Vinclat: Evaluating Reasoning, Cognition and Culture in One Game

Marc Pàmies **Javier Aula-Blasco** **Aitor Gonzalez-Agirre** **Marta Villegas**
Barcelona Supercomputing Center
`mpamies@bsc.es`

Abstract

This paper introduces *Vinclat*, a novel evaluation dataset for Catalan carefully designed to assess the reasoning capabilities and cultural knowledge of LLMs. It comprises 1,000 high-quality instances, meticulously crafted and reviewed by human annotators. Each instance presents a complex riddle that requires a two-step reasoning process involving inferential and abductive reasoning, along with other cognitive skills such as lexical retrieval, paraphrasing, flexibility in interpretation, pattern recognition, and associative thinking. Given four independent clues, models should infer intermediate concepts which, despite being seemingly unrelated, can be creatively connected to reach a final solution. The task targets a unique blend of capabilities, distinguishing it from existing NLP benchmarks. Our evaluation of state-of-the-art models reveals that these still fall significantly short of human-level reasoning, although scaling trends suggest that the performance gap may narrow over time. This indicates that *Vinclat* provides a robust and long-term challenge, resisting the rapid saturation that is commonly observed in many existing evaluation datasets.

1 Introduction

In recent years, large language models (LLMs) have made groundbreaking progress in natural language processing (NLP), achieving state-of-the-art performance in tasks ranging from straightforward factual knowledge (Krathwohl, 2002) question answering (Mihaylov et al., 2018; Petroni et al., 2021), to more complex tasks at the human cognitive level such as multi-step reasoning (Cobbe et al., 2021; Srivastava et al., 2023). More recently, increasing attention has been directed towards enhancing the multilingual reasoning capabilities of these models, as demonstrated by recent releases that highlight the significant impact of improved reasoning in solving complex problems across languages (Guo et al., 2025; Zheng et al., 2025).

Despite these advances, assessing further complex reasoning capabilities entangled with deep contextual knowledge remains a significant challenge, particularly in creative and multilingual contexts, where conventional benchmarks often fall short (Ghosh et al., 2025; Shojaee et al., 2025). There is a notable scarcity of downstream tasks that have been explicitly crafted to evaluate the equivalent to complex human cognitive skills in LLMs, leading to the widespread use of coding (Gu et al., 2024; Zhuo et al., 2024) and mathematical (Mishra et al., 2022; Glazer et al., 2024) datasets as proxies. While these datasets can provide some insight, they are insufficient to comprehensively capture the nuanced and multifaceted reasoning abilities required for broader cognitively complex tasks. This gap underscores the need for a more diverse suite of reasoning-intensive NLP tasks that pave the way for holistic evaluation frameworks, which is critical to advance LLMs towards more human-like intelligence.

We believe that problem-solving datasets derived from games can serve as effective tools to evaluate progress in this area. Despite being often regarded as mere entertainment, games have a long history of being used as valuable test-beds for artificial intelligence research (Hu et al., 2024; Silver et al., 2016, 2017; Vinyals et al., 2019). This is primarily due to the fact that games offer a safe, controlled environment that simulates real-world scenarios and can often be generalized to more complex domains. Moreover, the well-defined rules and objectives inherent in games make it relatively straightforward to assess success and measure performance. Building on this belief, we introduce *Vinclat*[1], a Catalan-language dataset for multi-step problem solving that employs a game-based structure to evaluate both the reasoning capabilities of LLMs and the depth of their cultural awareness.

---

[1] `https://hf.co/datasets/projecte-aina/vinclat`

Proceedings of the First Workshop on Multilingual Multicultural Evaluation, pages 49–66
March 28, 2026 ©2026 Association for Computational Linguistics

2 The Task

2.1 Task Definition

The main objective is to guess a target word or term (*solution*) based on a set of *keywords* that are obtained from four given *hints* (see Figure 1). Notably, the length of the target word(s) is provided in advance, which introduces a structural constraint that guides the reasoning process. However, successfully solving the task is more complex than it might appear at first glance, as it involves two distinct steps, each requiring different types of reasoning abilities (see §2.2).

- **Step 1.** First, a plausible term must be inferred for each hint. These intermediate terms, or *keywords*, are not constrained by a fixed length, and it is not strictly necessary to identify all of them correctly to arrive at the final solution (see Table 1).
- **Step 2.** In this phase, the set of seemingly unrelated terms obtained in Step 1 must be jointly interpreted to uncover a hidden semantic or conceptual link. This requires synthesizing the intermediate answers into a single unifying term that captures their commonality, often involving associative and categorical reasoning (see Table 2).

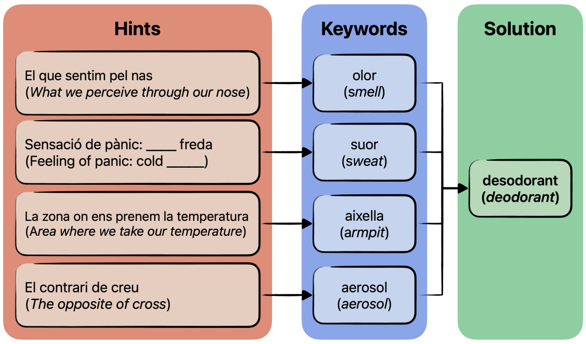
Hints
Keywords
Solution
El que sentim pel nas
(What we perceive through our nose)
Sensació de pànic: ____ freda
(Feeling of panic: cold _____)
La zona on ens prenem la temperatura
(Area where we take our temperature)
El contrari de creu
(The opposite of cross)
olor
(smell)
suor
(sweat)
aixella
(armpit)
aerosol
(aerosol)
desodorant
(deodorant)

Figure 1: Example of a *Vinclat*. Here, only the *Hints* and number of letters of the *Solution* are provided. The *Keywords* and the actual *Solution* are to be guessed.

2.2 Involved Cognitive Abilities

Solving a *Vinclat* requires a broad set of human-like cognitive skills. LLMs evaluated with *Vinclat* need to be able to jointly demonstrate all of these capabilities in order to successfully complete the task. First, players need to access precise lexical items from memory, especially when clues are vague or indirect —a process known as lexical retrieval (Levelt, 1989), often mirrored in factual recall benchmarks like LAMA (Petroni et al., 2019) or MMLU (Hendrycks et al., 2020). Beyond direct recall, *Vinclat* challenges players to draw inferences. For instance, inferring that *“fa llum però no crema”* (it lights up but doesn’t burn) refers to an “LED bulb” requires bridging gaps between literal meaning and plausible explanation, similar to what ARC (Clark et al., 2018) or StrategyQA (Geva et al., 2021) test in LLMs. These clues often follow implicit patterns such as recognizing that *“martell”* (hammer), *“trepant”* (drill), *“tornavís”* (screwdriver) and *“serra”* (saw) all belong to the category *“eines”* (tools), demanding a form of analogy-making and semantic pattern recognition

Hint	**Keyword**	**Explanation**
La filla de la paciència *(The daughter of patience)*	ciència *(science)*	Based on the Catalan proverb *“La paciència és la mare de la ciència”* (which translates to “patience is the mother of science”), it should be inferred that the daughter of *patience* is *science*.
Roma cap per avall *(Rome upside down)*	amor *(love)*	The word *“Roma”* (Rome) should be read backwards as *“amor”* (Love), despite not being explicitly stated in the hint.
El 25 de desembre, ___, ___, ___ *(On December 25, ___ ___ ___)*	fum *(smoke)*	Successfully deciphering this fill-in-the-blank hint requires knowledge of the lyrics of a traditional Catalan Christmas carol.
De sol, digital o d’agulles *(sun, digital or needle)*	Rellotge *(clock)*	All three words refer to types of clocks.

