UC Merced

Proceedings of the Annual Meeting of the Cognitive Science Society

Title

Measuring IA Cognitive Competences: A Gap in Current Benchmarks

Permalink

https://escholarship.org/uc/item/4zx0k1b6

Journal

Proceedings of the Annual Meeting of the Cognitive Science Society, 48(0)

Authors

Amigó, Enrique
Morante, Roser
Gonzalo, Julio
et al.

Publication Date

2026

Copyright Information

This work is made available under the terms of a Creative Commons Attribution License, available at https://creativecommons.org/licenses/by/4.0/

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Measuring IA Cognitive Competences: A Gap in Current Benchmarks

Enrique Amigó[1], **Roser Morante**[2], **Julio Gonzalo**[1], **Laura Plaza**[1]
[1] NLP&IR Group – UNED, Spain; [2]HiTZ Center – Ixa, University of the Basque Country UPV/EHU

Abstract

To what extent do current LLM benchmarks pose the same reasoning challenges that arise in real-world settings? Instead of relying on anthropocentric notions, here we analyze the cognitive competences demanded by a benchmark in terms of (a) formal complexity (e.g. number of inference steps required to reach the desired answer) and (b) amount of contextual interpretation needed to reduce language ambiguity to a level where the required inference can be performed. Our analysis of a representative sample of benchmarks reveals that there are almost no benchmarks with both, high formal complexity and high contextual interpretation requirements. This limitation may help explain the discrepancy between the strong benchmark performance of language models and the shortcomings observed when these models are deployed in real world applications.

Keywords: Large language models; cognitive competences; reasoning; meaning; artificial intelligence.

Introduction

Large language models (LLMs) have become central tools for the study of linguistic and cognitive processes, prompting extensive discussion about whether they possess cognitive competences comparable to those observed in humans. Despite the achievements of LLMs, their high performance on many established benchmarks contrasts sharply with the inconsistent behavior observed in real world settings (Wu et al., 2024; Kazemi et al., 2025; Wei et al., 2026). This disparity suggests that high benchmark scores do not necessarily reflect the breadth of capabilities required in practical applications, especially when dataset artifacts, memorization effects, or shortcut strategies allow models to perform well without engaging in the conceptual operations that benchmarks are intended to assess.

One reason for this mismatch is that prevailing definitions of competences are largely grounded in human-centered cognitive skills. Many studies assess the cognitive capabilities of AI systems through broad and loosely defined reasoning categories, such as logical, argumentative, deductive, abductive, critical, decompositional, multi hop, or mathematical reasoning (Guo et al., 2023; H. Liu et al., 2025). Relying exclusively on these anthropocentric categories risks obscuring the distinctive capacities of large language models and can simultaneously underestimate their strengths while overestimating their ability to reproduce human-like reasoning (Mitchell, 2019; Millière & Rathkopf, 2026).

Moreover, benchmarks that are supposed to assess the same reasoning skill may rely on substantially different cognitive demands. For example, two benchmarks may be equally described as testing *deductive* competences, but one may only requires combining explicit rules provided in the input (e.g. a purely mathematical task), whereas the other may also involve strong contextual disambiguation that exploits real-world knowledge.

Other works address less formal aspects, such as flexible reasoning (Kim et al., 2025), coherence of conceptual structures across tasks (Suresh et al., 2023), or adaptability to novel situations (Valiente & Pilly, 2025; Zhu et al., 2024). Although the literature consistently shows that these aspects remain significant limitations of current AI systems, it is still necessary to specify which underlying cognitive mechanisms are intended to be measured through evaluation on existing benchmarks.

In this work, we conduct a qualitative analysis of a representative set of existing benchmarks, with a specific focus on their **meaning reduction** requirements. That is, given the full space of possible meanings of the texts involved in the task, performing complex inference requires reducing this space, on the basis of contextual information, to the meaning interpretations that allow to establish logical implications.

From this perspective, we characterize (reasoning) cognitive competences along two complementary dimensions, and analyze benchmarks by placing them within these two dimensional space. On the one hand, **formal complexity** captures the number of elements and deterministic relations, or the length of reasoning chains, that must be handled to solve a task. On the other hand, the **contextual scope of meaning reduction** refers to the type of context that must be handled to carry out meaning reduction, ranging from implication relations explicitly stated in the input to the ambiguity reduction that requires applying real-world contextual constraints.

Our contributions are the following: (i) We propose to characterize the cognitive competences required by LLM evaluation benchmarks on the basis of their *meaning reduction* and *formal complexity* requirements. (ii) We analyze over fifty existing benchmarks in terms of the proposed schema. (iii) Our analysis reveals that **there is a substantial lack of benchmarks with tasks that combine both high formal complexity and context-aware meaning reduction**. This result suggests that, despite the remarkable capabilities exhibited by current generative AI technology, there are fundamental aspects of its cognitive competences that are not being properly measured. This inadequacy may help explaining the discrepancy between the scores obtained in evaluation benchmarks and the limitations observed when these systems are deployed in real world settings.

The Need for Meaning Reduction in Cognitively Demanding Tasks

The observation that identifying and applying inferential relations between textual expressions or statements re-

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society* ©2026 the author(s). This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY).

quires narrowing the space of possible meanings is not new (MacCartney & Manning, 2007). However, in this section we argue that meaning reduction becomes increasingly central as machine learning systems reach higher levels of cognitive competence. To this end, we introduce the notions of *meaning space* and *meaning reduction*, and examine their role in cognitively demanding tasks. Finally, we specify the two dimensions along which benchmarks are analyzed, namely formal complexity and contextual scope.

The Meaning Space

We assume that LLMs evaluated through cognitive competence benchmarks take as input a collection of textual statements, both during training and at inference time, and that each statement is associated with an underlying meaning. However, the concept of meaning is fuzzy, and it can be supported by different linguistic theories such as referential meaning theory (Lycan, 2018), use based, ideational (Gelepithis, 1988), or verificationist (Schrenk, 2008) theories.

Regardless of the theory of meaning we consider most appropriate, for our purposes we only need to assume that behind any expression lies a set of possible meanings. Note that these sets of possible meanings may be fuzzy, rather than discrete. This interpretation of *meaning* aligns with formal semantics frameworks such as *Situation Semantics* (Barwise & Perry, 1981), where there exists a universal space of situations, and a linguistic expression does not denote a single value but rather the set of situations in which the expression is true, appropriate, or informationally compatible. It also connects with Montague's *Intensional Semantics*, where meaning is defined as the set of worlds in which the expression is true (Montague, 1973). It also relates to Probabilistic Semantics, where an expression corresponds to a structured region within the semantic space, characterized by graded uncertainty (Goodman & Lassiter, 2015).

