POLYCROSS: A BENCHMARK FOR IDENTIFYING HIDDEN MEANINGS IN SHORT PHRASES WITH LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Identifying hidden meanings in short text snippets is a challenging task with applications in word puzzles. In this work we investigate the capacity of large language models (LLMs) to identify concealed meanings in short phrases, provide rationales for the existence of meanings identified, and translate phrases with multiple meanings into another language. For the empirical evaluation we introduce POLYCROSS, a benchmark dataset built from a special class of crosswords clues in Romanian which we call *polysemantic clues*. Each such clue carries multiple meanings, including a subtle meaning that leads to the actual crosswords solution. The baseline is THEMCROSS, a dataset of general-knowledge Romanian crosswords clues with no hidden meanings. We use the two data sets to evaluate several LLMs (GPT-OSS, LLaMA, Granite, Claude and LLMic) with progressively stronger hints about solutions. We assess LLM accuracy of finding crosswords solutions and the quality of LLM rationales to explain their solutions. Our results show that the performance of the LLMs evaluated is substantially lower on polysemantic clues (POLYCROSS) than on the baseline set (THEMCROSS). This highlights an important gap in current LLM capabilities and establishes POLYCROSS as a benchmark to measure future progress on this problem.

1 INTRODUCTION

The ability to discover a hidden meaning in a short phrase, in addition to a more obvious, common meaning, is important in crosswords puzzles, a widely used form of entertainment around the world. Crosswords puzzles often employ subtle plays on words that can add an entirely new meaning to a clue. Addressing such challenges is a non-trivial task for human users, requirering not only effort but also specialised skills and experience.

In this work we investigate the ability of large language models (LLMs) to accurately discover hidden meanings in short phrases, such as crosswords clues in the Romanian language. Specifically, we measure their ability to find the solution to a clue with a hidden meaning, to give an accurate rationale for why the clue matches the solution, and to translate clues with multiple meanings in such a way that multiple meanings are preserved in the translation.

To perform our LLM evaluation we introduce POLYCROSS, a benchmark dataset for LLM evaluation that consists of a special type of crosswords clues and their solutions. The structure (clue, solution) of each record makes our dataset akin to question-answering datasets from the literature. However, compared to typical question answering datasets, POLYCROSS clues have multiple meanings each. The meaning that leads to the solution is meant to be more subtle, more difficult to discover compared to others. For instance, an POLYCROSS clue often has two meanings, one more obvious, but unrelated to the solution, and one more subtle, which leads to the solution. We call these *polysemantic clues*[1] and we call puzzles with such clues *polysemantic crosswords puzzles*.

As an example of a polysemantic clue consider the clue "exponent of the industrial revolution" with the solution ROTOR. The clue leads the user into thinking of a historical period, the industrial revolution.

[1] In this work, we simply use the term "polysemantic" to characterize a word or phrase with multiple meanings or interpretations. Finer distinctions, such as polysemy versus homonymy, well established in the linguistic literature, are beyond the scope of our work.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

However, to arrive at the correct solution ROTOR, the word revolution should be understood to mean a physical rotation rather than a historical period. Since a rotor is a rotating part, often as part of an industrial machinery, the solution is, indeed, ROTOR.

While polysemy is present in additional types of datasets in the literature, such as crosswords cryptic clues (Friedlander & Fine, 2016) and puns (Delabastita, 2016; Su et al., 2025), our POLYCROSS clues are different from these in several important ways. Cryptic clues follow a domain-specific convention in which each clue contains two components—a definition and a wordplay cipher—that together form a fluent but misleading phrase (Rozner et al., 2021). By contrast polysemantic clues do not rely on any predefined structure or formal conventions: they exploit naturally occurring ambiguity in language, making them more representative of real-world semantic ambiguity. Although both polysemantic clues and puns leverage polysemy, puns are primarily designed to create humor, whereas polysemantic clues point to a specific solution that can serve as ground truth in the evaluation of language models.

We use POLYCROSS to evaluate LLM performance on the three tasks mentioned, namely finding solutions, providing rationales and translating a clue while preserving its meanings. For the first task we use the five LLMs: LLMic (Bădoiu et al., 2025) – the only publicly available Romanian trained LLM, gpt-oss-120b (OpenAI et al., 2024), Llama-3.3-70b (Touvron et al., 2023), Granite-4.0-small (Mishra et al., 2024) and Claude Opus 4.8 (Anthropic, 2026). They are prompted to find solutions to POLYCROSS clues in a range of experiments with gradually more hints given to them. Hints start with the length of the solution, which is then complemented with the first letter of the solution, then the second letter, the third and the fourth.

For the second task the LLMs mentioned are prompted to give a rationale for having chosen that solution. A human expert then evaluates and scores a sample of the rationales generated. The expert has over 35 years of experience with composing, publishing and solving crosswords puzzles in Romanian. The actual rationale of why a polysemantic clue matches a solution is quite obvious to a human expert in most cases. As such it is easy for a human expert to tell with a high confidence if and *why* a given solution is acceptable for a given polysemantic clue. In fact in our experiments the human expert was able to classify all LLM-provided rationales into three classes: poor, partially correct, and correct.

For the third task we prompt the LLMs to translate clues from Romanian to English and find the solution on the translated clue.

Our baseline dataset of clues and solutions is called THEMCROSS and has clues with no hidden meanings (also called direct clues). This dataset tests base knowledge similarly to trivia datasets rather then using plays on words to construct riddles. The baseline dataset has a number of commonalities with POLYCROSS, such as being in the same language and being extracted from crosswords puzzles as well. This is why we introduce THEMCROSS as a baseline dataset, rather than using some of the many datasets available in English, a collection of which is available at: `https://github.com/ad-freiburg/large-qa-datasets`. Our evaluation shows significantly better LLM performance, both in terms of solutions found, rationales given and translation, on the baseline datatset (THEMCROSS) than on the experimental dataset (POLYCROSS).

In summary this paper makes the following **contributions**. First, we introduce a polysemantic dataset of (clue, solution) records in which each clue has multiple meanings. Second, we use the dataset to evaluate LLM abilities to find solutions that correspond to a subtle meaning in a clue, to give a rationale of why a given solution matches a given clue, and to translate clues. We find that LLMs ability to find solutions in POLYCROSS is substantially below that in THEMCROSS. The highest accuracy on POLYCROSS is about 77% for gpt-oss-120b and 86% for Claude Opus 4.8 with the most detailed level of hinting versus 87% and 95% on THEMCROSS. We find that giving correct rationales is difficult in POLYCROSS for the LLMs evaluated even in those instances where the LLMs find the actual solution. Even in cases where the actual solution is found the highest percentage of correct rationales is about half of the cases versus 86-98% in THEMCROSS. Translating POLYCROSS clues turns out to yield a low accuracy, with differences in performance from THEMCROSS going to over one order of magnitude. While the magnitude of the difference is insightful, the low performance of polysemantic clue translation is not surprising. Translating POLYCROSS clues with a preservation of their multiple meanings is difficult in many cases, if even possible, for human experts. Overall, our

results indicate the need for a benchmark dataset to enable performance measurements in time for various LLM instances and to drive further progress in this challenging problem.