Table 1: Examples showcasing the reasoning involved in *Vinclat* ’s first step.

Keywords	Solution
Reina (*queen*) Egipte (*Egypt*) Llet (*milk*) Serp (*snake*)	Cleopatra
Arbre (*tree*) Bombardeig (*bombing*) Quadre (*painting*) País Basc (*Basque Country*)	Guernica

Table 2: Examples illustrating the associative and categorical reasoning required in *Vinclat* 's second step.

that benchmarks like ConceptARC (Lörer et al., 2023) and some BIG-bench sub-tasks (Srivastava et al., 2022) aim to capture. *Vinclat* also encourages players to mentally test possible solutions before committing: this trial-and-error process reflects hypothesis generation and testing (Klahr and Dunbar, 1988), a capability explored in datasets such as GSM8K (Cobbe et al., 2021) and CLUTRR (Sinha et al., 2019). Interestingly, interpreting figurative or idiomatic clues calls for metalinguistic awareness, the capacity to analyze language abstractly (Bialystok, 1986), a skill also critical for disambiguation tasks like the Winograd Schema Challenge in SuperGLUE (Wang et al., 2019). Lastly, when one or more clues are unclear, *Vinclat* players often rely on semantic guessing —filling in gaps based on context and heuristics, a behavior echoed in masked language models like BERT (Devlin et al., 2019) and rooted in classic work on memory interpolation (Bartlett, 1932). All these cognitive mechanisms reflect a rich interplay of language, reasoning, and abstraction that makes *Vinclat* a considerably challenging task.

3 The Dataset

The *Vinclat* dataset contains 1,000 instances that were carefully conceived by native Catalan speakers. As mentioned, every instance entails four independent hints, each pointing to an intermediate term or keyword. The model must first infer these intermediate terms and then uncover a final solution term, whose number of letters and words is provided as an additional hint. The final solution is designed so that its specified word and letter pattern is unique. Other plausible answers may exist, but none match the given structure.

3.1 Data Creation and Curation

The dataset was created by the original authors of the Vinclat game, who released a new instance per day on the official website[2]. The authors met on a weekly basis in collaborative sessions to produce the puzzles for the upcoming period. During these sessions, they aimed to:

- ensure topical diversity, avoiding repetition across instances,
- vary difficulty levels by including at least one relatively easy clue to support solvability,
- maintain linguistic and semantic coherence while encouraging creative associations.

We manually revised the original dataset to ensure that all instances can be interpreted in isolation, improving the fairness and reproducibility of the task. During this curation, we revised instances that were time-dependent or game-specific, as these relied on information unavailable to both LLMs and human annotators (see §4.5). For example, *"Les festes que gaudirem en un mes"* ("The holidays we'll enjoy in a month") was rewritten as *"Les festes que gaudirem el mes de desembre"* (The holidays we'll enjoy in December) to remove the need for knowing the date. Similarly, clues like *"La resposta d'ahir"* (Yesterday's answer) —which originally referred to *"cervell"* (brain)— were replaced with self-contained alternatives such as *"La part més gran de l'encèfal humà"* (The largest part of the human encephalon). This handcrafted nature of the dataset, along with its careful attention to cultural relevance and cognitive challenge, makes it a unique and high-quality resource for evaluating reasoning and association capabilities in LLMs.

Crucially, the dataset was never published in bulk or made publicly available in machine-readable format prior to this study. The only access point to individual instances was through the daily publication of a single puzzle on the website, which limits their presence in large-scale training corpora typically used for LLM pre-training. Since December 2nd, 2024, no new instances have been added, and all previously published puzzles have been removed from the website, making them unavailable online. Furthermore, the highly localized and manually authored nature of the dataset —including its use of Catalan cultural references, idiomatic expressions, and unpublished formulations— makes

---

[2] `https://vinclat.cat`

it extremely unlikely that LLMs have been exposed to it or anything similar during pre- or post-training. For these reasons, we consider the risk of data contamination to be negligible. To preserve long-term validity, we release only instance-level hints and maintain an external leaderboard[3] where models are evaluated and added over time.

3.2 Content Domains and Cultural Context

The dataset covers a wide range of topics to ensure both variation in difficulty and broad appeal. Hints and solutions touch upon areas such as arts and culture (e.g., theatre, literature, music, cinema), general knowledge (e.g., science, history, geography), daily life and common sense associations, politics and society, philosophical or abstract concepts, famous personalities, religion and spirituality, and Catalan popular sayings, idioms, traditions and folklore, among others. This topical diversity contributes to the complexity and richness of the task, as it requires both semantic flexibility and cultural awareness.

Although the dataset is deeply rooted in the Catalan cultural and linguistic context, not all instances rely exclusively on local knowledge. Many hints reflect Catalonia's traditions, idiomatic expressions, social references, and shared historical memory, which require a certain level of cultural grounding for successful resolution. At the same time, the dataset deliberately includes hints from more global or universal domains. This careful balance between local and global content reflects an intent to design a game that is culturally situated yet broadly accessible, showcasing the linguistic and conceptual nuances of Catalan while remaining solvable by educated native speakers. As a result, *Vinclat* provides a particularly rich resource for evaluating the interplay between language, culture, and reasoning in LLMs.

4 Evaluation

4.1 Baselines

As this task involves multi-step reasoning guided by explicit instructions, it fundamentally relies on the model's ability to understand and follow natural language directives. Recent findings suggest that *base* models (i.e., pre-trained solely via next-token prediction) often fail on instruction-heavy datasets, particularly when answers must conform to specific formats or require tool-like reasoning behaviors (Liang et al., 2023; Zhou et al., 2023). As such, we do not evaluate *base* models, as they have not been optimized for instruction-following behavior.

For baseline evaluation, we select the top-performing open-source models with demonstrated reasoning capabilities, as preliminary experiments revealed that non-reasoning LLMs, in spite of being instructed, yield remarkably low results. In particular, we evaluate the performance of DeepSeek-V3-0324 (Liu et al., 2024), DeepSeek-R1 (Guo et al., 2025) and the Qwen3 family of models (Zheng et al., 2025). The latter offers an especially compelling case study, as it allows us to analyze if performance scales with model size. Furthermore, Qwen3's ability to turn thinking on and off at will provides a controlled setup to quantify the benefits of forcing an LLM to deliberate before answering. The inclusion of DeepSeek-V3-0324, which can be considered the non-reasoning counterpart to DeepSeek-R1, allows us to investigate the performance gap attributable to reasoning capabilities within the same architecture. As new models become available, we plan to evaluate the most promising ones and publish their results on a public leaderboard.

4.2 Implementation Details

Evaluated models were served as endpoints using the vLLM library (v.0.8.5),[4] enabling seamless interaction with an OpenAI-compatible server. This setup allowed the models to function as if they were part of a hosted API service, facilitating inference requests in a standardized format compatible with the `openai` python library (v.1.60.2).[5] The compute nodes in the cluster used for evaluation are equipped with four 64GB H100 GPUs, which results in a limited VRAM capacity to accommodate the largest models in a single node. To address this, we use the dockerized version of vLLM[6] for distributed inference. Specifically, in such cases, four nodes were employed with both the tensor and pipeline parallel sizes set to 4, ensuring efficient distributed processing (Kwon et al., 2023).