Figure 1 exemplifies this idea. Given two expressions A and B, both are associated with a subset of possible meanings belonging to the global meaning space. Expression B is in principle more general than A, given that *soccer* is more specific than *sport*. Therefore, A connects with a smaller subset of possible meanings. However, there are possible meanings of the second expression that are not shared with the meanings of the first expression. For example, the second expression could refer to a computer game, in which case that meaning would not fit the first expression.

Meaning Reduction vs. AI Competences

Starting from this space of meanings, and setting aside anthropocentric considerations, we can analyze the different competences that AI paradigms have developed over time.

First, *meaning generalization* covers tasks that require broadening the meanings associated with the input so that expressions, text fragments, or documents in the input or in the collection can be matched. For instance, in Figure 1, a textual similarity detection task between expressions A and B can be

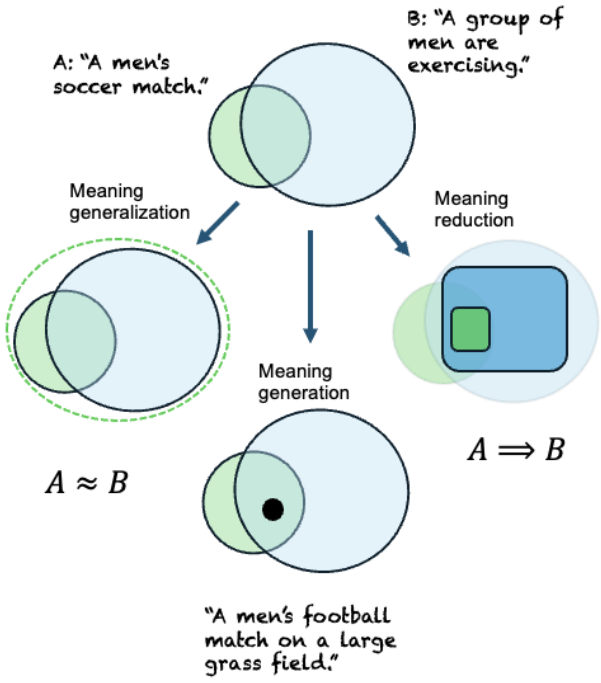

Figure 1: The meaning space and meaning oriented cognitive competences.

interpreted as requiring a generalization of both meanings so that the two expressions can be treated as equivalent.

Meaning generalization underlies many tasks addressed by traditional discriminative AI systems, such as text classification, in which input texts are generalized into predefined categories. Information retrieval can also be viewed within this paradigm, since a relevant document in the collection, which is typically more specific than the query, is generalized in order to match it with the query. A similar principle applies to clustering and topic detection tasks.

At a second competence level, the system not only establishes relations between existing meanings, but also generates text that conveys new meanings. We refer to this level as *meaning generation*. For example, in Figure 1, the system produces an extended text that increases the specificity of the input texts, in the sense that its meaning lies in the intersection of the sets of possible meanings associated with expressions A and B.

Meaning generation is actually the natural competence of generative systems based on LLM's. Note that, in general, extending a sequence $(w_{i-n}, \ldots, w_{i-1})$ with a continuation $(w_i, \ldots, w_{i+m})$ that maximizes $P(w_i, \ldots, w_{i+m} \mid w_{i-n}, \ldots, w_{i-1})$ yields an extended text whose meaning is more specific than that of the original sequence.[1] Although this competence level does not necessarily rely on explicit logical structures or symbolic inference, it is often sufficient

[1]Note that there are some exceptions, for instance, when adding disjunctive statements.

to support conversational systems and to exhibit argumentative behavior. However, as widely documented in the literature, it entails a substantial risk of hallucination, internal inconsistency, and limited interpretability.

In this work, we focus on a higher competence level, which we refer to as *meaning reduction*, which is closely related to the notion of deductive reasoning and rule induction. Identifying entailment relations between input statements is a fundamental component of many reasoning processes, whether the goal is to derive hypotheses, induce rules, or perform other forms of inference. Within the meaning space introduced above, entailment relationships can be conceptualized as an strict inclusion relation between meanings, which requires reducing their space of possible meanings.

For example, in the case of Figure 1, establishing a strict implication requires assuming that a soccer match necessarily involves physical exercise, so that the set of possible meanings of A can be regarded as a subset of possible meanings of B. This, however, requires introducing assumptions or imposing constraints. One must assume that the subjects are not playing a video game, or that they are actually playing rather than merely planning a match. In such situations, inference becomes possible only after restricting the sets of possible meanings in order to establish a subsumption relation.

According to this perspective, the *meaning reduction* competence level needs to be evaluated with tasks and benchmarks that require reducing the set of possible interpretations of the input expressions in order to establish strict subsumption relations, and therefore implications. Examples of tasks requiring meaning reduction are natural language inference and answering questionnaires involving logical problems.[2]

Note that the meaning reduction competence covers and goes beyond meaning generalization and meaning generation. Meaning reduction involves the progressive restriction of the possible meanings space, driven by task constraints, semantic context, and the need to identify deterministic inferential relationships between statements, rather than merely linking meanings or generating text based on patterns internalized during pretraining. Acquiring these competencies might require new paradigms, such as neuro-symbolic approaches.

In terms of evaluation benchmarks, ensuring that a task requires meaning reduction can be challenging. For instance, (Z. Liu, Yu, & He, 2025) show that removing example demonstrations does not reduce the ability to induce classification rules. This indicates that the LLM is in fact relying on model prior, rather than performing genuine meaning reduction and inference.

Formal Complexity vs. Contextual Scope

In this article, we conduct a qualitative analysis of benchmarks that require meaning reduction along two dimensions. The first dimension is *formal complexity*, which we take to capture aspects such as the number of variables involved in a single decision (Andrews & Halford, 2002), the arity and structure of the relations involved (Halford, Wilson, & Phillips, 1998), the number and structure of logical operators, (Salto, Requena, Álvarez-Merino, Antón-Toro, & Maestú, 2021), or the number of steps required by the inference (Johnson-Laird, 2006). For our purposes, we do not need to define a precise function. It is sufficient to identify types of benchmarks whose formal complexity increases along all these aspects with respect to another benchmark.

