Do large language models solve verbal analogies like children do?

Tamar Johnson[◇], **Mathilde ter Veen**[◇], **Rochelle Choenni**[†],
Han L. J. van der Maas[◇] **Ekaterina Shutova**[†] **and Claire E. Stevenson**[◇],
[◇]Psychological Methods, University of Amsterdam, the Netherlands
[†]ILLC, University of Amsterdam, the Netherlands
t.johnson@uva.nl
c.e.stevenson@uva.nl

Abstract

Analogy-making lies at the heart of human cognition. Adults solve analogies such as *horse belongs to stable like chicken belongs to . . . ?* by mapping relations (*kept in*) and answering *chicken coop*. In contrast, young children often use association, e.g., answering *egg*. This paper investigates whether large language models (LLMs) solve verbal analogies in A:B::C:? form using associations, similar to what children do. We use verbal analogies extracted from an online learning environment, where 14,006 7-12 year-olds from the Netherlands solved 872 analogies in Dutch. The eight tested LLMs performed at or above the level of children, with some models approaching adult performance estimates. However, when we control for solving by association this picture changes. We conclude that the LLMs we tested rely heavily on association like young children do. However, LLMs make different errors than children, and association doesn't fully explain their superior performance on this children's verbal analogy task. Future work will investigate whether LLMs associations and errors are more similar to adult relational reasoning.

1 Introduction

Analogy-making, using what you know about one thing to infer knowledge about a new, somehow related instance, lies at the heart of human intelligence and creativity and forms the core of educational practice (Gentner, 1988; Hofstadter, 1997; Holyoak, 2012). Given how important analogical reasoning is to learning and generalization, much research has focused on how this seemingly unique human ability emerges, develops, and can be improved (Goswami, 1991; Sternberg and Nigro, 1980; Stevenson and Hickendorff, 2018) as well as emulated in machines (Gentner and Forbus, 2011; Mitchell, 2021). Recently, large language models (LLMs), such as GPT-3 (Brown et al., 2020), have demonstrated surprisingly good performance in verbal analogy solving (e.g., *table is to legs as tree is to . . . ? chair, leaves, branches or roots?*) (Lu et al., 2022; Webb et al., 2023). The question then arises *how* LLMs solve these analogies. Is it similar to adult humans using relational mapping? Or perhaps more similar to the associative processes children tend to use?

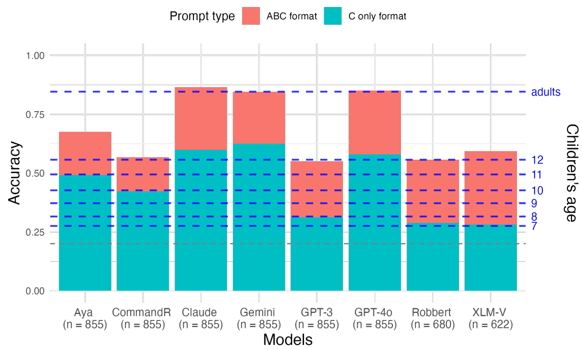


Figure 1: How well does each LLM perform? We see that when prompted with A:B::C:? many LLMs outperform children. However, LLMs can also solve most items by association, evidenced by correctly solving analogies when only prompted with C:?.

Earlier work shows that language models largely rely on semantic similarity between analogy terms to solve analogies (Rogers et al., 2020; Ushio et al., 2021b), which would indicate solving by association. In this paper we investigate whether LLMs use association or analogy to solve a set of Dutch verbal analogies. First, we examine how LLM performance compares to children and find that the best models outperform 12-year-olds, approaching adult performance estimates. Second, we examine whether LLM performance is influenced by the same item characteristics that affect children's analogy solving, where results confirmed that this is indeed the case, especially for lower performing models. Third, through a series of prompting experiments we show that these LLMs appear to use association to solve a large proportion of analogies.

Fourth, we compare error patterns of children with LLMs and find that LLMs are far more similar to each other (and those of similar architecture and size) than to children.