For generation parameters, we adhered to the default configurations recommended by model providers, thereby eliminating the need for extensive parameter tuning. Regarding the number of generation tokens, the limit was dynamically set to

---

[3] `https://hf.co/spaces/projecte-aina/vinclat-leaderboard`

[4] `https://github.com/vllm-project/vllm`

[5] `https://pypi.org/project/openai/`

[6] `https://hub.docker.com/r/vllm/vllm-openai`

the model's context window size minus the length of the largest prompt in the batch, providing the LLM with as many tokens as possible to carry out its reasoning process. It is worth noting that, given the nature of the dataset, prompt lengths exhibit minimal variation and are significantly shorter —around 300 tokens— than the available context window (i.e., 163,840 tokens for DeepSeek and 40,960 for Qwen3). The inference and evaluation scripts are fully available to the public.[7]

4.3 Prompt Templates

We designed two prompt templates that differ in how they convey the structure of the target solution: a fill-in-the-blanks format and a letter-count format. Each instance was evaluated using both templates, in two languages, Catalan and English, resulting in four total prompt variants per instance (see Appendix A). This allows us to analyze model performance across different representations of structural constraints and languages of instruction, while keeping the content (the four hints and the solution) consistently in Catalan.

Each prompt begins with the same task description, which provides models with a clear breakdown of the two-stage reasoning process required by the game. The prompt then diverges into one of two formats:

- **Letter-count.** This variant presents the expected structure of the solution in natural language (e.g., "The solution has 2 words with the following structure: 2 letters for the first word and 5 letters for the second word."). This evaluates whether the model can reason over symbolic constraints and integrate them with thematic associations from the hint words.

- **Fill-in-the-blanks.** This *cloze-style* variant encodes the structure of the solution using underscore placeholders for each letter in the solution. For example, for a two-word solution with two letters in the first word and three in the second, the prompt would include "The solution should fit here: __ ___" (i.e., two underscores followed by a whitespace and three underscores). This evaluates whether the model can interpret and use visual or subword-level cues to generate appropriate candidate solutions.

Using both formats allows us to assess whether models rely on abstract understanding of word constraints (e.g., word counts and lengths in text) versus those that leverage pattern recognition or visual structure matching (Bubeck et al., 2023; Li et al., 2024). As mentioned, both prompt formats were used in Catalan and English. This design choice enables us to evaluate models' ability to follow instructions in Catalan, a relatively low-resource language, while also isolating the effect of prompt language on performance. Given that many large language models are primarily post-trained in English, this comparison allows us to examine how language choice in prompting influences model behavior. At the same time, it reflects realistic usage scenarios in multilingual regions, where users may interact with systems in more than one language or receive instructions in one language while processing content in another.

4.4 Scoring Criteria

An instance is considered to be correctly solved if and only if the model produces the final solution, regardless of the correctness and number of keywords identified in the intermediate step. The rationale for not penalizing keywords is that they are not necessarily fixed terms; variations or synonyms can also contribute to a successful solution. Additionally, no penalty is applied for failing to identify all four keywords, as omitting unclear or misleading clues can be advantageous. In such cases, disregarding a specific hint is preferable to selecting an incorrect keyword that could divert the solution process away from the right final answer.

The final part of the prompt specifies that the response must be returned in JSON format, making it easier to parse the LLM's output. After extracting the dictionary —which, in reasoning models, appears immediately after the `</think>` tag— the correctness of the solution was verified using exact match. All responses were lowercased prior to validation in order to ensure that correct answers were not penalized due to differences in letter casing.

4.5 Human Baseline

To establish a human baseline for the task, we conducted a targeted human evaluation on a sample of 300 instances randomly drawn from the full dataset. This subset was double-checked and confirmed to balance feasibility and diversity while capturing a representative cross-section of the task's difficulty and content variety.

---

[7] `https://github.com/mapama247/vinclat_eval`

Model	EN_LC	EN_FB	CA_LC	CA_FB	Avg.	UNKs
Without Reasoning						
Qwen3-0.6B	0	0	0	0	**0**	2,575
Qwen3-1.7B	0	0.1	0	0.1	**0.05**	98
Qwen3-4B	0.2	0.3	0.2	0.4	**0.28**	2.25
Qwen3-8B	0.7	0.5	0.6	1.0	**0.7**	0
Qwen3-14B	0.7	1.4	1.9	1.4	**1.35**	13.25
Qwen3-32B	1.3	2.0	1.8	2.0	**1.78**	0.25
Qwen3-30B-A3B	1.3	1.3	1.2	1.4	**1.3**	3.75
Qwen3-235B-A22B	11.5	11.0	7.6	6.3	**9.1**	9.75
DeepSeek-V3-0324	13.1	7.6	9.4	8.3	**9.6**	29.25
With Reasoning						
Qwen3-0.6B	0.1	0	0	0.1	**0.05**	767.5
Qwen3-1.7B	0.4	0	0.1	0.1	**0.15**	402.25
Qwen3-4B	0.7	0.5	1.0	0.8	**0.75**	28.25
Qwen3-8B	2.6	1.0	1.7	1.2	**1.63**	122.25
Qwen3-14B	4.2	3.1	2.5	2.3	**3.03**	279.5
Qwen3-32B	7.0	5.7	6.5	5.2	**6.1**	98.67
Qwen3-30B-A3B	2.9	1.8	2.9	1.9	**2.38**	125
Qwen3-235B-A22B	22.3	13.4	21.2	13.0	**17.48**	103.5
DeepSeek-R1	33.0	22.7	33.5	23.8	**28.25**	109

Table 3: Accuracy per model. "EN" and "CA" refer to the prompt language; "LC" refers to the letter-count format and "FB" to fill-in-the-blanks. Column "UNKs" refers to the amount of keywords tagged as "unknown" by the model, averaged across prompts.

We recruited three human annotators to each complete all 300 selected instances under different conditions designed to simulate varying levels of resource availability and knowledge access:

- **Annotator A** completed the task without access to the internet, relying solely on personal knowledge and reasoning.
- **Annotator B** had full access to the internet, simulating an open-book approach to answering each clue and inferring the final keyword.
- **Annotator C** was allowed to use the internet for 50% of the instances (150 out of 300). These instances were selected randomly and balanced across the sample to mitigate potential order or topic bias.

This design allows us to assess not only the upper-bound performance of knowledgeable human participants with external support, but also the limits of unaided reasoning, and the mixed conditions in between. All annotators were native Catalan speakers, born and raised in Catalonia, and all held university degrees (with some having completed postgraduate studies). This ensured both linguistic and cultural familiarity with the game and its underlying associations. Annotators were asked to leave instances unanswered if they failed to find a solution within 5-7 minutes of trying, as this avoided prolonged effort on especially tricky instances.

Human performance varied substantially depending on internet access. In the offline condition, annotators answered on average 61% of the instances, with an accuracy of 82% on answered items, yielding an overall accuracy of 50% when accounting for unanswered cases. In the online condition, response rates increased to an average of 84%, with an accuracy of 85% on answered items and an overall accuracy of 71%. These results show that access to online resources markedly improves both coverage and total accuracy, primarily by reducing the number of skipped items.