The second is the *contextual scope* of meaning reduction, which refers to how explicit the contextual information is that determines the strict relations of implication among statements. It can be described by the following three categories:

i) Explicit Relational Structures: When meaning is determined by explicit logical relations or by strict formal rules stated in the prompt or in the training data. This occurs, for example, in mathematical problem solving. A characteristic of these benchmarks is that the symbols or expressions referring to the related entities can be modified without affecting the problem itself.[3]

ii) Context Independent Semantics: When meaning depends on the interpretation of phrases within natural language in isolation from any particular scenario. In this case, implication relations are not explicitly stated but follow from the general semantics of the input expressions. For example, *"John eats an apple"* versus *"John nourishes himself"*. The semantic interpretation involved here is general and operates at the sentence level, independently of any broader context.

iii) Context Dependent Semantics: When meaning interpretation requires a specific situational or domain context. In this case, implication relations cannot be derived from linguistic meanings alone. The specificity of the required context forms a continuum, ranging from broad intellectual frameworks such as *Descartes' philosophy* to highly concrete scenarios such as administrative procedures in a particular institution or company.

Notice that the relationship between formal complexity and meaning reduction is bidirectional, in the sense that implication relations are not merely the result of meaning reduction. Rather, the process of reducing meanings in order to solve a problem is itself guided by the implication relations that can be established among the candidate interpretations.

Analysis of Cognitive Competence Benchmarks

In this and the following subsections, we analyze benchmarks that require meaning reduction. Figure 2 presents the benchmarks organized according to the two factors previously described. As a point of reference, a high level of formal complexity with no context meaning dependence is reached by a Prolog compilation or by an inductive logic programming system. At the other extreme, ontology learning involves a high degree of meaning reduction, although not formal com-

[2]Note that mathematical functions also fall under this interpretation, since they can be understood semantically as implication relations, for example $2 \wedge 3 \wedge sum \Longrightarrow 5$.

[3]For example, in a logical problem where implication relations between propositions are specified, each of these propositions can be replaced by symbols.

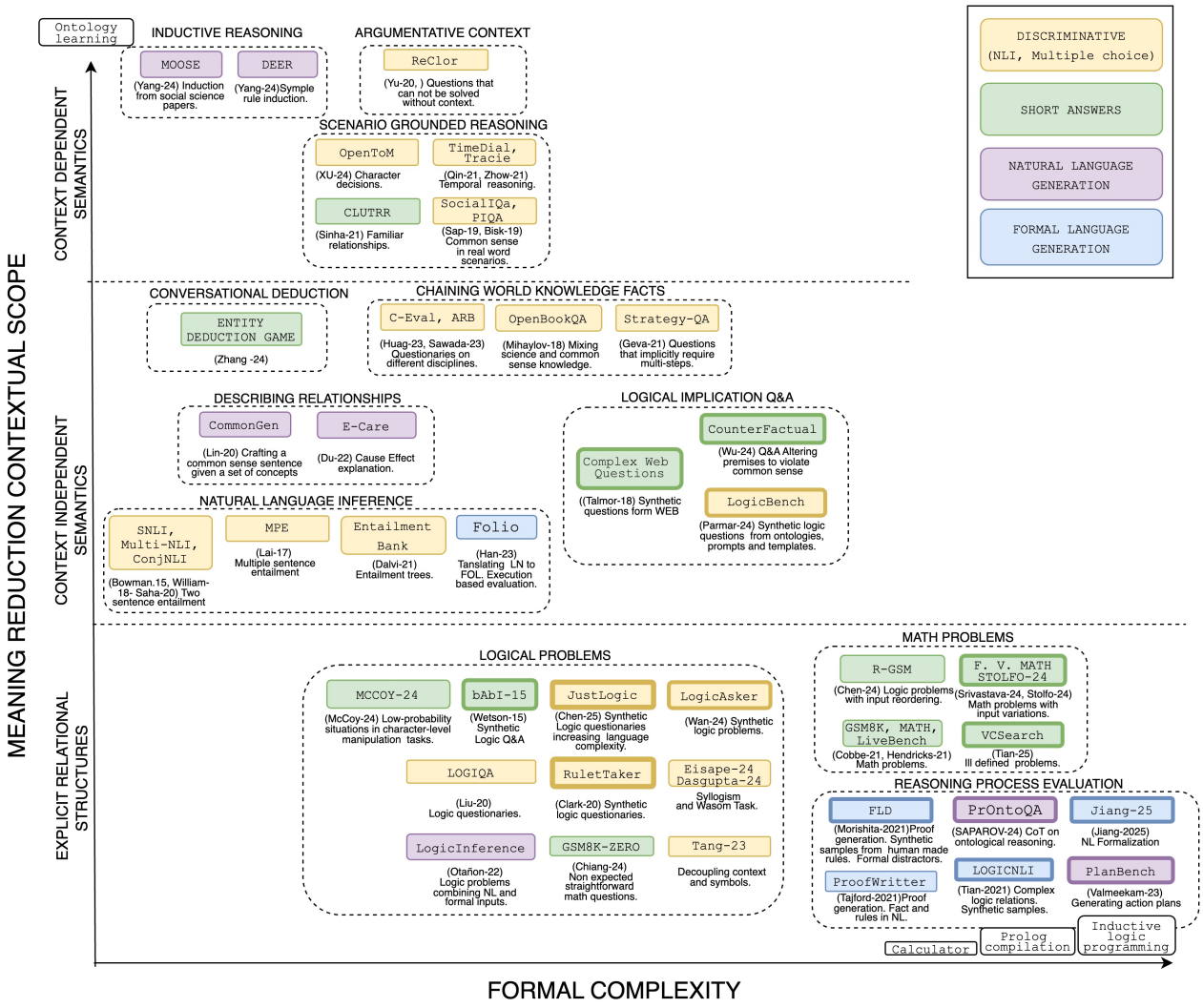

Figure 2: Categorization of core competence benchmarks according to two factors: context needed to perform meaning reduction (vertical axis) and formal complexity (horizontal axis). The type of output is color-encoded (see box in the upper right corner). Benchmarks with bold borders include automatically generated test cases.

plexity in terms of problem solving.

Two important considerations must be kept in mind when interpreting this figure. First, the figure is not the result of applying any quantitative metric, but rather of ordering the benchmark sets according to the comparative analysis described throughout this section. Second, our goal is not to provide an exhaustive survey, but to examine a representative subset that allows us to shed light on what type of tasks and data are included in benchmarks in order to evaluate the targeted cognitive competences.

The datasets are color-coded according to the type of output, namely: discriminative tasks in yellow (e.g., NLI, multiple-choice), short free-text answers in green, long answers in purple, and formal language outputs in blue.[4]

[4]Note that some benchmarks appear more than once, as they may be associated with different tasks. If the benchmark does not have a name, we refer to it with the name of the author or model which was presented with.