This paper contributes to the study of analogical reasoning in LLMs in three ways: (1) it is the first to directly compare LLM verbal analogy solving performance to that of children; (2) we use experiments to tap into whether LLMs solve analogies using association like young children; and (3) we use Dutch rather than English language items and examine performance in multilingual LLMs.

2 Theoretical Background

2.1 The Analogical Reasoning Process

Although there are different cognitive models of analogical reasoning—varying in the order of processing steps and whether these occur sequentially or in parallel—there is a general consensus on which processes are involved. Taking the example of *"body is to feet as tree is to . . . ?"* (or more abstractly, A:B::C:?), the basic analogy information processing steps are generally considered to be: (1) encoding relevant information about the base (A:B) and target (C) domains; (2) searching and retrieving relationships and similarities between the analogy elements in the base domain, A and B (e.g., *"stands on"* for body and feet); (3) aligning the base and target domains (*"body and tree are things that stand"*) and mapping the mostly likely relationship between A and B, to the target domain, C, to come up with D; and (4) evaluating the validity of the predicted solution (Gentner and Hoyos, 2017; Sternberg, 1977; Thibaut and French, 2016).

2.2 Factors Affecting People's Verbal Analogy Solving

The basic analogy solving steps are consistently found in people from about 12 years and up (Thibaut and French, 2016). When adults make mistakes there are three main factors that lead to errors: (1) the relation type (causal is more difficult than categorical), (2) a large conceptual distance between analogy base and target domains, and (3) salient distractors amongst the multiple-choice options (Jones et al., 2022).

Type of Relation Jones et al. (2022) grouped analogical relations into three types: categorical, causal and compositional. They found that adults perform better on categorical analogies (e.g., *tarantula:spider::bee:insect*) than causal (e.g., *fracture:cast::incision:scar*) or compositional (e.g., *fingernail:finger::knee:leg*) analogies. Children's performance follows a similar pattern, assuming sufficient domain knowledge is in place (e.g., Sternberg and Nigro, 1980; Goswami and Brown, 1990; Alexander and Kulikowisch, 1991).

Conceptual Distance Between Base and Target Domains The greater the distance between an analogy base and target domain the more difficult the analogy is for adults and children to solve (Jones et al., 2022; Thibaut and French, 2016). For example, *bowl:dish::spoon:silverware* is easier for people to solve than *wrench:tool::sad:mood*.

Distractor Salience People are sometimes lured to choose a distracting incorrect response in multiple choice verbal analogies, and are most easily distracted by answer options that have a strong semantic association with the C term (Kucwaj et al., 2022). Jones et al. (2022) defines distractor salience as the relation between C:D relative to each of the C:D', where D' represents each distractor option. Distractor salience is high, when the semantic similarity between C and one of the incorrect answers D' is greater than the semantic similarity between C and the correct answer D. High distractor salience leads to lower performance in adults (Ichien et al., 2020; Jones et al., 2022) and this is even more apparent in children (Richland et al., 2006; Thibaut and French, 2016).

2.3 Analogical Reasoning Development

Children's verbal analogical reasoning improves with age, where a gradual shift occurs around 4-8 years of age from reasoning based on surface similarities and associations to reasoning based on (abstract) relations (Gentner, 1988; Stevenson and Hickendorff, 2018; Gentile et al., 1977). For example, if we ask a four-year-old *"horse belongs to stable like chicken belongs to . . . ?"* they may use association and reply *"egg"*, relying on the strong connection between the words chicken and egg to solve the problem. In contrast, older children and adults will likely give the intended relational response *"chicken coop"*, using the underlying relation structure to solve the analogy.

However, even when children can solve these analogies, two main factors that seem to affect the transition from associative to relational reasoning are increased domain knowledge (Goswami and

Brown, 1990; Gentner, 1988; Alexander and Kulikowisch, 1991) and improved executive functions (working memory and inhibition control; Doumas et al., 2018; Thibaut and French, 2016).