To quantify the uncertainty of these estimates, we computed 95% confidence intervals for the proportion of correct responses using the standard normal approximation for proportions:

$$CI = \hat{p} \pm Z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

where $\hat{p}$ is the observed accuracy, n the number of instances annotated, and $Z = 1.96$ corresponds to a 95% confidence level (Agresti and Coull, 1998). The resulting intervals are 47-59.4% for offline, and 66.9-75.5% for online, supporting that the measured performance provides a statistically robust estimate of human capabilities on the full dataset. It is worth noting that all original *Vinclat* instances were confirmed to be solvable by the original authors, as records from the online game show that there was always a subset of players who successfully solved each instance.

To complement the quantitative evaluation, a human annotator different to the three above but with a similar profile conducted a qualitative error analysis of model outputs. For each model-language-prompt configuration, a minimum of 25 incorrectly solved instances were randomly selected. The annotator was allowed to explore more instances for any configuration if they deemed it necessary to complete their analysis, totaling over 1,800 cases analyzed. The annotator examined the reasoning traces (when available) and final answers, identifying recurrent error types such as hint misinterpretation, linguistic or cultural misunderstanding, hallucination, and structural inconsistency. This manual review provided fine-grained insights into the nature of model failures and supported the interpretation of quantitative trends discussed in §5.

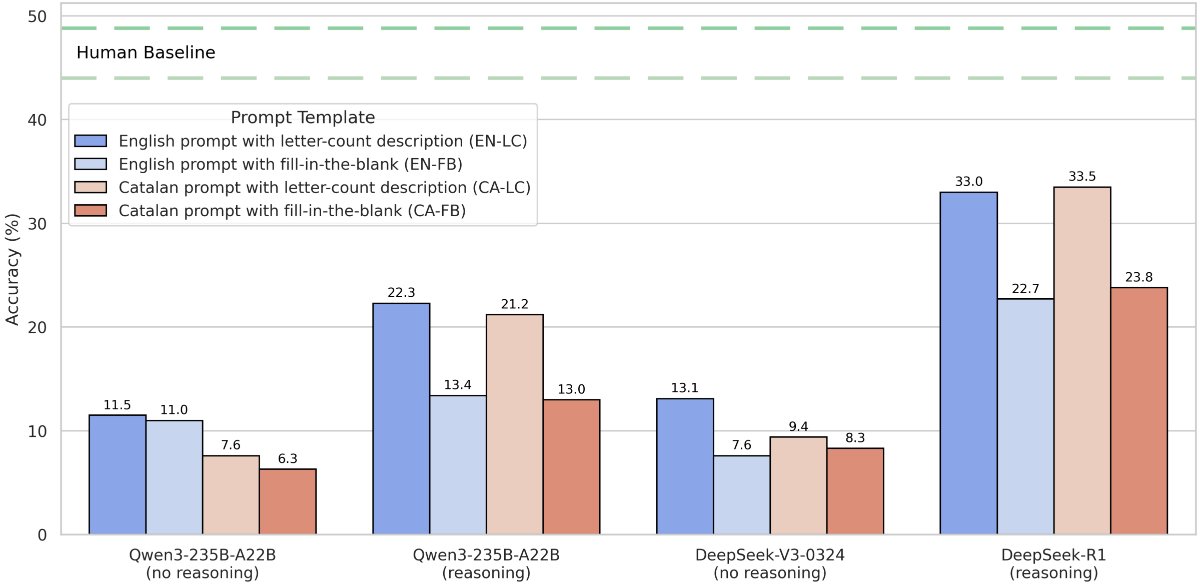

Figure 2: Accuracy scores based on prompt format and reasoning capabilities.

5 Results and Discussion

Comparison to human baselines Across all evaluations, large language models perform significantly worse than human annotators, regardless of internet access (see Table 3 and Figure 2). The highest-performing model only reaches 33.5% accuracy, while the majority of evaluated models struggle to exceed 20%, and several models remain below 10%. In contrast, human annotators, both with and without full access to online resources, performed considerably better, confirming that human-level commonsense reasoning and cultural background knowledge —especially in a task deeply rooted in local culture— remain out of reach for current LLMs. This mirrors findings in other abductive or commonsense-heavy tasks, where humans considerably outperform models (Bhagavatula et al., 2020; Loredo Lopez et al., 2025), suggesting that such deficits are not merely language- or culture-specific but reflect broader limitations in the reasoning mechanisms of current large language models.

Impact of model size and reasoning There is a clear correlation between model size and performance in the *Vinclat* task, as shown in Figure 3. Larger models consistently achieve higher accuracy than smaller ones, although the overall performance remained modest. This trend is more pronounced in models equipped with explicit reasoning capabilities, which show a more noticeable upward slope in performance with increasing size than non-reasoning models, consistent with prior work showing that emergent abilities such as abstraction and multi-hop reasoning often surface at scale (Wei et al., 2022a). By comparing identical architectures with and without reasoning enabled, it becomes clear that reasoning mechanisms enhance the model's ability to handle complex tasks like *Vinclat* (see examples in Appendix B).

Another interesting finding is that reasoner models often label hint keywords as "unknown", suggesting a higher degree of calibration and a learned tendency to abstain when uncertain —an ability associated with more robust and reliable models (Kadavath et al., 2022). The only outlier to this trend is Qwen3-0.6B, but we suspect this may be due to its limited capabilities, as the model is relatively small in size.

Impact of prompt format The choice of prompt format affects model performance in systematic ways (see Figure 2). In general, models tend to achieve higher accuracy with the "Letter-count" format, which explicitly states the number of words and letters, than with the "Fill-in-the-blanks" (underscore-based) version. A likely reason for this is that the former format provides a more explicit constraint, reducing ambiguity and helping models focus on finding a semantically coherent solution of the correct structure. In contrast, the latter requires symbolic interpretation and spatial reasoning, which appears to challenge smaller or less specialized models.

Interestingly, reasoner models handle the "Fill-in-the-blanks" format better than non-reasoning models, suggesting they can internalize structural constraints more effectively. These results align with findings that structured or symbolic formats require planning-like behavior (Yao et al., 2023) and that standard models struggle when format interpretability is low (Wei et al., 2022b). This highlights the importance of prompt format in evaluating model capabilities and suggests that different prompting strategies may probe distinct aspects of model reasoning and language understanding.

Impact of architecture Mixture-of-Experts (MoE) models consistently outperformed their dense counterparts of comparable nominal size (see Figure 3). While MoE architectures dynamically route tokens through different expert sub-networks, their actual parameter count is larger in aggregate, offering greater representational capacity (Shazeer et al., 2017). The advantage was especially notable in reasoner MoEs, suggesting a possible synergy between structured reasoning components and sparse activation patterns. These findings support recent literature showing that MoEs can yield high performance on diverse tasks when well-routed and supported by instruction-following objectives (Du et al., 2022). Nevertheless, they also raise open questions about efficiency and cultural robustness, as the better performance of MoEs does not yet translate to human-like capabilities in culturally rich or open-ended inference tasks like *Vinclat*.