2 PROBLEM FORMULATION

Consider an ordered set of (clue,solution) pairs $\mathcal{D} = \{(c_i, s_i)\}_{i=1}^{N}$ with an associated list of hints $(h_1, \ldots, h_N)$. Solving a clue c_i means mapping it and an optional hint h_i to s_i. The corresponding performance measure, the *solution accuracy*, is the percentage of clues successfully mapped to their solutions in $\mathcal{D}$.

A solver taking a clue c_i (and optional hint h_i) and outputting the solution s_i' can also produce a textual rationale r_i explaining its choice of s_i' given c_i and h_i. A human expert then evaluates the quality of the rationale by assigning integer q_i from a fixed scale to the tuple (c_i, h_i, s_i', r_i). The corresponding performance measure is then the mean *rationale quality* which is q_i averaged over tuples (c_i, h_i, s_i', r_i) in which the solver's solution s_i' is correct: $s_i' = s_i$.

We also consider a translation-based approach to solving a clue c_i. In this setting, the clue c_i—and optionally the associated hint h_i—is first translated from the source language (e.g., Romanian) into a target language, yielding c_i' (and h_i'). A solution s_i' is then obtained in the target language and subsequently translated back into the source language. We evaluate this approach using the same solution-accuracy metric as in the original setting.

To illustrate, consider the following actual example from POLYCROSS translated from Romanian to English. The clue c is "a burning desire" associated with the solution s being CIGARETTE. A hint h may consist of the solution length (9) and the first three letters (i.e., the prefix) CIG. The rationale $r =$ "a cigarette is something many people desire or crave, and it is burning while being consumed" would be assigned the highest quality value.

3 RELATED WORK

Cryptic crossword solving has been studied as a challenging benchmark for compositional language understanding, combining semantic interpretation with symbolic wordplay and strict character-level constraints. More recently large language models have substantially improved raw clue-solving accuracy via prompting and chain-of-thought reasoning (Sadallah et al., 2025; Saha et al., 2025), yet still struggle with solving semantically more complex cryptic crosswords. To address this reasoning-focused pipelines have reframed cryptic solving as a hypothesis-and-verification task, formalizing wordplay into executable code checked by symbolic verifiers (Andrews & Witteveen, 2025), yielding improved correctness and interpretability. Overall prior work highlights a trade-off between fluency and rigor: LLMs generate plausible answers, while verification-based methods improve reliability at the cost of complexity, positioning cryptic crosswords as a stringent testbed for constraint-aware and compositional reasoning.

Leng et al. (2025) introduce CrossWordBench, a crosswords benchmark for LLMs and for large vision-language models (LVLMs). CrossWordBench uses pre-existing clue-solution pairs to generate puzzles that are then provided in textual and visual formats which can further be used to evaluate LLMs and LVLMs on puzzle solving. In contrast our focus is on introducing a novel benchmark with a distinctive type of clues, and using the benchmark to perform an in-depth evaluation of LLM capabilities to identify hidden meanings in the clues.

Xu et al. (2024) evaluate LLMs on their ability of recognizing, explaining, and generating puns. They report that models often struggle on all three tasks, though a good performance, such as explanations competitive with human explanations, is observed in some cases. Zangari et al. (2025) introduce new datasets with puns. These are used for an LLM evaluation, which further points out the limits of current technology in this domain. On the other hand we address a different problem of finding and explaining solutions to polysemantic clues where the meaning that leads to the solution is not necessarily obvious.

4 THE DATASETS

In this section we detail our two datasets, POLYCROSS and THEMCROSS[2].

POLYCROSS is the main benchmark dataset that we propose in this work. As argued in the introduction, POLYCROSS is important to evaluate the ability of LLMs to identify a hidden meaning in a short phrase, give a rationale that can be used to assess the LLM ability to understand the hidden meaning and perform a translation while preserving the multiple meanings present in the phrase at hand.

As a control we introduce a baseline dataset called THEMCROSS. Both datasets are in the same language and they are sourced from crosswords puzzles. The key difference is that THEMCROSS clues are not polysemantic. THEMCROSS is extracted from so-called thematic puzzles meant to test general knowledge. In THEMCROSS clues have a clear meaning with no additional hidden meanings and no intention to trick the user. We call these monosemantic clues. Thus, THEMCROSS is in fact a question-answering dataset, akin to trivia datasets.

To construct POLYCROSS we parsed 897 polysemantic crosswords puzzles published in Romanian crosswords publications (Reb, a;b;c) and made available online (Reb, d). Polysemantic puzzles should not contain monosemantic clues, and this requirement is checked by editors prior to publication. We extracted 30,829 clue-solution pairs from the 897 polysemantic puzzles considered. Out of these we excluded from our analysis 4,645 records where the solution is a two-letter sequence, many of which are not proper words. A very simple example in English would be "The limits of the Universe!", with the solution UE (i.e., first and last letter of word "Universe"). Observe how this relies on domain-specific rules such as looking at individual letters inside a word, which is beyond the scope of our work. In the rest of this paper we assume that POLYCROSS has 26,184 records, all having a solution length of 3 or greater.

THEMCROSS contains 1,172 (clue, solution) pairs obtained from 34 thematic crosswords puzzles Reb (c). This excludes any records with two-letter solutions, for the same reason as in the case of POLYCROSS.

5 EVALUATION & DISCUSSION

In this section we empirically evaluated the proposed POLYCROSS benchmark using open-source large language models from the IBM Granite (Mishra et al., 2024), Meta LLaMA (Touvron et al., 2023) and OpenAI GPT-OSS (OpenAI et al., 2024) families. Our evaluation includes the Granite-4.0-small, Llama-3.3-70b, and gpt-oss-120b models, covering a range of model architectures and scales (from 30 to 120 billion parameters). We used a hosted inference environment equipped with eight A100 80GB GPUs capable of serving 1,500 requests per minute using vLLM . In addition, we also used LLMic-v2-3b Bădoiu et al. (2025), the only publicly available Romanian trained LLM. For reference we ran Claude Opus 4.8 (Anthropic, 2026) a multi-lingual frontier model. Throughout our experiments we used greedy decoding (temperature 0).