Explicit Relational Structure Benchmarks

The lower part of the figure includes benchmarks in which meaning reduction is achieved through formal relations explicitly expressed in the input and training data. In a first category, we find benchmarks based on logical problems, where the input combines logical expressions and natural language, such as LogicInference (Ontanon, Ainslie, Cvicek, & Fisher, 2022) or LogiQA (J. Liu et al., 2020). Some benchmarks such as RuleTaker (Clark, Tafjord, & Richardson, 2020), JustLogic (M. K. Chen, Zhang, & Tao, 2025), LogicAsker (Wan et al., 2024) and bAbI (Weston et al., 2015) avoid contamination by using synthetically generated logic questions. Other benchmarks increase difficulty by modifying linguistic expressions so that words no longer convey natural semantic cues (Tang et al., 2023), or by manip-

ulating term probabilities to weaken distributional semantic signals (McCoy, Yao, Friedman, Hardy, & Griffiths, 2024). Additional formal complexity is introduced through classic logic problems such as syllogisms or the Wason selection task (Eisape et al., 2024; Dasgupta et al., 2024).

From logic problems, the benchmark landscape expands to mathematical reasoning tasks such as those found in GSM8K (Cobbe et al., 2021) and MATH (Hendrycks et al., 2021). We have placed this group on the right assuming that mathematical problems involve greater formal complexity aspects than logical problems. Interestingly, explicitly including the solution in the input can lead models to over reason, as shown in GSM8K ZERO (Chiang & Lee, 2024). Further sources of complexity include the reordering of premises in R-GSM (X. Chen, Chi, Wang, & Zhou, 2024), functional transformations (Srivastava et al., 2024; Stolfo, Jin, Shridhar, Schoelkopf, & Sachan, 2023), and ill defined problem settings such as those in VCSEARCH (S.-Y. Tian et al., 2025).

A final group of benchmarks evaluates not only the final result, but also the reasoning process itself. This increases formal complexity insofar as it requires identifying the number of reasoning steps required. However, it requires benchmarks to explicitly specify entailment relations. We have placed this group accordingly.

PRONTOQA poses first order logical problems generated from synthetic world models, producing instances that require long chains of reasoning steps (Saparov & He, 2023). Other benchmarks require the system output to be expressed in a formal language, making it possible to validate both the final answer and the full chain of reasoning. In PROOFWRITTER (Tafjord, Dalvi, & Clark, 2021), RULETAKER is extended to evaluate the ability of systems to produce formal proofs that are assessed in terms of proof correctness. In FLD (Morishita, Morio, Yamaguchi, & Sogawa, 2023), formal complexity is increased by instructing annotators to generate implication rules and by introducing factual distractors. PLANBENCH introduces small robot and object worlds in which actions must be chained step by step in order to achieve a goal (Valmeekam, Marquez, Olmo, Sreedharan, & Kambhampati, 2023). (Jiang et al., 2025) evaluate LLM reasoning chains across multiple formats, including Python, natural language, formal language, and mixed representations. In LOGICNLI (J. Tian et al., 2021), rules are generated automatically, including contradictions and paradoxes. While the system does not need to explicitly generate the chain of reasoning, the traceability of the answer is still evaluated.

Context Independent Semantics Benchmarks

In the benchmarks reviewed below, entailment relations require the semantic interpretation of textual statements. With a minimal degree of logical complexity (left area in the figure), we find textual entailment benchmarks. Some examples include SNLI (Bowman, Angeli, Potts, & Manning, 2015), MULTI-NLI (Williams, Nangia, & Bowman, 2018) and CONJ-NLI (Saha, Nie, & Bansal, 2020). The latter introduces a higher degree of formal complexity through the use of coordinating conjunctions. Formal complexity increases significantly in multiple sentence entailment benchmarks such as MPE (Lai, Bisk, & Hockenmaier, 2017), and in entailment tree benchmarks such as ENTAILMENT-BANK (Dalvi et al., 2021). FOLIO, in turn, incorporates the translation of natural language into formal languages (Han et al., 2024).

In other benchmarks, the task consists of describing relations between concepts or statements in textual format, as in COMMONGEN (Lin et al., 2020) or E-CARE (Du, Ding, Xiong, Liu, & Qin, 2022). Within tasks that require sentence level semantic interpretation, the context can be further enriched by introducing conversational settings, as in the case of ENTITY DEDUCTION GAME (Zhang, Lu, & Jaitly, 2024). These problems are more demanding than NLI tasks in terms of meaning reduction, although not in terms of formal complexity, so we place them above the previous group.

Another category comprises logical implication question answering benchmarks in which the answer requires deducing implication relations between facts. These tasks involve greater formal complexity and a higher degree of meaning reduction than NLI. In COMPLEXWEBQUESTION, questions are generated from information extracted from the web (Talmor & Berant, 2018). In LOGICBENCH, formal complexity is increased by constructing questions from ontologies (Parmar et al., 2024). COUNTERFACTUAL further complicates semantic reduction by creating an alternative world context in which entity types are preserved while their relations are altered (Wu et al., 2024).

Finally, there are also benchmark datasets based on questionnaires that require chaining world knowledge facts, such as C-EVAL (Huang et al., 2023) or ARB (Sawada et al., 2023). Some of these combine scientific knowledge with common sense, such as OPENBOOKQA (Mihaylov, Clark, Khot, & Sabharwal, 2018), while others select questions that require multi step reasoning to reach the answer, as in STRATEGYQA (Geva et al., 2021). These benchmarks are more challenging than NLI in terms of both formal complexity and meaning reduction. However, they do not reach the level of formal complexity of logical implication question answering or benchmarks with explicit relational structures.

Context Dependent Semantics Benchmarks

Context dependent semantics benchmarks are those in which the task requires more than the semantic interpretation of individual sentences, as meaning reduction is governed by a specific semantic context.

Some of these scenarios, which can be referred to as *scenario grounded reasoning benchmarks*, are explicitly defined in the test data, such as family relations in CLUTRR (Sinha, Sodhani, Dong, Pineau, & Hamilton, 2019), human mental states in OPENTOM (Xu, Zhao, Zhu, Du, & He, 2024), common sense reasoning in real world scenarios as in SOCIALIQA (Sap, Rashkin, Chen, Le Bras, & Choi, 2019) or PIQA (Bisk, Zellers, Bras, Gao, & Choi, 2019), or contexts requiring temporal reasoning as in TIMEDIAL (Qin et al., 2021) or TRACIE (Zhou et al., 2021). In these benchmarks,

concepts acquire meaning and relational structure within a specific context, although one that remains largely explicit. The benchmark RECLOR also imposes a context, but it can be assigned to a separate category of argumentative reasoning, since the context is a paragraph. We place these benchmarks at the third level, with a level of formal complexity similar to FOLIO or ENTAILMENTBANK.

