-4.1 in the api. https://openai.com/index/gpt-4-1/, April 2025.

OpenAI. Introducing GPT-5.2. OpenAI Blog, Dec 2025. URL https://openai.com/index/introducing-gpt-5-2/.

Gustaw Opiełka, Anna Rogers, Djamé Seddah, and Aleksandr Drozd. Conceptvectors: A new approach to interpretability in language models. *arXiv preprint arXiv:2503.03666*, 2025. URL https://arxiv.org/abs/2503.03666.

Steven T. Piantadosi, Harry Tily, and Edward Gibson. Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, 108(9):3526–3529, 2011. doi: 10.1073/pnas.1012551108. URL https://www.pnas.org/doi/10.1073/pnas.1012551108.

Anthony Piotte, Anna Rogers, and Djamé Seddah. Morphoops: Investigating morphological robustness of neural language models. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pp. 212–225, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.conll-1.17. URL https://aclanthology.org/2021.conll-1.17.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI Technical Report*, 2019. URL https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

Philipp Romanov and Rozita Khusainova. Evaluating the morphological sensitivity of word embeddings. *Computational Linguistics*, 49(2):267–293, 2023. doi: 10.1162/coli_a_00468. URL https://aclanthology.org/2023.cl-2.2.

Stuart J. Russell. A quantitative analysis of analogy by similarity. *Proceedings of the Fifth National Conference on Artificial Intelligence (AAAI-86)*, pp. 284–288, 1986a. URL https://cdn.aaai.org/AAAI/1986/AAAI86-047.pdf.

Stuart J. Russell. *Analogical and Inductive Reasoning*. PhD thesis, Stanford University, 1986b.

Stuart J. Russell. Analogy by similarity. In David H. Helman (ed.), *Analogical Reasoning: Perspectives of Artificial Intelligence, Cognitive Science, and Philosophy*, pp. 251–269. Kluwer Academic Publishers, Dordrecht, 1988. URL https://people.eecs.berkeley.edu/~russell/papers/helman88-similarity.pdf.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are emergent abilities of large language models a mirage? *arXiv preprint arXiv:2304.15004*, 2023. URL https://arxiv.org/abs/2304.15004.

João Sedoc, Rachel Rudinger, and Sebastian Ostermann. Retrofitting contextualized word embeddings to wordnet ontology. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 5656–5671, Online, 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.naacl-main.450.

Claude E. Shannon. Prediction and entropy of printed english. *Bell System Technical Journal*, 30(1):50–64, 1951. doi: 10.1002/j.1538-7305.1951.tb01366.x.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. LLaMA: Open and efficient foundation language models, 2023. URL https://arxiv.org/abs/2302.13971.

Jamie Turner and Adina Williams. Tokenization bottlenecks in language model training. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 7890–7902, Singapore, 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.emnlp-main.487.

Peter D. Turney. The latent relation mapping engine: Algorithm and experiments. *Journal of Artificial Intelligence Research*, 33:615–655, 2008. URL https://auld.aaai.org/Papers/JAIR/Vol33/JAIR-3317.pdf.

Ekaterina Vylomova, Trevor Cohn, Xuanli He, and Gholamreza Haffari. Word representation models for morphologically rich languages in neural machine translation. In *Proceedings of the First Workshop on Subword and Character Level Models in NLP*, pp. 103–108, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-4115. URL https://aclanthology.org/W17-4115/.

Taylor Webb, Keith J. Holyoak, and Hongjing Lu. Emergent analogical reasoning in large language models. *arXiv preprint arXiv:2212.09196*, 2022. URL https://arxiv.org/abs/2212.09196.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, 2022. URL https://arxiv.org/abs/2201.11903.

Hanwen Yang, Eric Zhou, and Tianhao Wu. SEGON: Segmenting out-of-distribution tokens for robust language modeling. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pp. 1345–1357, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.74. URL https://aclanthology.org/2024.acl-long.74.

Siyu Yuan, Jiangjie Chen, Changzhi Sun, Jiaqing Liang, Yanghua Xiao, and Deqing Yang. ANALOGYKB: Unlocking analogical reasoning of language models with a million-scale knowledge base. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pp. 1249–1265, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.68. URL https://aclanthology.org/2024.acl-long.68.