5.1 SOLUTION FINDING AND TRANSLATION

We designed five prompt variants to infer solutions for the clues in the benchmark dataset, out of which one focuses on translation. Specifically, we considered:

- **v1**: A few-shot prompt that instructs the model to guess the solution to a given clue (see Figure 4 in the appendix).
- **v2**: A variant of **v1** augmented with a hint specifying the number of letters in the answer (see Figure 5 in the appendix).
- **v3**: An extension of **v2** that additionally provides a short prefix of up to four letters of the correct answer (see Figure 6 in the appendix). Thus, we test four sub-variants of this main prompt, each having a different of the solution prefix given as a hint in addition to the solution length.

[2]Both datasets are included in the supplementary material.

Table 1: Accuracy (%) on the POLYCROSS and THEMCROSS datasets. Best per column in bold. No = no hint (v1); Len = length hint (v2); Tr. = translate (v5).

			v3: length + prefix				v4: v3 + CoT				v5
Model	No	Len	k=1	k=2	k=3	k=4	k=1	k=2	k=3	k=4	Tr.
PolyCross											
Granite-4.0-small	0.6	1.1	5.1	11.8	28.7	40.8	3.5	7.0	20.3	32.7	0.7
Llama-3.3-70B	2.5	4.1	12.3	25.0	43.3	59.8	14.1	25.4	46.9	58.8	4.6
gpt-oss-120B	3.8	7.9	23.0	39.7	61.4	76.6	23.5	40.0	62.9	76.3	8.0
Claude Opus 4.8	**5.2**	**8.7**	**28.4**	**50.7**	**75.8**	**85.6**	**29.6**	**50.7**	**75.9**	**86.1**	**8.8**
LLMic-v2-3B	0.4	0.1	1.5	3.9	22.8	43.6	–	–	–	–	–
ThemCross											
Granite-4.0-small	11.1	13.6	20.1	27.2	40.4	54.4	11.4	15.4	19.3	30.1	9.8
Llama-3.3-70B	30.9	35.2	46.1	52.7	63.9	74.3	48.1	53.3	63.8	73.8	39.4
gpt-oss-120B	35.2	46.6	56.9	64.4	77.6	87.5	58.9	65.6	76.5	88.2	47.6
Claude Opus 4.8	**59.1**	**65.1**	**83.0**	**88.3**	**92.8**	**95.5**	**83.0**	**88.3**	**93.5**	**95.6**	**73.5**
LLMic-v2-3B	4.7	3.7	7.4	11.9	26.7	47.3	–	–	–	–	–

- **v4**: A chain-of-thought prompt that encourages explicit reasoning before producing an answer. This variant uses the same hints as **v3**, and the model is instructed to reason in Romanian (see Figure 7 in the appendix). Similarly to **v3**, we test four sub-variants of this prompt.
- **v5**: A chain-of-thought prompt that instructs the model to translate the clue and reason in English. In this case, the hint consists of the number of letters in the Romanian answer (see Figure 8 in the appendix).

Note that LLMic is a 3B-parameter model without instruction tuning or a chat template so it cannot follow the JSON instruction prompts above. Instead, we evaluated it with simpler few-shot *completion* prompts that carry the same three worked examples and the same hint wording as **v1**–**v3**, ending in "Raspuns:"; the answer is its greedy decoding based continuation up to the first newline. Since **v4** and **v5** rely on following instructions they are omitted ("–" in the tables).

For our experiments, we sampled uniformly at random 3,000 POLYCROSS items with answer lengths of at least three letters and retained all 1,148 THEMCROSS items satisfying the same criterion. Furthermore, to account for Romanian diacritics and have a uniform evaluation across all models, we removed all diacritical marks from both gold answers and model predictions before evaluation.

Table 1 reports the accuracy obtained on the THEMCROSS and POLYCROSS benchmark datasets. For each prompt variant and LLM we report the solution accuracy measured as defined in the problem formulation. For prompt variants **v3** and **v4** we further vary the length of the prefix given as a hint from one to four letters. Furthermore, for POLYCROSS the average solution length is 5.24 and the median is 5. For THEMCROSS the average is 4.93 and the median is 5, respectively.

Looking at the results obtained on POLYCROSS we observe that the benchmark is highly challenging for all LLMs including the powerful frontier one when no hints or only the answer length are provided. In these conditions the accuracy remains below 10% even for the largest model, Claude Opus 4.8, suggesting that current models often fail to recover the intended hidden meaning of the clues. The task remains difficult when short prefixes are introduced: when only the first two letters are given, gpt-oss-120b and Claude Opus 4.8 achieve 39.7% 50.7% accuracy, respectively. The smaller models perform much worse in this case.

We can see that accuracy improves as the prefix length increases. With the longest prefix of four letters in the hint the accuracy reaches 76.6% for gpt-oss-120b and 85.6% for Claude Opus 4.8 on POLYCROSS. To better understand the contribution of the hint (word length and prefix), note that 45.40% of all solutions in POLYCROSS have no more than four letters. Thus the hint with four letters has enough information to uniquely identify the solution in all those cases. An additional 20.1% of the solutions have a length of 5 in which case the prefix gives 80% of the correct letters. 13.42% of the solutions have a length of 6. All these cases with words of up to 6 letters, where the hint with four letters give two thirds of the required letters or more, add up to 78.83% of all solutions.

Table 2: v3 accuracy (%) by the number of letters the hint leaves hidden. Gray: share of items in the stratum at each k.