It is also worth highlighting the group of inductive reasoning benchmarks, where instead of applying complex formal structures, the goal is to abstract simple structures from textual information. This applies to hypothesis induction from social science papers on the benchmark used for evaluating MOOSE (Yang, Du, et al., 2024), or to induction of natural language rules from factual statements on DEER (Yang, Dong, et al., 2024). The formal complexity in these cases is simpler than in the previous benchmarks, but the challenge in terms of meaning reduction is higher than in any of the earlier benchmarks.

The Trade-off between Formal Complexity and Contextual Scope

Although our analysis does not apply precise quantitative criteria to distribute the benchmarks within a space, the results of our pairwise comparisons along two dimensions allows us to make certain observations:

(i) When relational structures are explicitly specified in the input, either as logical or mathematical problems, it becomes relatively straightforward to design synthetic benchmarks with arbitrarily high levels of formal complexity. Current LLMs already exhibit strong performance in these settings, often reaching human level results.

(ii) When entailment relations depend on the semantic interpretation of texts rather than on explicit formal structures, as in the Context Independent Semantics level of Figure 2, the formal complexity of tasks is significantly lower. In such cases, complexity is typically limited to a sequence of reasoning steps in question answering tasks or to a set of relations defined by an ontology.

(iii) Finally, at the Context Dependent Semantics level, we find benchmarks that emulate a context, whether social, temporal, or grounded in real world relations among entities. In these scenarios, formal complexity decreases even further, being reduced to a small set of predefined entities and relations.

Based on these observations, we conclude that in current benchmarks there appears to be a trade off between the formal complexity and the level of context awareness required by the tasks. As the formal complexity of the statements and relations involved in solving a problem increases, these relations tend to be made more explicit in the input. Conversely, in cases where strong meaning reduction based on context is required, formal complexity tends to decrease.

Conclusions

According to our analysis, the notions of meaning generalization, generation, and reduction make it possible to structure the cognitive competences of AI from the standpoint of the system itself, aligning with the fundamental mechanisms that have shaped different paradigms of intelligent systems, such as discriminative, generative, and recent neurosymbolic approaches. In particular, reducing meanings in such a way that inferences can be performed constitutes a foundational competence that underpins, first and foremost, different forms of formal reasoning, such as deduction, induction, and abduction. Furthermore, the ability to dynamically reduce meaning across contexts supports higher-level abilities, such as flexible reasoning, maintenance of conceptual coherence across tasks, and adaptation to new situations, where current AI systems, according to the literature, still show notable limitations.

Our analysis of recent benchmarks shows that some tasks require meaning reduction to infer the implication relations needed for reasoning. However, highly formal tasks, such as mathematical or logical problems, usually make these relations explicit, whereas tasks requiring their automatic inference tend to have lower formal complexity. We therefore argue that current benchmarks involve a trade off between the contextual scope of meaning reduction and the formal complexity of the problems addressed.

As a result, the ability of LLMs to build world models in specific contexts that require complex reasoning remains largely unquantified. We do not claim that LLMs lack this ability, but that designing reliable benchmarks to measure it is still an open challenge. Persistent problems such as hallucinations, lack of precision, and weak justification may arise in contexts where knowledge and reasoning structures are not explicitly represented in the training data. This reinforces the need for benchmarks that evaluate these capabilities.

In summary, despite clear progress in areas such as machine translation, image and video generation, or mathematical problem solving, the AI capabilities remain largely unquantified in real complex domain specific environments where no standardized formal languages apply. Benchmarks are therefore needed to evaluate meaning reduction in specific contexts and system performance on complex real world problems. These are not exceptional cases, but common scenarios for AI deployment in industrial settings.

Acknowledgments

This work was supported by the Spanish Ministry of Science, Innovation and Universities (projects ANNOTATE (PID2024-156022OB-C31) and TRAIL-LLMs/HumanAIze (AIA2025-163322-C65), both funded by MICIU/AEI/10.13039/501100011033 and the European Social Fund Plus (ESF+)), by ODESIA – Observatory for Artificial Intelligence in Spanish, RED.ES Agreement CO39/21-OT, partially funded by European MRR funds, and the European Research Council under Horizon Europe, grant number 10113572 (LUMINOUS).

References

Andrews, G., & Halford, G. S. (2002). A cognitive complexity metric applied to cognitive development. *Cognitive Psychology*, *45*(2), 153-219. doi: 10.1016/S0010-0285(02)00002-6

Barwise, J., & Perry, J. (1981). Situations and attitudes. *The Journal of Philosophy*, *78*(11), 668–691. Retrieved 2025-12-12, from http://www.jstor.org/stable/2026578

Bisk, Y., Zellers, R., Bras, R. L., Gao, J., & Choi, Y. (2019). *Piqa: Reasoning about physical commonsense in natural language.* Retrieved from https://arxiv.org/abs/1911.11641

Bowman, S. R., Angeli, G., Potts, C., & Manning, C. D. (2015, September). A large annotated corpus for learning natural language inference. In L. Màrquez, C. Callison-Burch, & J. Su (Eds.), *Proceedings of the 2015 conference on empirical methods in natural language processing* (pp. 632–642). Lisbon, Portugal: Association for Computational Linguistics. Retrieved from https://aclanthology.org/D15-1075/ doi: 10.18653/v1/D15-1075

Chen, M. K., Zhang, X., & Tao, D. (2025). *Justlogic: A comprehensive benchmark for evaluating deductive reasoning in large language models.* Retrieved from https://arxiv.org/abs/2501.14851

Chen, X., Chi, R. A., Wang, X., & Zhou, D. (2024). *Premise order matters in reasoning with large language models.* Retrieved from https://arxiv.org/abs/2402.08939

Chiang, C.-H., & Lee, H.-y. (2024, March). Over-reasoning and redundant calculation of large language models. In Y. Graham & M. Purver (Eds.), *Proceedings of the 18th conference of the european chapter of the association for computational linguistics (volume 2: Short papers)* (pp. 161–169). St. Julian's, Malta: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2024.eacl-short.15/

Clark, P., Tafjord, O., & Richardson, K. (2020). *Transformers as soft reasoners over language.* Retrieved from https://arxiv.org/abs/2002.05867

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., ... Schulman, J. (2021). Training verifiers to solve math word problems. *ArXiv*, *abs/2110.14168*. Retrieved from https://api.semanticscholar.org/CorpusID:239998651

Dalvi, B., Jansen, P., Tafjord, O., Xie, Z., Smith, H., Pipatanangkura, L., & Clark, P. (2021, November). Explaining answers with entailment trees. In M.-F. Moens, X. Huang, L. Specia, & S. W.-t. Yih (Eds.), *Proceedings of the 2021 conference on empirical methods in natural language processing* (pp. 7358–7370). Online and Punta Cana, Dominican Republic: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.emnlp-main.585/ doi: 10.18653/v1/2021.emnlp-main.585