Wenhao Zhou, Ben Zhou, Kyle Richardson, and Nathan Schneider. Least-to-most prompting enables complex reasoning in large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pp. 1288–1300, Toronto, Canada, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.72. URL https://aclanthology.org/2023.acl-long.72.

George K. Zipf. The psycho-biology of language: An introduction to dynamic philology. 1935.

George K. Zipf. *Human Behavior and the Principle of Least Effort*. Addison-Wesley, 1949.

Appendix A Dataset Statistics

A.1 Token-length distribution

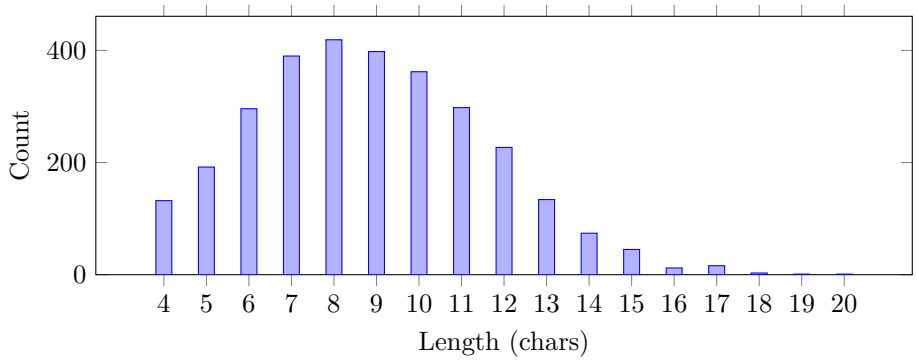


Figure A.1: Histogram of query-word lengths. Median = 9, 90th percentile = 12.

Table A.1: Counts of query-word lengths.

Length	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Count	132	192	296	390	419	398	362	298	227	134	74	45	12	16	3	1	1

Table A.2: Descriptive statistics for query-word length.

Count	**Mean**	**Median**	**90-pct**	**Min**	**Max**
3 000	8.85	9	12	4	20

Ten percent of queries contain 13–20 characters, adding morphological variety that can hinder surface memorisation.

A.2 Part-of-speech breakdown

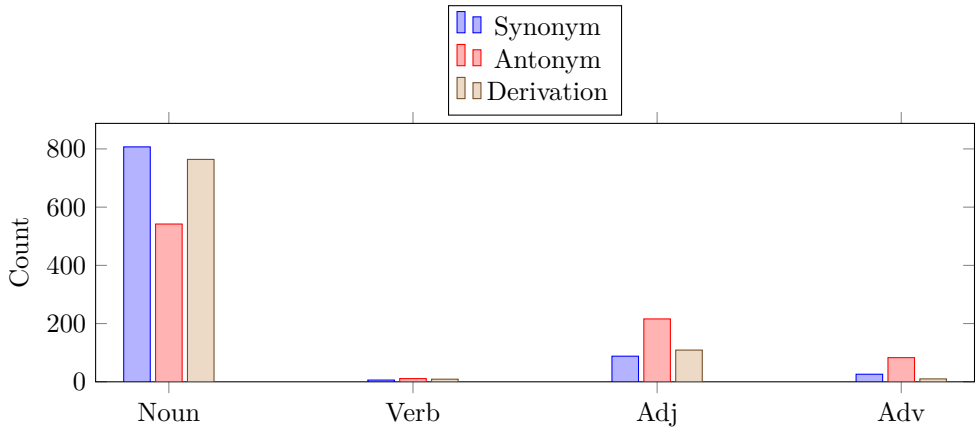


Figure A.2: POS distribution per relation.

Synonym and derivation queries are noun-heavy, while antonyms skew toward adjectives, mirroring WordNet polarity pairs such as *hot–cold*.

A.3 Candidate-set sizes

Table A.3: WordNet candidate-set size $|R_r(E)|$ for query lemmas.

Relation	Mean	St Dev	Min	Max
Synonym	3.93	4.96	1	54
Antonym	1.35	0.71	1	5
Derivation	3.53	4.13	1	62

Antonym queries are nearly single-choice; synonym and derivation queries present much larger target sets, demanding stronger relational inference.