Children tend to fail in analogy solving if they are unfamiliar with the elements or relations in the analogy (Gentner and Hoyos, 2017; Goswami and Brown, 1990; Goddu et al., 2020). If children are shown to possess the required domain knowledge and are provided clear instructions on how to solve the task then they can successfully solve verbal analogies (in the form of pictures) as early as 3-years-old (Goswami, 1991; Goddu et al., 2020).

However, even when children can solve these analogies, evidence from scene analogy problems (Richland et al., 2006) and eye-tracking studies (Thibaut and French, 2016) shows that children up to 8 years-old tend to focus first on the C term when solving analogies, sometimes ignoring A and B altogether (Thibaut and French, 2016). This appears to be related to limited working memory capacity (Richland et al., 2006; Stevenson et al., 2013; Stevenson, 2017) and limits in inhibition- and executive control (Thibaut and French, 2016; Doumas et al., 2018). Performance improves when interventions are used that support children's processing capacities (Stevenson and Hickendorff, 2018) and when children are forced to focus first on the A:B pair (Glady et al., 2017).

2.4 Verbal Analogy Solving in LLMs

The extent to which LLMs can solve analogies is a subject of debate. Most of this work has focused on comparing models in terms of overall accuracy on benchmarks such as the Bigger Analogy Test Set (BATS; Mikolov et al., 2013b) and verbal analogies from the Scholastic Assessment doTest (SAT; Turney et al., 2003) and investigating the types of relations they can solve (e.g., syntactic versus semantic). More importantly, when LLMs demonstrate analogy solving abilities, it is unclear how they achieved these solutions (e.g., Webb et al., 2023), whether this is through relational reasoning or another process, such as the associative strategy often employed by young children.

Word embeddings Over a decade ago, Mikolov et al. (2013b) published their seminal paper showing that pre-trained word embeddings (e.g., Word2Vec Mikolov et al., 2013a) could be used to solve verbal analogies in the form of A:B::C:? using vector arithmetic, the most famous example being: $embed(king) - embed(man) + embed(woman) \approx embed(queen)$, where $embed$ represents the word embedding obtained from the pre-trained neural network. This milestone was tempered by Gladkova et al. (2016), who made clear that this method was limited in the breadth of relations that it could process. For example, the capitol-country relation was solved quite successfully, but others such as animal-sound and part-whole, were solved less successfully.

Transformer language models With the rise of the Transformer architecture, featuring language models such as BERT (Devlin et al., 2018), verbal analogy solving remained a challenge. Earlier work transferred the verbal analogy datasets, such as the BATS to the sentence level, and showed that BERT-based models and GPT-2 (Radford et al., 2019) performed at a similar level to GloVe (Pennington et al., 2014), a word embedding model, on analogies containing relations such as capitol-country and male–female pairs (Zhu and de Melo, 2020). More recently, Czinczoll et al. (2022) developed a dataset containing scientific and metaphor analogies (SCAN). Here there was a clear advantage of transformer models over analogy solving with word embeddings, where GPT-2, BERT and M-BERT outperformed GloVe on the analogy items containing metaphors such as *career:mountain::success:ascent*. Also, Petersen and van der Plas (2023) showed that by changing the training objective of LLMs to maximize relational similarity, LLM performance improves. Yet, the general conclusion remained that verbal analogy solving is more challenging for LLMs than people.

People versus LLMs in analogy solving Recent research has shown that LLMs can solve verbal analogies with similar accuracy to people. For example, Ushio et al. (2021b) showed that LLMs such as GPT-2 and RoBERTa generally perform well on analogies designed for 4th to 10th graders (9-16 year-olds). Also, Webb et al. (2023) concluded that GPT-3 and GPT-4 generally perform around the same level as adults on two verbal analogy datasets.