Dasgupta, I., Lampinen, A. K., Chan, S. C. Y., Sheahan, H. R., Creswell, A., Kumaran, D., ... Hill, F. (2024). *Language models show human-like content effects on reasoning tasks.* Retrieved from https://arxiv.org/abs/2207.07051

Du, L., Ding, X., Xiong, K., Liu, T., & Qin, B. (2022, May). e-CARE: a new dataset for exploring explainable causal reasoning. In S. Muresan, P. Nakov, & A. Villavicencio (Eds.), *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 432–446). Dublin, Ireland: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2022.acl-long.33/ doi: 10.18653/v1/2022.acl-long.33

Eisape, T., Tessler, M., Dasgupta, I., Sha, F., Steenkiste, S., & Linzen, T. (2024, June). A systematic comparison of syllogistic reasoning in humans and language models. In K. Duh, H. Gomez, & S. Bethard (Eds.), *Proceedings of the 2024 conference of the north american chapter of the association for computational linguistics: Human language technologies (volume 1: Long papers)* (pp. 8425–8444). Mexico City, Mexico: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2024.naacl-long.466/ doi: 10.18653/v1/2024.naacl-long.466

Gelepithis, P. A. (1988). Survey of theories of meaning. *Cognitive Systems*, *2*(2), 141–162.

Geva, M., Khashabi, D., Segal, E., Khot, T., Roth, D., & Berant, J. (2021). Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, *9*, 346–361. Retrieved from https://aclanthology.org/2021.tacl-1.21/

Goodman, N. D., & Lassiter, D. (2015). Probabilistic semantics and pragmatics uncertainty in language and thought. In *The handbook of contemporary semantic theory* (p. 655-686). John Wiley & Sons, Ltd. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118882139.ch21

Guo, Z., Jin, R., Liu, C., Huang, Y., Shi, D., Supryadi, ... Xiong, D. (2023). *Evaluating large language models: A comprehensive survey.*

Halford, G. S., Wilson, W. H., & Phillips, S. (1998). Processing capacity defined by relational complexity: Implications for comparative, developmental, and cognitive psychology. *Behavioral and Brain Sciences*, *21*(6), 803–831. doi: 10.1017/S0140525X98001769

Han, S., Schoelkopf, H., Zhao, Y., Qi, Z., Riddell, M., Zhou, W., ... Radev, D. (2024, November). FOLIO: Natural language reasoning with first-order logic. In Y. Al-Onaizan, M. Bansal, & Y.-N. Chen (Eds.), *Proceedings of the 2024 conference on empirical methods in natural language processing* (pp. 22017–22031). Miami, Florida, USA: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2024.emnlp-main.1229/ doi: 10.18653/v1/2024.emnlp-main.1229

Hendrycks, D., Burns, C., Kadavath, S., Arora, A., Basart, S., Tang, E., ... Steinhardt, J. (2021). *Measuring mathematical problem solving with the math dataset.* Retrieved from https://arxiv.org/abs/2103.03874

Huang, Y., Bai, Y., Zhu, Z., Zhang, J., Zhang, J., Su, T., ... He, J. (2023). *C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models.* Retrieved from https://arxiv.org/abs/2305.08322

Jiang, J., Wang, J., Yan, Y., Liu, Y., Zhu, J., Zhang, M., & Gao, L. (2025, November). Do large language models excel in complex logical reasoning with formal language? In C. Christodoulopoulos, T. Chakraborty, C. Rose, & V. Peng (Eds.), *Proceedings of the 2025 conference on empirical methods in natural language processing* (pp. 16889–16914). Suzhou, China: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2025.emnlp-main.855/ doi: 10.18653/v1/2025.emnlp-main.855

Johnson-Laird, P. (2006). Deductive reasoning. In *Encyclopedia of cognitive science.* John Wiley Sons, Ltd. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1002/0470018860.s00512 doi: https://doi.org/10.1002/0470018860.s00512

Kazemi, M., Fatemi, B., Bansal, H., Palowitch, J., Anastasiou, C., Mehta, S. V., ... Firat, O. (2025, July). BIG-bench extra hard. In W. Che, J. Nabende, E. Shutova, & M. T. Pilehvar (Eds.), *Proceedings of the 63rd annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 26473–26501). Vienna, Austria: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2025.acl-long.1285/ doi: 10.18653/v1/2025.acl-long.1285

Kim, J., Podlasek, A., Shidara, K., Liu, F., Alaa, A., & Bernardo, D. (2025, 11). Limitations of large language models in clinical problem-solving arising from inflexible reasoning. *Scientific Reports*, *15*. doi: 10.1038/s41598-025-22940-0

Lai, A., Bisk, Y., & Hockenmaier, J. (2017, November). Natural language inference from multiple premises. In G. Kondrak & T. Watanabe (Eds.), *Proceedings of the eighth international joint conference on natural language processing (volume 1: Long papers)* (pp. 100–109). Taipei, Taiwan: Asian Federation of Natural Language Processing. Retrieved from https://aclanthology.org/I17-1011/

Lin, B. Y., Zhou, W., Shen, M., Zhou, P., Bhagavatula, C., Choi, Y., & Ren, X. (2020, November). CommonGen: A constrained text generation challenge for generative commonsense reasoning. In T. Cohn, Y. He, & Y. Liu (Eds.), *Findings of the association for computational linguistics: Emnlp 2020* (pp. 1823–1840). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2020.findings-emnlp.165/ doi: 10.18653/v1/2020.findings-emnlp.165

Liu, H., Fu, Z., Ding, M., Ning, R., Zhang, C., Liu, X., & Zhang, Y. (2025). *Logical reasoning in large language models: A survey.* Retrieved from https://arxiv.org/abs/2502.09100

Liu, J., Cui, L., Liu, H., Huang, D., Wang, Y., & Zhang, Y. (2020). *Logiqa: A challenge dataset for machine reading comprehension with logical reasoning.* Retrieved from https://arxiv.org/abs/2007.08124

Liu, Z., Yu, D., & He, H. (2025, November). On the role of model prior in real-world inductive reasoning. In C. Christodoulopoulos, T. Chakraborty, C. Rose, & V. Peng (Eds.), *Proceedings of the 2025 conference on empirical methods in natural language processing* (pp. 10560–10583). Suzhou, China: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2025.emnlp-main.534/ doi: 10.18653/v1/2025.emnlp-main.534

Lycan, W. G. (2018). *Philosophy of language: A contemporary introduction*. Routledge.