		PolyCross				ThemCross			
Items	Model	k=1	k=2	k=3	k=4	k=1	k=2	k=3	k=4
$\lvert a\rvert \leq k$ *(answer fully revealed)*		0%	0%	21%	44%	0%	0%	19%	45%
	Granite-4.0-small	–	–	59.2	57.4	–	–	66.2	69.4
	Llama-3.3-70B	–	–	63.9	73.5	–	–	85.2	86.4
	gpt-oss-120B	–	–	99.7	99.5	–	–	99.1	99.0
	Claude Opus 4.8	–	–	99.1	98.9	–	–	97.2	98.8
	LLMic-v2-3B	–	–	87.4	93.6	–	–	79.6	91.0
$\lvert a\rvert - k \in \{1, 2\}$		21%	44%	45%	35%	19%	45%	48%	40%
	Granite-4.0-small	9.9	17.9	28.3	35.3	30.1	32.4	37.6	47.9
	Llama-3.3-70B	17.4	34.2	47.1	56.9	56.5	62.2	62.7	69.1
	gpt-oss-120B	33.1	50.5	60.1	65.8	63.4	68.6	74.5	80.6
	Claude Opus 4.8	41.6	62.8	80.5	88.1	88.4	92.2	92.6	94.3
	LLMic-v2-3B	2.7	6.1	7.7	6.0	12.5	16.4	17.2	13.5
$\lvert a\rvert - k \geq 3$		79%	56%	34%	21%	81%	55%	33%	15%
	Granite-4.0-small	3.8	7.0	10.2	15.0	17.8	23.0	29.8	27.3
	Llama-3.3-70B	10.9	17.7	25.5	35.5	43.7	45.0	53.3	52.8
	gpt-oss-120B	20.3	31.1	39.1	46.1	55.4	60.9	69.9	72.2
	Claude Opus 4.8	24.9	41.3	55.2	52.9	81.8	85.2	90.5	88.6
	LLMic-v2-3B	1.1	2.3	2.4	1.1	6.2	8.3	10.3	8.0
$\lvert a\rvert > k$ *(informative)*		100%	100%	79%	56%	100%	100%	81%	55%
	Granite-4.0-small	5.1	11.8	20.4	27.8	20.1	27.2	34.4	42.2
	Llama-3.3-70B	12.3	25.0	37.7	49.0	46.1	52.7	58.9	64.6
	gpt-oss-120B	23.0	39.7	51.0	58.5	56.9	64.4	72.6	78.3
	Claude Opus 4.8	28.4	50.7	69.5	75.1	83.0	88.3	91.7	92.8
	LLMic-v2-3B	1.5	3.9	5.4	4.2	7.4	11.9	14.4	12.0

We further find that chain-of-thought (CoT) prompting tends to improve performance but only slightly. This behavior aligns with known limitations of reasoning-based prompting, where models that do not quickly identify a correct solution tend to generate lengthy reasoning traces (i.e., overthinking) that often culminate in incorrect answers Sui et al. (2025).

In contrast the control THEMCROSS dataset is substantially less difficult once hints are introduced, particularly for larger models gpt-oss-120b and Claude Opus 4.8 which achieve 87.5% and 95.5% accuracy for the longest hint. This behavior can be attributed to the fact that the clues typically have a single, unambiguous interpretation, enabling larger models to infer the correct solution more reliably. However, the smaller models like Granite-4.0-small and LLMic-v2-3B still have low accuracy.

Finally, as seen in column **v5** in Table 1, prompting the model to translate the clues and reason in English yields inferior performance, indicating that cross-lingual reasoning introduces additional difficulty for the LLMs, even for the powerful frontier one. Comparing values in column **v5**, we observe that the translation performance on POLYCROSS is much lower than on THEMCROSS. The particularly low accuracy on translating POLYCROSS clues stems not only from limitations in the LLM technology but also from an important characteristic of the dataset. Specifically, many polysemantic clues are difficult or even impossible to translate even for a human expert. Take for example a clue such as "Deschise la culoare" which can mean at the same time "having a light color" (more obvious) or "open on a hallway" (more subtle). The latter points to the solution UȘI (i.e., doors in English). This double meaning is possible in Romanian because word "deschise" can mean both "light" (referring to colors) and "open". Likewise, "culoare" can mean both "color" and "hallways". Clearly, this is very hard, if possible at all, to be translated in English while preserving both meanings. Other cases can be translated more easily as seen with the two examples (involving ROTOR and CIGARETTE) earlier in the paper.

Thus, the low accuracy in polysemantic clue translation is not surprising. Despite this, we argue that the translation experiment presented is important to benchmark the performance compared to translating monosemantic clues, and to motivate future research in translating polysemantic phrases across languages.

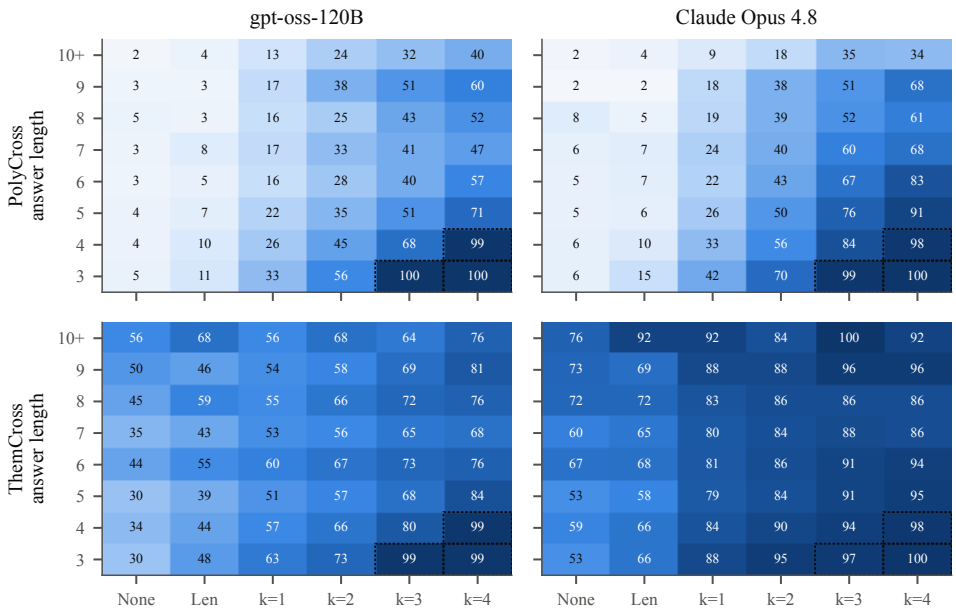


Figure 1: Accuracy (%) by answer length and hint level for the two strongest models.

5.2 Further analysis of model behaviour

We next examine how hints affect model behavior. As shown in Table 1, accuracy increases sharply with the number of revealed letters k. For example, on PolyCross, a four-letter hint improves Claude Opus 4.8's accuracy from 8.7% to 85.6%. To determine whether these gains reflect genuine clue understanding or simply a reduced search space, we stratify instances by the number of unrevealed letters, $|a| - k$, where a is the answer. The results are summarized in Table 2.

Consider instances with fully revealed answers ($|a| \leq k$). At $k = 4$, 44.0% of PolyCross and 44.7% of ThemCross items fall into this category, making the hint itself the correct answer. As expected, Claude Opus 4.8 and gpt-oss-120B achieve near-perfect accuracy (97–100%) on these items. In contrast, Granite-4.0-small and Llama-3.3-70B correctly reproduce the hint only 57–69% and 64–86% of the time, respectively, often generating a longer word that begins with the provided prefix. These errors reflect failures to follow the hint rather than failures of clue reasoning. LLMic-v2-3B copies the hint much more reliably, on 87% ($k = 3$) and 94% ($k = 4$) of fully revealed PolyCross items. However, these cases account for nearly all of its gains: although it reaches 43.6% accuracy at $k = 4$, its accuracy drops to 4.2% on the remaining items, the lowest among all evaluated models.