Impact of prompt language Prompt language had only a minor effect on performance (see Table 3 and Figure 2). Differences between English and Catalan versions were small compared to those driven by reasoning capabilities or prompt format. This suggests that once models understood the task structure, their performance depended more on internal reasoning mechanisms than on the surface language of the instructions. The minimal gap may also indicate that high-performing multilingual or instruction-tuned models can transfer comprehension of well-specified tasks across languages (Scao et al., 2023). However, this apparent robustness should be interpreted with caution. Since *Vinclat* itself is culturally rooted in Catalan, limited variation by prompt language may simply reflect the dominant role of cultural and associative reasoning over linguistic comprehension.

Qualitative error analysis The error typology revealed consistent patterns across architectures and prompting conditions. The most frequent errors were misinterpretation of hints and incorrect final solutions despite reasonable partial reasoning, each accounting for roughly 40–60% of observed cases. Linguistic or cultural misunderstandings such as mistranslating idiomatic expressions or missing references to Catalan media, literature, or humor were also highly prevalent, especially in non-reasoning and English-prompted runs.

All models exhibited systematic reasoning artifacts. In DeepSeek-R1, a primacy effect was observed when prompted in Catalan (the model fixating on its first hypothesis even when incorrect), whereas the English version showed a recency effect, tending to choose the last proposed answer. Models prompted in English also displayed a tendency to overthink, producing excessively long and convoluted chains of thought that ultimately led it away from the correct answer. Both DeepSeek versions suffered from semantic overgeneralization, proposing overly broad concepts (e.g., *autism* for *Asperger*) and frequent cultural hallucinations when literally translating figurative Catalan.

The Qwen3-A22B model family showed a different profile. While they displayed stronger structural understanding —accurately reproducing the letter and word constraints— they often produced linguistic hallucinations, fabricating Catalan words from Spanish or English cognates (e.g., *meixó* from *mechón*). Orthographic errors and culturally induced failures were common, particularly when interpreting tongue-twisters, idioms, or jokes well known in Catalan society. Interestingly, several "failed" hints were semantically valid alternatives rather than true errors (e.g., guessing *tabú* instead of *secret*), suggesting that some model reasoning chains were plausible but misaligned with the original puzzle key.

Across models, reasoner variants generally produced more logically coherent explanations, whereas non-reasoning versions were more erratic and hallucination-prone. However, even the best-performing setups struggled with cultural grounding, reinforcing the view that commonsense and culturally situated reasoning remain unsolved challenges for current LLMs.

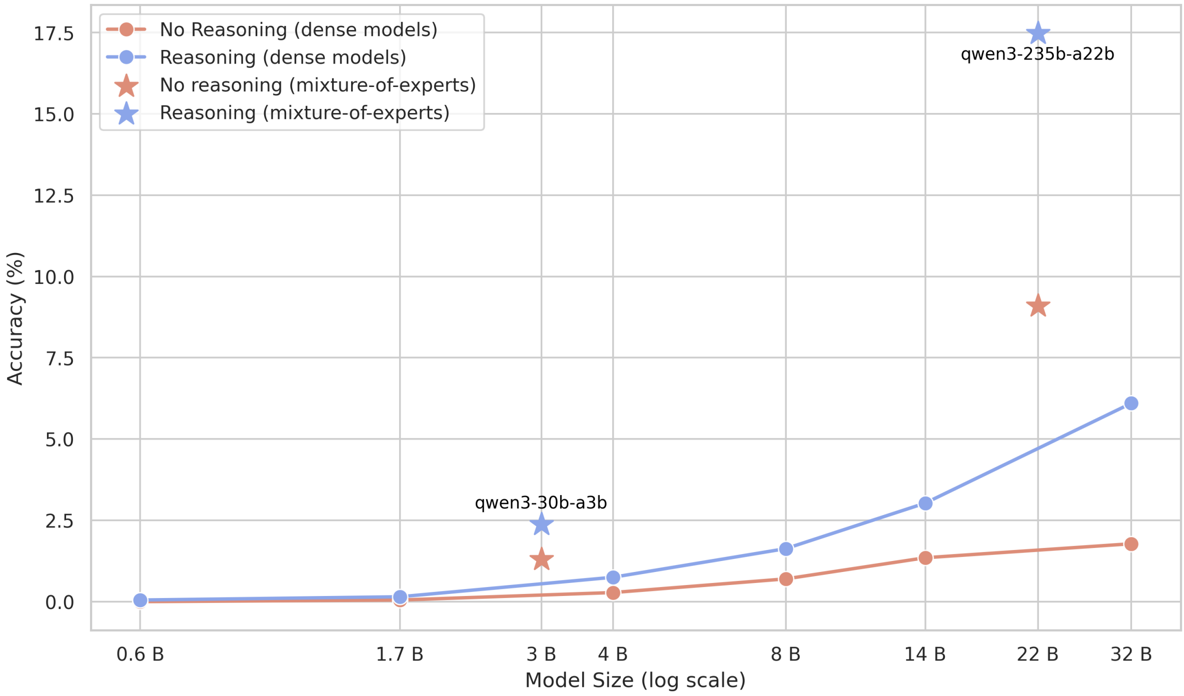

Figure 3: Accuracy scores based on model size and architecture for Qwen3 models.

6 Conclusion

This paper introduced a novel evaluation task for language models based on *Vinclat*, a Catalan word-association game that requires interpreting multiple hints and synthesizing them to retrieve a final solution under structural constraints. The resulting dataset of 1,000 high-quality instances —spanning domains such as culture, science, idioms, and history— offers a rich, culturally grounded benchmark for evaluating multi-hop reasoning, abstraction, and lexical-semantic association in LLMs.

Despite recent advances in large-scale language modeling, our results reveal that *Vinclat* remains a difficult challenge: even the strongest models tested achieved only modest accuracy, and performance fell significantly short of human baselines. The qualitative error analysis further showed that models frequently fail due to misinterpretation of clues, hallucinations, and a lack of cultural grounding, even when their reasoning structure appears coherent. Models also showed sensitivity to prompt format and language, with reasoners benefiting from clear structural prompts and showing less degradation when prompted in Catalan. Mixture-of-expert and reasoning-enabled models outperformed both their dense and non-reasoning counterparts, further supporting the need for explicit planning and compositional capabilities in solving complex tasks.

These findings point to clear limitations in current LLMs' abilities to integrate indirect clues, deal with partial information, and apply structural constraints over multiple reasoning steps, all essential capabilities for nuanced language understanding. *Vinclat* offers a promising diagnostic tool for probing such skills, especially in multilingual and culturally diverse contexts.

Future work may explore whether task-specific training, retrieval augmentation, or fine-tuned reasoning strategies can close the performance gap. Incorporating targeted cultural or linguistic knowledge could also improve model grounding and interpretability. Finally, expanding the benchmark to other languages or game formats could help test generalization and transfer capabilities in multilingual and cross-cultural LLM evaluation.

Limitations

Cultural and linguistic specificity *Vinclat* is deeply rooted in the Catalan language and culture. While this enriches its value as a culturally-grounded evaluation benchmark, it also poses challenges. Many hints rely on background knowledge that may not be well represented in LLM pretraining corpora, especially for mid-resource languages like Catalan. This makes it difficult to disentangle whether poor performance is due to lack of reasoning ability, insufficient cultural exposure, or both.

Single-turn setup The evaluation relied on a single-turn setting, which —although methodologically clean— does not reflect the interactive capabilities of modern LLMs. Multi-turn setups could allow models to revise initial guesses, ask for clarification, or reason iteratively. Especially for tasks involving ambiguity or multi-step reasoning, this restricted setup may underestimate true performance potential.