MacCartney, B., & Manning, C. D. (2007, June). Natural logic for textual inference. In S. Sekine, K. Inui, I. Dagan, B. Dolan, D. Giampiccolo, & B. Magnini (Eds.), *Proceedings of the ACL-PASCAL workshop on textual entailment and paraphrasing* (pp. 193–200). Prague: Association for Computational Linguistics. Retrieved from https://aclanthology.org/W07-1431/

McCoy, R. T., Yao, S., Friedman, D., Hardy, M. D., & Griffiths, T. L. (2024). Embers of autoregression show how large language models are shaped by the problem they are trained to solve. *Proceedings of the National Academy of Sciences*, *121*(41), e2322420121. Retrieved from https://www.pnas.org/doi/abs/10.1073/pnas.2322420121

Mihaylov, T., Clark, P., Khot, T., & Sabharwal, A. (2018, October-November). Can a suit of armor conduct electricity? a new dataset for open book question answering. In E. Riloff, D. Chiang, J. Hockenmaier, & J. Tsujii (Eds.), *Proceedings of the 2018 conference on empirical methods in natural language processing* (pp. 2381–2391). Brussels, Belgium: Association for Computational Linguistics. Retrieved from https://aclanthology.org/D18-1260/ doi: 10.18653/v1/D18-1260

Millière, R., & Rathkopf, C. (2026, 01). Anthropocentric bias in language model evaluation. *Computational Linguistics*, 1-10. Retrieved from https://doi.org/10.1162/COLI.a.582 doi: 10.1162/COLI.a.582

Mitchell, M. (2019). *Artificial intelligence: a guide for thinking humans*. New York: Farrar, Straus and Giroux. Retrieved from https://melaniemitchell.me/aibook/

Montague, R. (1973). The proper treatment of quantification in ordinary English. In K. J. J. Hintikka, J. Moravcsic, & P. Suppes (Eds.), *Approaches to natural language* (p. 221-242). Dordrecht: Reidel.

Morishita, T., Morio, G., Yamaguchi, A., & Sogawa, Y. (2023). *Learning deductive reasoning from synthetic corpus based on formal logic.* Retrieved from

https://arxiv.org/abs/2308.07336

Ontanon, S., Ainslie, J., Cvicek, V., & Fisher, Z. (2022). *Logicinference: A new dataset for teaching logical inference to seq2seq models.* Retrieved from https://arxiv.org/abs/2203.15099

Parmar, M., Patel, N., Varshney, N., Nakamura, M., Luo, M., Mashetty, S., ... Baral, C. (2024, August). LogicBench: Towards systematic evaluation of logical reasoning ability of large language models. In L.-W. Ku, A. Martins, & V. Srikumar (Eds.), *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 13679–13707). Bangkok, Thailand: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2024.acl-long.739/ doi: 10.18653/v1/2024.acl-long.739

Qin, L., Gupta, A., Upadhyay, S., He, L., Choi, Y., & Faruqui, M. (2021, August). TIMEDIAL: Temporal commonsense reasoning in dialog. In C. Zong, F. Xia, W. Li, & R. Navigli (Eds.), *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: Long papers)* (pp. 7066–7076). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.acl-long.549/ doi: 10.18653/v1/2021.acl-long.549

Saha, S., Nie, Y., & Bansal, M. (2020, November). ConjNLI: Natural language inference over conjunctive sentences. In B. Webber, T. Cohn, Y. He, & Y. Liu (Eds.), *Proceedings of the 2020 conference on empirical methods in natural language processing (emnlp)* (pp. 8240–8252). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2020.emnlp-main.661/ doi: 10.18653/v1/2020.emnlp-main.661

Salto, F., Requena, C., Álvarez-Merino, P., Antón-Toro, L. F., & Maestú, F. (2021). Brain electrical traits of logical validity. *Scientific Reports*, *11*. Retrieved from https://api.semanticscholar.org/CorpusID:233221470

Sap, M., Rashkin, H., Chen, D., Le Bras, R., & Choi, Y. (2019, November). Social IQa: Commonsense reasoning about social interactions. In K. Inui, J. Jiang, V. Ng, & X. Wan (Eds.), *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (emnlp-ijcnlp)* (pp. 4463–4473). Hong Kong, China: Association for Computational Linguistics. Retrieved from https://aclanthology.org/D19-1454/ doi: 10.18653/v1/D19-1454

Saparov, A., & He, H. (2023). *Language models are greedy reasoners: A systematic formal analysis of chain-of-thought.* Retrieved from https://arxiv.org/abs/2210.01240

Sawada, T., Paleka, D., Havrilla, A., Tadepalli, P., Vidas, P., Kranias, A., ... Komatsuzaki, A. (2023). *Arb: Advanced reasoning benchmark for large language models.* Retrieved from https://arxiv.org/abs/2307.13692

Schrenk, M. (2008). Verificationist theory of meaning. In U. Windhorst, M. Binder, & N. Hirowaka (Eds.), *Encyclopaedic reference of neuroscience.* Springer.

Sinha, K., Sodhani, S., Dong, J., Pineau, J., & Hamilton, W. L. (2019, November). CLUTRR: A diagnostic benchmark for inductive reasoning from text. In K. Inui, J. Jiang, V. Ng, & X. Wan (Eds.), *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (emnlp-ijcnlp)* (pp. 4506–4515). Hong Kong, China: Association for Computational Linguistics. Retrieved from https://aclanthology.org/D19-1458/ doi: 10.18653/v1/D19-1458

Srivastava, S., B, A. M., V, A. P., Menon, S., Sukumar, A., T, A. S., ... Thomas, S. (2024). *Functional benchmarks for robust evaluation of reasoning performance, and the reasoning gap.* Retrieved from https://arxiv.org/abs/2402.19450

Stolfo, A., Jin, Z., Shridhar, K., Schoelkopf, B., & Sachan, M. (2023, July). A causal framework to quantify the robustness of mathematical reasoning with language models. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 545–561). Toronto, Canada: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2023.acl-long.32/ doi: 10.18653/v1/2023.acl-long.32

Suresh, S., Mukherjee, K., Yu, X., Huang, W.-C., Padua, L., & Rogers, T. (2023, December). Conceptual structure coheres in human cognition but not in large language models. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Proceedings of the 2023 conference on empirical methods in natural language processing* (pp. 722–738). Singapore: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2023.emnlp-main.47/ doi: 10.18653/v1/2023.emnlp-main.47

Tafjord, O., Dalvi, B., & Clark, P. (2021, August). ProofWriter: Generating implications, proofs, and abductive statements over natural language. In C. Zong, F. Xia, W. Li, & R. Navigli (Eds.), *Findings of the association for computational linguistics: Acl-ijcnlp 2021* (pp. 3621–3634). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.findings-acl.317/ doi: 10.18653/v1/2021.findings-acl.317