Restricting evaluation to informative-hint instances ($|a| > k$) reveals how much performance depends on the hint itself. On PolyCross, accuracy at $k = 4$ drops from 85.6% to 75.1% for Claude Opus 4.8 and from 76.6% to 58.5% for gpt-oss-120b. When three or more letters remain hidden, accuracy falls further to 52.9% and 46.1%, respectively. In contrast, applying the same restriction on ThemCross reduces Claude's performance by fewer than three percentage points, indicating that direct clues remain largely solvable even with no hints or partial hints. The much larger degradation on PolyCross suggests that it requires stronger semantic disambiguation and cannot be solved reliably from partial letter information alone, further underscoring the difficulty of the dataset.

Figure 1 further breaks down accuracy by answer length. On PolyCross, performance without prefix hints remains in the single digits across all lengths. As more letters are revealed, shorter answers become substantially easier, while longer answers remain challenging. For example, at $k = 4$, Claude Opus 4.8 solves 91% of five-letter answers but only 34% of answers with ten or more letters. In contrast, answer length has a much smaller effect on performance on ThemCross, indicating that its clues remain relatively tractable even for longer answers.

To gain further insight into model failures, we categorize each incorrect prediction using the first applicable error type: no answer (empty or an output that is a random string), wrong length (the specified answer length is not respected), wrong prefix (the provided letters are not preserved), one

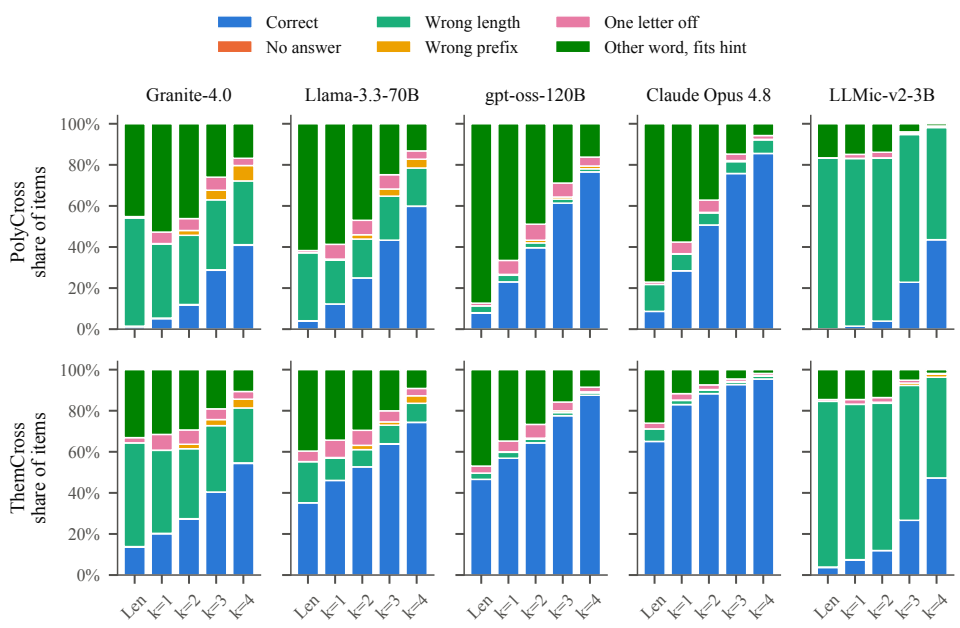


Figure 2: Share of all items that are correct or fall into each error category, per model and hint level.

letter off (edit distance one from the gold answer, considered only when the solution is at least two letters longer than the hint, as otherwise every incorrect completion is trivially one edit away), and other word, fits hint (a completion that matches the hint but is not the target answer).

Figure 2 shows the resulting error breakdown. We notice that hint violations dominate for the weaker models. Granite-4.0-small gives an answer of the wrong length on 31–53% of PolyCross items at every hint level, and Llama-3.3-70B on 19–33%. Among answered items, gpt-oss-120B and Claude Opus 4.8 respect both the length and the prefix 92–99% of the time. LLMic-v2-3B keeps the given prefix on 91–100% of items but the given length on only 17–45% on PolyCross (19–51% on ThemCross): the few-shot completion conveys the letters, but the model does not treat the stated length as a constraint. Wrong length is therefore its dominant error at every hint level. An exception is Claude Opus 4.8 at v2 on PolyCross, which violates the length on 13.1% of items, more than gpt-oss-120B (3.4%): without letters to anchor it, it sometimes prefers a semantically related word of the wrong length. Table 3 in Appendix lists examples of errors encountered during the evaluation.

5.3 Evaluating Rationales for Solutions

So far we reported the LLM solution accuracy on the two datasets with various levels of hinting, including the case with translation. Note that more informative hints limit the space of possible solutions significantly and thus may reduce the need to discover a hidden meaning in the clue. Thus in this section we present results of an additional investigation where we instructed the LLMs to provide a rationale for their solutions which a human expert then rated. To make computing the rationale quality tractable we limit the considerations only to cases where the LLM produced a correct solution.

The score scale used by the human was integers from 0 to 5 as follows. A score of 0 or 1 (category poor) reflects that the rationale has no or little match with the actual reason why the solution is correct. A score of 2 or 3, grouped together into the partially-correct category, indicates a partial match of the rationale given with the actual reason. Scores 4 and 5, grouped into the correct category, indicate a good and very good rationale provided by the LLM.

Figure 3 summarizes the results as a histogram of scores for each dataset and each LLM. The x-axis is the rationale quality on the scale 0-5. For each LLM the percentage on the y-axis for a given score on the x-axis is the percentage of cases with that score. The left histogram is for the baseline ThemCross dataset. All LLMs give very good rationales in a majority of the cases. The percentage of records with a maximum score (i.e., score 5) is 95% for gpt-oss-120b, 93% for Llama-3.3-70b and

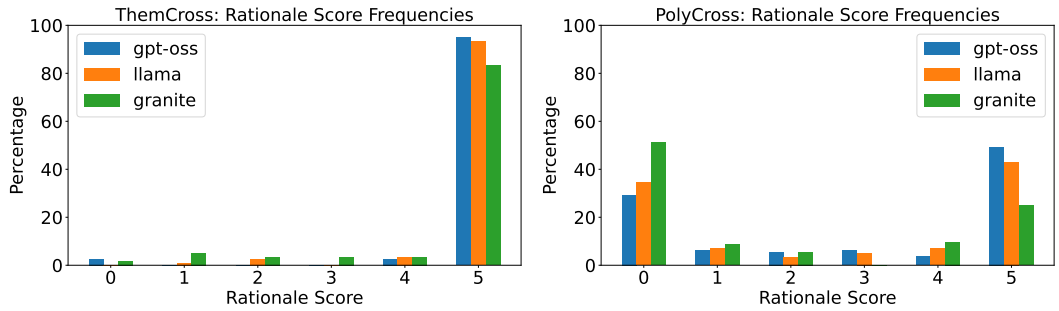


Figure 3: Rationale scores for THEMCROSS baseline (left) and POLYCROSS (right), measured on records where LLMs give correct solutions. High-quality rationales convincingly dominate in the THEMCROSS baseline, whereas performance on POLYCROSS suffers.