Scope of model types We focused on a representative set of open-weight models across different sizes and architectures. However, due to administrative and bureaucratic constraints, we did not include commercial, instruction-tuned models (e.g., GPT-4, Claude, Gemini), which may demonstrate stronger performance. In any case, the ultimate contribution of this work is the introduction of a new, complex dataset and to derive meaningful insights about the data, rather than to conduct an exhaustive model benchmarking exercise.

Human baseline While helpful as a reference point, the human baselines —especially the ones without internet access— may not fully reflect the capabilities of an average native speaker under realistic conditions. Although participants were trained on the task, they were not expert players and may have lacked familiarity with the types of strategies often needed to solve *Vinclat* puzzles. The internet-assisted baseline depends on the search strategies and tools used, which introduces variability. As such, these baselines are indicative rather than definitive upper bounds.

Acknowledgments

We sincerely thank Quim Roura, Alba Ferran and Blai Samitier for their creativity and dedication in crafting the riddles included in the dataset. We also extend our gratitude to Carlota Alba Foz for her meticulous review and corrections, and to BSC's Annotation Team for their role in establishing the human performance baseline.

This work was promoted and financed by *Generalitat de Catalunya* through the Aina project.

7 Bibliographical References

References

Alan Agresti and Brent A. Coull. 1998. Approximate is better than "exact" for interval estimation of binomial proportions. *The American Statistician*, 52(2):119–126.

Frederic C. Bartlett. 1932. *Remembering: A study in experimental and social psychology*. Cambridge University Press.

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Scott Wen tau Yih, and Yejin Choi. 2020. Abductive commonsense reasoning. *Preprint*, arXiv:1908.05739.

Ellen Bialystok. 1986. Factors in the growth of linguistic awareness. *Child development*, pages 498–510.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *Preprint*, arXiv:2303.12712.

Peter Clark, Isaac Cowhey, Oren Etzioni, and 1 others. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *NAACL-HLT*.

Nan Du, Yanping Huang, Andrew M. Dai, Simon Tong, Dmitry Lepikhin, Yuanzhong Xu, Maxim Krikun, Yanqi Zhou, Adams Wei Yu, Orhan Firat, Barret Zoph, Liam Fedus, Maarten Bosma, Zongwei Zhou, Tao Wang, Yu Emma Wang, Kellie Webster, Marie Pellat, Kevin Robinson, and 8 others. 2022. Glam: Efficient scaling of language models with mixture-of-experts. *Preprint*, arXiv:2112.06905.

Mor Geva, Daniel Khashabi, Wen-tau Chen, and Dan Roth. 2021. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics*.

Akash Ghosh, Debayan Datta, Sriparna Saha, and Chirag Agarwal. 2025. The multilingual mind: A survey of multilingual reasoning in language models. *arXiv preprint arXiv:2502.09457*.

Elliot Glazer, Ege Erdil, Tamay Besiroglu, Diego Chicharro, Evan Chen, Alex Gunning, Caroline Falkman Olsson, Jean-Stanislas Denain, Anson Ho, Emily de Oliveira Santos, and 1 others. 2024. Frontiermath: A benchmark for evaluating advanced mathematical reasoning in ai. *arXiv preprint arXiv:2411.04872*.

Alex Gu, Baptiste Rozière, Hugh Leather, Armando Solar-Lezama, Gabriel Synnaeve, and Sida I Wang. 2024. Cruxeval: A benchmark for code reasoning, understanding and execution. *arXiv preprint arXiv:2401.03065*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Dan Hendrycks, Steven Burns, Sam Basart, Collin Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Chengpeng Hu, Yunlong Zhao, Ziqi Wang, Haocheng Du, and Jialin Liu. 2024. Games for artificial intelligence research: A review and perspectives. *IEEE Transactions on Artificial Intelligence*.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, and 17 others. 2022. Language models (mostly) know what they know. *Preprint*, arXiv:2207.05221.

David Klahr and Kevin Dunbar. 1988. The dual space search hypothesis: An alternative to the problem space hypothesis for scientific reasoning. In *Proceedings of the 10th Annual Conference of the Cognitive Science Society*.

David R. Krathwohl. 2002. A revision of Bloom's Taxonomy: An overview. *Theory Into Practice*, 41(4):212–218.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. *Preprint*, arXiv:2309.06180.

Willem J.M. Levelt. 1989. *Speaking: From intention to articulation*. MIT Press.

Jiawei Li, Yizhe Yang, Yu Bai, Xiaofeng Zhou, Yinghao Li, Huashan Sun, Yuhang Liu, Xingpeng Si, Yuhao Ye, Yixiao Wu, , Bin Xu, Ren Bowen, Chong Feng, Yang Gao, and Heyan Huang. 2024. Fundamental capabilities of large language models and their applications in domain scenarios: A survey. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11116–11141, Bangkok, Thailand. Association for Computational Linguistics.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, and 31 others. 2023. Holistic evaluation of language models. *Preprint*, arXiv:2211.09110.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Angel Yahir Loredo Lopez, Tyler McDonald, and Ali Emami. 2025. NYT-connections: A deceptively simple text classification task that stumps system-1 thinkers. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 1952–1963, Abu Dhabi, UAE. Association for Computational Linguistics.

Nikolai Lörer, Norman Lee, and Douwe Kiela. 2023. Conceptarc: Evaluating abstract reasoning in language models with the abstraction and reasoning corpus. *arXiv preprint arXiv:2306.01774*.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2381–2391, Brussels, Belgium. Association for Computational Linguistics.

Swaroop Mishra, Matthew Finlayson, Pan Lu, Leonard Tang, Sean Welleck, Chitta Baral, Tanmay Rajpurohit, Oyvind Tafjord, Ashish Sabharwal, Peter Clark, and 1 others. 2022. Lila: A unified benchmark for mathematical reasoning. *arXiv preprint arXiv:2210.17517*.

Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, Vassilis Plachouras, Tim Rocktäschel, and Sebastian Riedel. 2021. KILT: a benchmark for knowledge intensive language tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2523–2544, Online. Association for Computational Linguistics.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, and 373 others. 2023. Bloom: A 176b-parameter open-access multilingual language model. *Preprint*, arXiv:2211.05100.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *Preprint*, arXiv:1701.06538.

Parshin Shojaee, Iman Mirzadeh, Keivan Alizadeh, Maxwell Horton, Samy Bengio, and Mehrdad Farajtabar. 2025. The illusion of thinking: Understanding the strengths and limitations of reasoning models via the lens of problem complexity. *arXiv preprint arXiv:2506.06941*.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, and 1 others. 2016. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489.

David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, and 1 others. 2017. Mastering the game of go without human knowledge. *nature*, 550(7676):354–359.

Koustuv Sinha, Shagun Sodhani, and 1 others. 2019. Clutrr: A diagnostic benchmark for inductive reasoning from text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, and 432 others. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Preprint*, arXiv:2206.04615.

Aarohi Srivastava and 1 others. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georgiev, and 1 others. 2019. Grandmaster level in starcraft ii using multi-agent reinforcement learning. *nature*, 575(7782):350–354.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, and 1 others. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022a. Emergent abilities of large language models. *Preprint*, arXiv:2206.07682.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022b. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *Preprint*, arXiv:2305.10601.