Talmor, A., & Berant, J. (2018, June). The web as a knowledge-base for answering complex questions. In M. Walker, H. Ji, & A. Stent (Eds.), *Proceedings of the 2018 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (long papers)* (pp. 641–651). New Orleans, Louisiana: As-

sociation for Computational Linguistics. Retrieved from https://aclanthology.org/N18-1059/ doi: 10.18653/v1/N18-1059

Tang, X., Zheng, Z., Li, J., Meng, F., Zhu, S.-C., Liang, Y., & Zhang, M. (2023). *Large language models are in-context semantic reasoners rather than symbolic reasoners.* Retrieved from https://arxiv.org/abs/2305.14825

Tian, J., Li, Y., Chen, W., Xiao, L., He, H., & Jin, Y. (2021, November). Diagnosing the first-order logical reasoning ability through LogicNLI. In M.-F. Moens, X. Huang, L. Specia, & S. W.-t. Yih (Eds.), *Proceedings of the 2021 conference on empirical methods in natural language processing* (pp. 3738–3747). Online and Punta Cana, Dominican Republic: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.emnlp-main.303/ doi: 10.18653/v1/2021.emnlp-main.303

Tian, S.-Y., Zhou, Z., Yu, K.-Y., Yang, M., Jia, L.-H., Guo, L.-Z., & Li, Y.-F. (2025, November). VCSearch: Bridging the gap between well-defined and ill-defined problems in mathematical reasoning. In C. Christodoulopoulos, T. Chakraborty, C. Rose, & V. Peng (Eds.), *Proceedings of the 2025 conference on empirical methods in natural language processing* (pp. 12721–12742). Suzhou, China: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2025.emnlp-main.642/ doi: 10.18653/v1/2025.emnlp-main.642

Valiente, R., & Pilly, P. (2025, 09). Metacognition for unknown situations and environments (muse). *Neural Networks*, *194*, 108131. doi: 10.1016/j.neunet.2025.108131

Valmeekam, K., Marquez, M., Olmo, A., Sreedharan, S., & Kambhampati, S. (2023). Planbench: an extensible benchmark for evaluating large language models on planning and reasoning about change. In *Proceedings of the 37th international conference on neural information processing systems.* Red Hook, NY, USA: Curran Associates Inc.

Wan, Y., Wang, W., Yang, Y., Yuan, Y., Huang, J.-t., He, P., ... Lyu, M. (2024, November). LogicAsker: Evaluating and improving the logical reasoning ability of large language models. In Y. Al-Onaizan, M. Bansal, & Y.-N. Chen (Eds.), *Proceedings of the 2024 conference on empirical methods in natural language processing* (pp. 2124–2155). Miami, Florida, USA: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2024.emnlp-main.128/ doi: 10.18653/v1/2024.emnlp-main.128

Wei, T., Li, T.-W., Liu, Z., Ning, X., Yang, Z., Zou, J., ... He, J. (2026). *Agentic reasoning for large language models.* Retrieved from https://arxiv.org/abs/2601.12538

Weston, J., Bordes, A., Chopra, S., Rush, A. M., van Merriënboer, B., Joulin, A., & Mikolov, T. (2015). *Towards ai-complete question answering: A set of prerequisite toy tasks.* Retrieved from https://arxiv.org/abs/1502.05698

Williams, A., Nangia, N., & Bowman, S. (2018, June). A broad-coverage challenge corpus for sentence understanding through inference. In M. Walker, H. Ji, & A. Stent (Eds.), *Proceedings of the 2018 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (long papers)* (pp. 1112–1122). New Orleans, Louisiana: Association for Computational Linguistics. Retrieved from https://aclanthology.org/N18-1101/ doi: 10.18653/v1/N18-1101

Wu, Z., Qiu, L., Ross, A., Akyürek, E., Chen, B., Wang, B., ... Kim, Y. (2024, June). Reasoning or reciting? exploring the capabilities and limitations of language models through counterfactual tasks. In K. Duh, H. Gomez, & S. Bethard (Eds.), *Proceedings of the 2024 conference of the north american chapter of the association for computational linguistics: Human language technologies (volume 1: Long papers)* (pp. 1819–1862). Mexico City, Mexico: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2024.naacl-long.102/ doi: 10.18653/v1/2024.naacl-long.102

Xu, H., Zhao, R., Zhu, L., Du, J., & He, Y. (2024, August). OpenToM: A comprehensive benchmark for evaluating theory-of-mind reasoning capabilities of large language models. In L.-W. Ku, A. Martins, & V. Srikumar (Eds.), *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 8593–8623). Bangkok, Thailand: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2024.acl-long.466/ doi: 10.18653/v1/2024.acl-long.466

Yang, Z., Dong, L., Du, X., Cheng, H., Cambria, E., Liu, X., ... Wei, F. (2024, March). Language models as inductive reasoners. In Y. Graham & M. Purver (Eds.), *Proceedings of the 18th conference of the european chapter of the association for computational linguistics (volume 1: Long papers)* (pp. 209–225). St. Julian's, Malta: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2024.eacl-long.13/

Yang, Z., Du, X., Li, J., Zheng, J., Poria, S., & Cambria, E. (2024, August). Large language models for automated open-domain scientific hypotheses discovery. In L.-W. Ku, A. Martins, & V. Srikumar (Eds.), *Findings of the association for computational linguistics: Acl 2024* (pp. 13545–13565). Bangkok, Thailand: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2024.findings-acl.804/ doi: 10.18653/v1/2024.findings-acl.804

Zhang, Y., Lu, J., & Jaitly, N. (2024, August). Probing the multi-turn planning capabilities of LLMs via 20 question games. In L.-W. Ku, A. Martins, & V. Srikumar (Eds.), *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 1495–1516). Bangkok, Thailand: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2024.acl-long.82/ doi:

10.18653/v1/2024.acl-long.82

Zhou, B., Richardson, K., Ning, Q., Khot, T., Sabharwal, A., & Roth, D. (2021, June). Temporal reasoning on implicit events from distant supervision. In K. Toutanova et al. (Eds.), *Proceedings of the 2021 conference of the north american chapter of the association for computational linguistics: Human language technologies* (pp. 1361–1371). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.naacl-main.107/ doi: 10.18653/v1/2021.naacl-main.107

Zhu, R., Guo, D., Qi, D., Chu, Z., Yu, X., & Li, S. (2024, June). A survey of trustworthy representation learning across domains. *ACM Trans. Knowl. Discov. Data*, *18*(7). Retrieved from https://doi.org/10.1145/3657301 doi: 10.1145/3657301