83% for Granite-4.0-small. Records with a score of either 4 or 5 (good and very good rationales) add up to 98% for gpt-oss-120b, 96% for Llama-3.3-70b and 86% for Granite-4.0-small.

The right histogram in Figure 3 shows the distribution of scores for POLYCROSS dataset. Here all LLMs frequently have lower rationale scores. To illustrate gpt-oss-120b model gives very good rationales (score 5) only in 49% of cases, compared to 95% for the THEMCROSS dataset. Model Lllama-3.3-70b gives a very good rationale in 43% of cases in POLYCROSS, compared to 93% in THEMCROSS. The rationales obtained with model Granite-4.0-small receive a maximum score in 25% of cases in POLYCROSS, versus 83% in the THEMCROSS dataset.

Furthermore, the score of 0 (very poor rationale) is the case for 29% of all records with gpt-oss-120b in use, 34% of all records with Llama-3.3-70b, and 51% of all records with Granite-4.0-small. Very poor or poor rationales (score 0 or 1) correspond to 35% of cases with gpt-oss-120b, 41% of cases with Llama-3.3-70b, and 60% of cases with Granite-4.0-small. In summary all three LLMs provide good rationale for their solutions for the baseline dataset but have lower rationale scores in the POLYCROSS dataset.

6 CURRENT SHORTCOMINGS AND FUTURE WORK

The results presented above demonstrate interesting insights into LLM's use of hidden meanings in crosswords clues. They invite follow-up research addressing the current shortcomings. For instance, we used only a single self-selected expert and the datasets were in a single language. The expert scored the rationales in a specific order with no consideration of priming effects. Future work will address these shortcomings as well as consider using LLMs for high-quality translation of short phrases with hidden meanings such as crosswords clues in POLYCROSS to other languages. Finally, the low solution accuracy of LLMs on crosswords clues with hidden meanings can possibly be used as a proxy of their polysemantic quality which may enable procedural content generation of high-quality polysemantic clues.

7 CONCLUSIONS

In this work we introduced POLYCROSS, a benchmark dataset of polysemantic crossword clues used to evaluate LLM performance on understanding hidden meanings in short phrases. Unlike standard question-answering benchmarks POLYCROSS challenges LLMs to identify a subtle, non-obvious interpretation of a clue while ignoring a more obvious but misleading one. We evaluated five LLMs including a multi-lingual frontier model by evaluating solution accuracy, the quality of generated rationale, and the quality of translating polysemantic clues. We further evaluated LLM capacity to translate polysemantic phrases while preserving all intended meanings. Our results show that LLMs can find solutions to clues, give rationales and translate clues with a much higher accuracy for monosemantic (i.e., with a single meaning) clues as compared to polysemantic clues. As more powerful LLMs are developed our POLYCROSS dataset will continue serving as a measure of LLMs being able to find solutions to polysemantic clues, translate polysemantic clues in a new language while preserving their multiple meanings, and explain their output, a step towards explainable AI.

AI USE STATEMENT

We used generative AI tools to assist with language editing and grammar correction. The authors take full responsibility for the final content of this work.

REFERENCES

Rebus, a. Monthly print periodical.

Rebus Elixir, b. Monthly print periodical.

Rebus Ideal, c. Monthly print periodical.

Rebus Online. https://www.rebusonline.com, d. Accessed 5 May 2026.

Martin Andrews and Sam Witteveen. A Reasoning-Based Approach to Cryptic Crossword Clue Solving. In *Proceedings of the 42nd International Conference on Machine Learning*, ICML'25. JMLR.org, 2025.

Anthropic. Claude opus (version 4.8). AI Large Language Model, 2026. URL https://claude.ai. Accessed: September 26, 2026.

Vlad-Andrei Bădoiu, Mihai-Valentin Dumitru, Alexandru M. Gherghescu, Alexandru Agache, and Costin Raiciu. Llmic: Romanian foundation language model, 2025. URL https://arxiv.org/abs/2501.07721.

D. Delabastita. *Traductio: Essays on Punning and Translation*. Taylor & Francis, 2016. ISBN 9781134959372. URL https://books.google.hn/books?id=dL_sCwAAQBAJ.

Kathryn J. Friedlander and Philip A. Fine. The Grounded Expertise Components Approach in the Novel Area of Cryptic Crossword Solving. *Frontiers in Psychology*, 7:567, 2016. doi: 10.3389/fpsyg.2016.00567.

Jixuan Leng, Chengsong Huang, Langlin Huang, Bill Yuchen Lin, William W. Cohen, Haohan Wang, and Jiaxin Huang. CrossWordBench: Evaluating the Reasoning Capabilities of LLMs and LVLMs with Controllable Puzzle Generation. *CoRR*, abs/2504.00043, 2025. doi: 10.48550/ARXIV.2504.00043. URL https://doi.org/10.48550/arXiv.2504.00043.

Mayank Mishra, Matt Stallone, Gaoyuan Zhang, Yikang Shen, Aditya Prasad, Adriana Meza Soria, Michele Merler, Parameswaran Selvam, Saptha Surendran, Shivdeep Singh, Manish Sethi, Xuan-Hong Dang, Pengyuan Li, Kun-Lung Wu, Syed Zawad, Andrew Coleman, Matthew White, Mark Lewis, Raju Pavuluri, Yan Koyfman, Boris Lublinsky, Maximilien de Bayser, Ibrahim Abdelaziz, Kinjal Basu, Mayank Agarwal, Yi Zhou, Chris Johnson, Aanchal Goyal, Hima Patel, Yousaf Shah, Petros Zerfos, Heiko Ludwig, Asim Munawar, Maxwell Crouse, Pavan Kapanipathi, Shweta Salaria, Bob Calio, Sophia Wen, Seetharami Seelam, Brian Belgodere, Carlos Fonseca, Amith Singhee, Nirmit Desai, David D. Cox, Ruchir Puri, and Rameswar Panda. Granite Code Models: A Family of Open Foundation Models for Code Intelligence, 2024. URL https://arxiv.org/abs/2405.04324. arXiv:2405.04324 https://arxiv.org/abs/2405.04324.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan

Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. GPT-4 Technical Report, 2024. URL https://arxiv.org/abs/2303.08774. arXiv:2303.08774 https://arxiv.org/abs/2303.08774.