Xingyu Zheng, Yuye Li, Haoran Chu, Yue Feng, Xudong Ma, Jie Luo, Jinyang Guo, Haotong Qin, Michele Magno, and Xianglong Liu. 2025. An empirical study of qwen3 quantization. *arXiv preprint arXiv:2505.02214*.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023. Lima: Less is more for alignment. *Preprint*, arXiv:2305.11206.

Terry Yue Zhuo, Minh Chien Vu, Jenny Chim, Han Hu, Wenhao Yu, Ratnadira Widyasari, Imam Nur Bani Yusuf, Haolan Zhan, Junda He, Indraneil Paul, and 1 others. 2024. Bigcodebench: Benchmarking code generation with diverse function calls and complex instructions. *arXiv preprint arXiv:2406.15877*.

A Prompt formats

English prompt with fill-in-the-blanks format (EN-FB)

Let’s play a game in Catalan! Your goal is to find a "solution" word or words of a specific length, which will be given to you. You will also receive four numbered hints. Try to solve each one to get a "hint word", which doesn’t need to match the solution’s length. Then, try to think about the common theme or connection between the "hint words" that you found. The final solution should fit the required letter count and is related to the "hint words" you identified.

It is important to note that you don’t necessary need all the "hint words" to get to the final solution. If you’re struggling with a hint or suspect your guess might be wrong, it’s often better to focus on the "hint words" you are sure about. A wrong one can send you down the wrong path!

Here are your hints:
1. `{hint_1}`
2. `{hint_2}`
3. `{hint_3}`
4. `{hint_4}`

The "solution" should fit here: `{underscores}`. What’s your guess? Return a JSON object with the following fields: *’hint_word_1’, ’hint_word_2’, ’hint_word_3’, ’hint_word_4’, ’solution’*. If you could not find the word associated to some hint, simply keep that field as *’unknown’*.

Catalan prompt with fill-in-the-blanks format (CA-FB)

Et proposo jugar a un joc en català! El teu objectiu és trobar una paraula o paraules "solució" d’una longitud específica, que et serà donada. També rebràs quatre pistes numerades. Intenta resoldre cadascuna per obtenir una "paraula pista", que no té perquè tenir la mateixa longitud que la solució. Després, intenta pensar en el tema comú o la connexió entre les "paraules pista" que has trobat. La solució final ha de tenir la longitud requerida i estar relacionada amb les "paraules pista" que hagis identificat.

És important tenir en compte que no necessàriament necessites totes les "paraules pista" per arribar a la solució final. Si tens dificultats amb una pista o sospites que la teva conjectura pot ser incorrecta, sovint és millor centrar-se en les "paraules pista" de les quals estàs segur. Una resposta incorrecta pot ser contraproduent, ja que et farà anar pel camí equivocat!

Aquí tens les teves pistes:
1. `{hint_1}`
2. `{hint_2}`
3. `{hint_3}`
4. `{hint_4}`

La solució hauria d’encaixar aquí: `{underscores}`. Quina diries que és la resposta correcta? Retorna un objecte JSON amb els següents camps: *’hint_word_1’, ’hint_word_2’, ’hint_word_3’, ’hint_word_4’, ’solution’*. Si no has pogut trobar la paraula associada a alguna pista, simplement deixa aquest camp com a *’unknown’*.

English prompt with letter-count format (EN-LC)

Let’s play a game in Catalan! Your goal is to find a "solution" word or words of a specific length, which will be given to you. You will also receive four numbered hints. Try to solve each one to get a "hint word", which doesn’t need to match the solution’s length. Then, try to think about the common theme or connection between the "hint words" that you found. The final solution should fit the required letter count and is related to the "hint words" you identified.

It is important to note that you don’t necessary need all the "hint words" to get to the final solution. If you’re struggling with a hint or suspect your guess might be wrong, it’s often better to focus on the "hint words" you are sure about. A wrong one can send you down the wrong path!

Here are your hints:
1. `{hint_1}`
2. `{hint_2}`
3. `{hint_3}`
4. `{hint_4}`

The "solution" has `{letter_count}`. What’s your guess? Return a JSON object with the following fields: *’hint_word_1’, ’hint_word_2’, ’hint_word_3’, ’hint_word_4’, ’solution’*. If you could not find the word associated to some hint, simply keep that field as *’unknown’*.

Catalan prompt with letter-count format (CA-LC)

Et proposo jugar a un joc en català! El teu objectiu és trobar una paraula o paraules "solució" d’una longitud específica, que et serà donada. També rebràs quatre pistes numerades. Intenta resoldre cadascuna per obtenir una "paraula pista", que no té perquè tenir la mateixa longitud que la solució. Després, intenta pensar en el tema comú o la connexió entre les "paraules pista" que has trobat. La solució final ha de tenir la longitud requerida i estar relacionada amb les "paraules pista" que hagis identificat.

És important tenir en compte que no necessàriament necessites totes les "paraules pista" per arribar a la solució final. Si tens dificultats amb una pista o sospites que la teva conjectura pot ser incorrecta, sovint és millor centrar-se en les "paraules pista" de les quals estàs segur. Una resposta incorrecta pot ser contraproduent, ja que et farà anar pel camí equivocat!

Aquí tens les teves pistes:
1. `{hint_1}`
2. `{hint_2}`
3. `{hint_3}`
4. `{hint_4}`

La "solució" té `{letter_count}`. Quina diries que és la resposta correcta? Retorna un objecte JSON amb els següents camps: *’hint_word_1’, ’hint_word_2’, ’hint_word_3’, ’hint_word_4’, ’solution’*. Si no has pogut trobar la paraula associada a alguna pista, simplement deixa aquest camp com a *’unknown’*.

B Selected instances

Problem

Hint 1: Pot ser de vàter, cuina, diari, seda...
Hint 2: El cinema n'és el setè
Hint 3: Els trencaclosques et fan unir-ne moltes
Hint 4: El que pots fer si prems Ctrl+V

Solution

Keyword 1: Paper
Keyword 2: Art
Keyword 3: Peces
Keyword 4: Enganxar
Final solution: Collage

LLM's reasoning trace

Okay, let's tackle this Catalan puzzle. I need to find the solution word based on the four hints provided. The solution has a specific length, but the hint words don't necessarily need to match that. Let me go through each hint one by one.

1. **Pot ser de vàter, cuina, diari, seda...** The first hint mentions things that can be "de vàter" (toilet), "cuina" (kitchen), "diari" (daily), "seda" (silk). I'm thinking these are types of paper. Toilet paper, kitchen paper, daily paper (newspaper), silk paper. So the hint word here might be "paper".

2. **El cinema n'és el setè** The cinema is the seventh. Hmm, maybe referring to the "seventh art". Cinema is often called the seventh art. In Catalan, that's "el setè art". So the hint word here is "art".

3. **Els trencaclosques et fan unir-ne moltes** Puzzles (trencaclosques) make you put together many... Pieces? Parts? The Catalan word for pieces is "peces" or "peces de trencaclosques". So the hint word might be "peces".

4. **El que pots fer si prems Ctrl+V** If you press Ctrl+V, that's the paste function. In Catalan, "enganxar" means to paste. So the hint word could be "enganxar".

Now, looking at the hint words: paper, art, peces, enganxar. What's the common theme? Maybe combining these concepts. Paper is a material, art is creative, pieces are parts, and paste is sticking them together. Maybe the solution is related to collage. A collage involves pasting pieces of paper or other materials to create art. The Catalan word for collage is "collage" (same as English) but let me check the length. The solution blank is 7 letters. "Collage" has 7 letters. So that fits. Therefore, the solution should be "collage".