Appendix B Generation and Analysis Scripts [4]

Listing 1: Deterministic item-generation script

```python
import json, random
from nltk.corpus import wordnet as wn

def get_synonyms(w):
    s = set()
    for syn in wn.synsets(w):
        for l in syn.lemmas():
            if l.name().lower() != w.lower():
                s.add(l.name().replace('_',' '))
    return list(s)

def get_antonyms(w):
    s = set()
    for syn in wn.synsets(w):
        for l in syn.lemmas():
            for a in l.antonyms():
                if a.name().lower() != w.lower():
                    s.add(a.name().replace('_',' '))
    return list(s)

def get_derivations(w):
    s = set()
    for syn in wn.synsets(w):
        for l in syn.lemmas():
            for d in l.derivationally_related_forms():
                if d.name().lower() != w.lower():
                    s.add(d.name().replace('_',' '))
    return list(s)

def pick_related(word, func):
    lst = func(word)
    return None if not lst else random.choice(lst)

words = [w for w in set(wn.all_lemma_names())
         if w.isalpha() and w.islower() and len(w) > 3]

random.seed(42)
dataset = []

for func, label in [(get_synonyms,"synonym"),
```

[4] Full code will be released on GitHub upon publication.

```
                    (get_antonyms,"antonym"),
                    (get_derivations,"derivation")]:
    cnt = 0
    while cnt < 1000:
        chosen = set()
        triples = []
        for _ in range(3):
            tries = 0
            while tries < 50:
                c = random.choice(words)
                if c not in chosen:
                    r = pick_related(c,func)
                    if r and r not in chosen:
                        triples.append((c,r))
                        chosen.update((c,r))
                        break
                tries += 1
        if len(triples) != 3: continue
        (A,B),(C,D),(E,F) = triples
        if len({A,B,C,D,E,F}) < 6: continue
        # ensure no word is a substring of another
        if any(x in y or y in x for x in chosen for y in chosen if x != y):
            continue
        dataset.append({"few_shot":[{"input":A,"output":B},
                                    {"input":C,"output":D}],
                        "question":f"{E} : ?",
                        "relation":label})
        cnt += 1

with open("dataset/lexical_dataset.jsonl","w",encoding="utf-8") as f:
    for entry in dataset:
        f.write(json.dumps(entry)+"\n")
```

The following script computes per-length accuracies and correlation:

Listing 2: Length-wise accuracy and Spearman correlation

```
import pandas as pd
from scipy.stats import spearmanr

data = [
    (4, 40.2), (5, 44.3), (6, 43.6), (7, 42.8), (8, 48.9),
    (9, 47.0), (10, 53.9), (11, 53.0), (12, 58.6), (13, 63.4),
    (14, 64.9), (15, 73.3), (16, 41.7), (17, 81.2), (18, 66.7),
    (19, 0.0), (20, 100.0)
]

df = pd.DataFrame(data, columns=["length", "accuracy"])
r, temp = spearmanr(df["length"], df["accuracy"])
print(f"Spearman r = {r:.2f}")
```

Appendix C Prompt Templates

Default prompt template (main experiments). Each dataset item is rendered as two support pairs followed by the query line:

```
A : B
C : D
E : ?
```

We then append the following answer-only guard (verbatim) to the user message:

```
Answer (only output the single word answer, absolutely nothing else).  IT
IS IMPERATIVE THAT YOU DO NOT RESPOND WITH ANYTHING OTHER THAN THE ONE WORD
```

```
ANSWER. DO NOT REPEAT THE QUESTION. DO NOT SAY 'THE ANSWER IS'. DO NOT SAY
'ANSWER:' OR 'ANSWER (OR ALTERNATIVE ANSWER/DESCRIPTION)', and do not show
your thought process.  It will only be a one-word answer, not multiple
words, phrases, or sentences.  DO NOT RESPOND WITH ANYTHING OTHER THAN THE
ANSWER. Additionally, do not add anything else from the definition or show
any of your thought process.  Do not include any part of the analogy you
are answering other than the answer.  If you output anything other than the
one-word answer, it will be marked as incorrect.  What is the answer?  The
answer is:
```

Prompt-variant ablations (Figure/Table reference in main text). To quantify prompt sensitivity, we also evaluate the same two-shot exemplars under two simplified variants:

```
Variant 1:  A : B
C : D
E : ?
?
Variant 2:  A : B
C : D
Given the above analogies, infer the missing word for E : ?.  Respond with
a single English word only.
```