Joshua Rozner, Christopher Potts, and Kyle Mahowald. Decrypting Cryptic Crosswords: Semantically Complex Wordplay Puzzles as a Target for NLP. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2021. Curran Associates Inc. ISBN 9781713845393.

Abdelrahman Sadallah, Daria Kotova, and Ekaterina Kochmar. What Makes Cryptic Crosswords Challenging for LLMs? In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert (eds.), *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 5102–5114, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-main.342/.

Soumadeep Saha, Sutanoya Chakraborty, Saptarshi Saha, and Utpal Garain. Language Models are Crossword Solvers. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 2074–2090. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.naacl-long.104. URL http://dx.doi.org/10.18653/v1/2025.naacl-long.104.

Yuchen Su, Yonghua Zhu, Ruofan Wang, Zijian Huang, Diana Benavides-Prado, and Michael J. Witbrock. A Survey of Pun Generation: Datasets, Evaluations and Methodologies. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025, Suzhou, China, November 4-9, 2025*, pp. 7375–7395. Association for Computational Linguistics, 2025. URL https://aclanthology.org/2025.findings-emnlp.389/.

Yang Sui, Yu-Neng Chuang, Guanchu Wang, Jiamu Zhang, Tianyi Zhang, Jiayi Yuan, Hongyi Liu, Andrew Wen, Shaochen Zhong, Na Zou, Hanjie Chen, and Xia Hu. Stop overthinking: A survey on efficient reasoning for large language models, 2025. URL https://arxiv.org/abs/2503.16419.

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. arXiv:2302.13971 https://arxiv.org/abs/2302.13971.

Zhijun Xu, Siyu Yuan, Lingjie Chen, and Deqing Yang. "A Good Pun Is Its Own Reword": Can Large Language Models Understand Puns? In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 11766–11782, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.657. URL https://aclanthology.org/2024.emnlp-main.657/.

Alessandro Zangari, Matteo Marcuzzo, Andrea Albarelli, Mohammad Taher Pilehvar, and Jose Camacho-Collados. Pun Unintended: LLMs and the Illusion of Humor Understanding. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 27936–27971, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1419. URL https://aclanthology.org/2025.emnlp-main.1419/.

A Appendices and Supplementary Material

A.1 Prompts

In this section, we present the prompts used in our experiments.

Prompt Template

You are an expert at solving crossword puzzles. Given a clue in Romanian (i.e., a short definition) you will provide the answer to that clue.
Rules:
- The answer to the given clue must be in Romanian.
- The clue definition may have multiple meanings and you must select the most likely answer.
- Do not translate verbatim the clue definition in English.
- You must analyze carefuly the clue to determine the most likely meaning for your answer.
- Provide the rationale behind your answer.
- Format the response as a JSON object with the following structure.
"'json "answer": "<your answer here>", "rationale": "<your rationale here>" "'
Use the following examples to learn your task better.
Example 1: CLUE: Prins asupra faptului ANSWER: "'json "answer": "OCUPAT", "rationale": "Prins asupra faptului inseamna ca persoana respectiva era ocupata cu o anumita activitate la momentul respectiv. Deci, OCUPAT este cel mai probabil raspuns." "'
Example 2: CLUE: Marcată de o purtare abuzivă ANSWER: "'json "answer": "ROASA", "rationale": "De exemplu, purtarea abuziva a unei perechi de pantofi implica a uzura pronuntata a pantofilor. Roasa inseamna uzat sau uzura." "'
Example 3: CLUE: Unitate de morărit ANSWER: "'json "answer": "SAC", "rationale": "Activitatea de morărit implica macinarea graului in faina. De obicei, faina rezultata este pusa intr-un SAC." "'
CLUE: <clue text>
ANSWER:

Figure 4: Prompt version v1. Few-shots examples and no additional hints provided.

Prompt Template

You are an expert at solving crossword puzzles. Given a clue in Romanian (i.e., a short definition) you will provide the answer to that clue. As a hint, you are given the number of letters that the answer must have.
Rules:
- The answer to the given clue must be in Romanian.
- The length of your answer must match the number given as hint.
- Do not use abbreviations.
- The clue definition may have multiple meanings and you must select the most likely answer.
- Do not translate verbatim the clue definition in English.
- You must analyze carefuly the clue to determine the most likely meaning for your answer.
- Provide the rationale behind your answer.
- Format the response as a JSON object with the following structure.
“‘json "answer": "<your answer here>", "rationale": "<your rationale here>" “‘
Use the following examples to learn your task better.
Example 1: CLUE: Prins asupra faptului HINT: 6 ANSWER: “‘json "answer": "OCUPAT", "rationale": "Prins asupra faptului inseamna ca persoana respectiva era ocupata cu o anumita activitate la momentul respectiv. Deci, OCUPAT este raspunsul corect pentru ca are 6 litere." “‘
Example 2: CLUE: Marcată de o purtare abuzivă HINT: 5 ANSWER: “‘json "answer": "ROASA", "rationale": "De exemplu, purtarea abuziva a unei perechi de pantofi implica a uzura pronuntata a pantofilor. ROASA inseamna uzat sau uzura. Mai mult, ROASA este raspunsul corect pentru ca are 5 litere." “‘
Example 3: CLUE: Unitate de morărit HINT: 3 ANSWER: “‘json "answer": "SAC", "rationale": "Activitatea de morărit implica macinarea graului in faina. De obicei, faina rezultata este pusa intr-un SAC iar cuvantul SAC are 3 litere." “‘
CLUE: clue text
HINT: number of letters in answer
ANSWER:

Figure 5: Prompt version v2. Few-shots examples and the number of letters provided as additional hint.