Figure 4: Response provided by DeepSeek-R1 to the instance with ID 453.

Problem

Hint 1: Hi ha qui en té als cacauets, als gats o a la pols
Hint 2: No te'ls donen sense recepta
Hint 3: L'arrosseguen les abelles de flor en flor
Hint 4: Estació que comença la segona meitat de març

Solution

Keyword 1: Al·lèrgia
Keyword 2: Medicaments
Keyword 3: Pol·len
Keyword 4: Primavera
Final solution: Antihistamínic

LLM's reasoning trace

Okay, let's try to figure this out step by step. The goal is to find a solution word that's 14 letters long, based on four clues. Each clue gives a "hint word," and the solution is related to all of them. Let's tackle each clue one by one.

First clue: "Hi ha qui en té als cacauets, als gats o a la pols." Translating that from Catalan, it's talking about something you can find in peanuts, cats, or dust. Hmm. Common things in those... Allergies? Because people can be allergic to peanuts, cats (pet dander), and dust mites. So maybe "al·lèrgia" (allergy) is the hint word here. Let me check the length. "Al·lèrgia" has 8 letters, but maybe in plural? Not sure yet.

Second clue: "No te'ls donen sense recepta." Which translates to "They don't give them to you without a prescription." So this is likely referring to medications, like pills or drugs. The Catalan word for pills is "píndoles" or "medicaments." But "medicaments" is longer. Wait, "píndoles" is 8 letters. But maybe "recepta" itself? No, the answer here is probably "píndoles" since they require a prescription.

Third clue: "L'arrosseguen les abelles de flor en flor." Bees drag this from flower to flower. Bees collect pollen, so "pol·len" in Catalan. That's 6 letters. That seems right. Pollen is what they carry.

Fourth clue: "Estació que comença la segona meitat de març." The season starting in the second half of March. In the Northern Hemisphere, spring starts around March 20th or 21st. So "primavera" (spring) in Catalan. That's 9 letters.

Now, the solution needs to be 14 letters long and related to all these hints. Let's look at the hint words again: allergy, pills, pollen, spring. How do these connect? Spring is when pollen is high, causing allergies. People take pills (antihistamines) for allergies. So the common theme is allergies. The solution might be a word related to allergies or antihistamines. But a 14-letter Catalan word... Let's think. "Antihistamínic" (antihistamine) has 14 letters. Let's check: A-N-T-I-H-I-S-T-A-M-Í-N-I-C. Yes, 14 letters. That makes sense. The hint words are all related to allergies: allergens (peanuts, cats, dust), medication (pills), pollen (from flowers), and spring (when pollen is high). So the solution is "antihistamínic".

Figure 5: Response provided by Qwen3-235B-A22B to the instance with ID 965.

Problem

Hint 1: Per molts l'homeopatia no ho és
Hint 2: La tercera pota que et falta d'un tamboret si ja tens els diners i l'amor
Hint 3: El percebem amb l'oïda
Hint 4: Sentir atentament

Solution

Keyword 1: Medicina
Keyword 2: Salut
Keyword 3: So
Keyword 4: Escoltar
Final solution: Estetoscopi

LLM's reasoning trace

Okay, let's tackle this puzzle step by step. The user wants me to find a solution word of a specific length (which isn't given here, but maybe the solution line has a blank that indicates the length) based on four clues. Each clue should lead to a "hint word," and the solution connects these hints.

Starting with clue 1: "Per molts l'homeopatia no ho és." Homeopathy is often considered a pseudoscience, so maybe the hint word here is "ciència" (science) because some people don't think homeopathy is scientific.

Clue 2: "La tercera pota que et falta d'un tamboret si ja tens els diners i l'amor." There's a saying that a stool needs three legs: "diner, amor, salut" (money, love, health). If they already have money and love, the missing third leg would be "salut" (health).

Clue 3: "El percebem amb l'oïda." That which we perceive with hearing is "so" (sound) or "música" (music). Maybe "so" is the hint here.

Clue 4: "Sentir atentament." To listen attentively could be "escoltar" (to listen) or "atenció" (attention). But "escoltar" is more directly related to listening carefully.

Now, connecting the hint words: ciència, salut, so, escoltar. The common theme might be something that combines science, health, sound, and listening. Maybe "salut" relates to health, "ciència" to medicine or biology, "so" to sound waves, and "escoltar" to using a stethoscope. The solution could be "estetoscopi" (stethoscope), which is a medical instrument used to listen to body sounds, related to health and science. Let's check the length: "estetoscopi" is 11 letters. If the blank is for an 11-letter word, that fits. Alternatively, maybe "fonendoscopi" which is another term, but less common. But "estetoscopi" is more widely known. So I think that's the solution.

Figure 6: Response provided by DeepSeek-R1 to the instance with ID 783.

Problem

Hint 1: El 10 de novembre celebrem el dia mundial de la...
Hint 2: El paper de Harrison Ford amb un fuet
Hint 3: Estar ben mort i ________
Hint 4: L'explicació del passat

Solution

Keyword 1: Ciència
Keyword 2: Indiana Jones
Keyword 3: Enterrat
Keyword 4: Història
Final solution: Arqueologia

LLM's reasoning trace

Okay, let's tackle this step by step. I need to find four hint words based on the clues provided and then figure out the solution word that's 11 letters long and connected to those hints.

Starting with hint 1: "El 10 de novembre celebrem el dia mundial de la..." November 10th is World Day of something. In Catalan, "dia mundial de la..." probably refers to a common world day. Let me think. World Day of what? Maybe "Consciència" (Consciousness) or "Creativitat" (Creativity)? Wait, but the most common one I know is World Science Day on November 10th. So "Ciència" (Science) in Catalan. So hint_word_1 could be "Ciència".

Hint 2: "El paper de Harrison Ford amb un fuet" (The role of Harrison Ford with a whip). Harrison Ford played Indiana Jones, who uses a whip. So the character is Indiana Jones, but in Catalan, maybe "Indiana Jones" or "Indi Jones"? Or perhaps "Arqueòleg" (archaeologist), but the main role with a whip is definitely Indiana Jones. So hint_word_2 is "Indiana Jones".

Hint 3: "Estar ben mort i ________" (To be very dead and ________). This sounds like a phrase. Maybe "estar ben mort i enterrat" (be very dead and buried)? "Enterrat" is buried. But the blank might be a single word. Alternatively, maybe "estar ben mort i sencer" (dead and whole?), but that doesn't make sense. Wait, common expressions: "Ben mort i enterrat" makes sense. So the hint_word_3 could be "enterrat".

Hint 4: "L'explicació del passat" (The explanation of the past). That could be "història" (history) because history explains the past. So hint_word_4 is "Història".

Now the solution is an 11-letter word related to these hints: Ciència (Science), Indiana Jones, Enterrat (Buried), Història (History). Let me think of the connections. Indiana Jones is an archaeologist, which relates to history and exploring the past. "Enterrat" refers to being buried, which could relate to archaeological finds. Science and history are both involved in studying the past. The common theme might be something like "Arqueologia" (Archaeology). Let's check the letters: A-R-Q-U-E-O-L-O-G-I-A. That's 11 letters. Yes! So the solution is "Arqueologia".

Figure 7: Response provided by Qwen3-235B-A22B to the instance with ID 876.