Prompt Template

You are an expert at solving crossword puzzles. Given a clue in Romanian (i.e., a short definition) you will provide the answer to that clue. As a hint, you are given the number of letters that the answer must have.
Rules:
- The answer to the given clue must be in Romanian.
- The length of your answer must match the number given as hint.
- Do not use abbreviations.
- The clue definition may have multiple meanings and you must select the most likely answer.
- Do not translate verbatim the clue definition in English.
- You must analyze carefuly the clue to determine the most likely meaning for your answer.
- Provide the rationale behind your answer.
- Format the response as a JSON object with the following structure.
“‘json "answer": "<your answer here>", "rationale": "<your rationale here>" “‘
Use the following examples to learn your task better.
Example 1: CLUE: Prins asupra faptului HINT: 6 ANSWER: “‘json "answer": "OCUPAT", "rationale": "Prins asupra faptului inseamna ca persoana respectiva era ocupata cu o anumita activitate la momentul respectiv. Deci, OCUPAT este raspunsul corect pentru ca are 6 litere." “‘
Example 2: CLUE: Marcată de o purtare abuzivă HINT: 5 ANSWER: “‘json "answer": "ROASA", "rationale": "De exemplu, purtarea abuziva a unei perechi de pantofi implica a uzura pronuntata a pantofilor. ROASA inseamna uzat sau uzura. Mai mult, ROASA este raspunsul corect pentru ca are 5 litere." “‘
Example 3: CLUE: Unitate de morărit HINT: 3 ANSWER: “‘json "answer": "SAC", "rationale": "Activitatea de morărit implica macinarea graului in faina. De obicei, faina rezultata este pusa intr-un SAC iar cuvantul SAC are 3 litere." “‘
CLUE: clue text
HINT: number of letters in answer
ANSWER:

Figure 6: Prompt version v3. Few-shots examples as well as the number of letters and the prefix provided as additional hints.

Prompt Template

You are an expert at solving crossword puzzles. Given a clue in Romanian (i.e., a short definition) you will provide the answer to that clue. As a hint, you are given the number of letters that the answer must have and the letters the answer must start with. You must use both hints to find the answer. You must also think step by step and provide the rationale behind your answer, explaining how you used the hints to find the answer.
Rules:
- The answer to the given clue must be in Romanian.
- The answer must have the number letters given as hint and must start with the letters given as hint.
- Do not use abbreviations.
- The clue definition may have multiple meanings and you must select the most likely answer.
- Do not translate verbatim the clue definition in English.
- You must analyze carefuly the clue to determine the most likely meaning for your answer.
- Provide the rationale behind your answer, explaining how you used the hints to find the answer.

When thinking step by step, you should first analyze the clue and provide a list of possible answers that match the clue definition, without using the hints. Then, you should use the hints to narrow down the list of possible answers and find the most likely answer. Finally, you should provide the rationale behind your answer, explaining how you used the hints to find the answer.
Mark your thinking process with the following tags: <think> ** your thoughts ** </think>
Your final answer should be the one that best matches the clue definition and satisfies the hints.
Wrap your final answer between square brackets like this: [your final answer here]
CLUE: clue text
HINT: Raspunsul are number of letters litere si incepe cu prefix text

Figure 7: Prompt version v4. Chain-of-Thought prompting without translating the clue in English with hints.

Prompt Template

You are an expert at solving crossword puzzles. Given a clue in Romanian (i.e., a short definition) you will provide the answer to that clue. You must also think step by step and provide the rationale behind your answer. You will be given a single hint that states the number of letters that the answer must have.
Rules:
- The clue definition may have multiple meanings and you must select the most likely answer.
- You must translate the clue definition in English and reason in English.
- You must analyze carefuly the clue to determine the most likely meaning for your answer.
- Do not use abbreviations.
- Provide the rationale behind your answer, explaining how you found the answer.
- Finally, you must provide the answer in Romanian, wrapped between square brackets like this: [your final answer here].

When thinking step by step, you should first analyze the clue and provide a list of possible answers that match the clue definition, without using the hints. Then, you should use the hints to narrow down the list of possible answers and find the most likely answer. Finally, you should provide the rationale behind your answer, explaining how you used the hints to find the answer.
When you generate the possible candidates, you should generate them in English, even if the final answer must be in Romanian. This is to ensure that you can reason about the meaning of the clue and the possible answers in a more precise way, without being limited by your Romanian vocabulary. After you have generated the possible candidates in English, you can then translate them to Romanian and check which ones satisfy the hint about the number of letters. Finally, you can select the most likely answer in Romanian and provide the rationale behind it.
Mark your thinking process with the following tags: <think> ** your thoughts ** </think>
Your final answer should be the one that best matches the clue definition and satisfies the hints.
Wrap your final answer between square brackets like this: [your final answer here]
CLUE: clue text HINT: The answer has number of letters letters.

Figure 8: Prompt version v5. Chain-of-Thought prompting with translating the clue in English.

Table 3: Wrong answers at v3 k=2 on POLYCROSS, sampled at random per error category.

Clue	Answer	Prediction	Model
One letter off			
Ultima rezervă de materie cenuşie	URNA	URMA	Claude Opus 4.8
Rezultatul adunării	ATOT	ATOM	Claude Opus 4.8
Animal... pentru tăiat de ocazie	LAMA	LADA	Claude Opus 4.8
Implicată într-o afacere în pierdere	DATA	DANA	gpt-oss-120B
Impun măsuri restrictive pentru limitarea debordărilor	LIMITATIVI	LIMITATIVE	gpt-oss-120B
O fată care stă la tușă	VARA	VACA	gpt-oss-120B
Other word, fits hint			
A avansa de la echipa mare	TRIA	TREC	Claude Opus 4.8
Active pe perioada alegerilor	SITE	SIGL	Claude Opus 4.8
Oglindiți în lumina adevărului	REALI	REDAT	Claude Opus 4.8
Pornire cu violenţă	ARTAG	ARUNC	gpt-oss-120B
Ataşat naval	VAS	VAL	gpt-oss-120B
Nimb prețios pentru sanctitatea iubirii (3 cuv.)	INELDEAUR	IN INIMA TA	gpt-oss-120B
Wrong length			
Iarba verde unde se iese cu grupul	PASCATOARE	PAJISTELE	Claude Opus 4.8
Dă semne că se face rău	AVERTIZARE	AVERTIZEA	Claude Opus 4.8
Disimularea unui om cu pretenții exagerate	INFUMURARE	INFATUARE	Claude Opus 4.8
Tensiunea instalată pentru senzori de vibraţii	NELINISTE	NECESITA	gpt-oss-120B
Instalat din start printre favoriți	ATU	ATOP	gpt-oss-120B
Atac la persoană	MISTOCARIE	MISOGENIE	gpt-oss-120B
Wrong prefix			
Substratul unui adevăr... pus la îndoială	VINCUSIFON	VERIDICITA	Claude Opus 4.8
Precedent pentru noi	NEO	NOU	Claude Opus 4.8
Odoare iubite de părinți (sing.)	ORAR	ODOR	Claude Opus 4.8
Pune bazele înălţării	CTITOR	CIMENT	gpt-oss-120B
Rezultatul adunării	NOIAN	NUMAR	gpt-oss-120B
Concept generic de acțiune în scădere	IEFTINIRE	DIMINUARE	gpt-oss-120B