Item factors affecting LLM verbal analogy solving There has been some research on the effect of *relationship type* on LLM's verbal analogy solving performance. Ushio et al. (2021a) showed that fine-tuned RoBERTa models performed slightly better on categorical relations (hypernymns) than compositional ones (meronymns). And Webb et al. (2023)

found that categorical relations in the SAT verbal analogies were easier for GPT-3 than compositional (function) relations and also that categorical relations were easier than both compositional and causal relations on the items from Jones et al. (2022). Similarly, Linford et al. (2022) found that categorical relations were easier for BERT models than causal relations, although performance on both was far lower than for human adults.

Similarly to people, LLMs have more difficulty as the *conceptual distance* between the domains in the analogy increases. For example, the LLMs in Czinczoll et al. (2022) performed better on the BATS analogies than on their SCAN dataset comprising scientific and metaphor based analogies, where the semantic distance between the base and target domains was greater. In addition the scientific analogies were solved better by LLMs than those based on metaphors, which was explained by there being a clearer correspondence between base and target domains in scientific analogies. Also, Webb et al. (2023), used the items from Jones et al. (2022) to investigate whether, like in people, a near conceptual distance between the base and target domains made analogies easier to solve for GPT-3 than far analogies; this was indeed the case. Interestingly,do humans outperformed GPT-3 on the far analogies.

There is less research on the effect of *distractor salience* on LLM analogy solving. In Petersen and van der Plas (2023) their best performing trained model appeared unaffected by low versus high distractor salience. In Musker et al. (2024), analogy tasks presented in an in-context-learning setting with interleaved distractors affected LLMs more than human adults. We expect that salient distractors, i.e. multiple-choice options that are semantically more similar to the analogy terms than the correct response, will have a greater chance of being "selected" by the LLMs.

3 Research Questions

In this study, with pre-registered hypotheses and methods, we examine how 8 multilingual LLMs solve 872 verbal analogies, also solved by 14,006 in an online learning environment.

RQ1: How well do LLMs perform compared to children ages 7-12 in verbal analogy solving? We expected recent LLMs to solve the analogies with similar accuracy to older children (12-year-olds) as this is similar to adult performance (hypothesis 1; Webb et al., 2023; Ushio et al., 2021a).

RQ2: Which item characteristics influence children's and LLM performance on verbal analogies? We expected the pattern of results found in adults also to be found in children and in LLMs. A growing strand of work shows that children, from a very young age, are remarkably sensitive to distributional regularities in their input and are adept at learning from this type of information (e.g., Saffran et al., 1996; Bresnan, 2007; Clark, 2014). Given the similarity (albeit to a limited extent) to how LLMs extract and track information from their input, we investigate whether the two learners are affected by shared item characteristics. First, we expect performance on categorical relations to be better than compositional and causal relations for both children (Sternberg and Nigro, 1980, hypothesis 2a1) and LLMs (Webb et al., 2023, hypothesis 2a2). Second, we expect analogies with a near conceptual distance between A:B to be easier than far analogies for children (Thibaut and French (2016); Hypothesis 2b1) and LLMs (Czinczoll et al., 2022; Webb et al., 2023, hypothesis 2b2). Third, we expect higher distractor salience to lead to more errors in children (Thibaut and French, 2016, hypothesis 2c1) and LLMs (Ushio et al., 2021b, hypothesis 2c2).

RQ3: Do LLMs choose associative or analogical solutions? We investigate this through a series of experiments comparing LLM performance on alternative formulations of the verbal analogies, where we control for associative responses.

4 Methods

LLM data and code and a selection of the children's data is publicly available. The full dataset is available upon request from Prowise Learn, the company that provided the children's data on the verbal analogies dataset.

4.1 Prowise Learn's Verbal Analogies Game

Prowise Learn is an online adaptive learning environment for elementary school children.

Verbal analogies is one of the games on the platform (see Appendix A for a screenshot of the game). The analogies are presented as text in "A:B::C:?" format, and the children must choose among five answer options, all five of which are semantically associated with C. For more information see Appendix A.

Data Collection with Children For this study, we extracted information on 14,006 7-12 year-old's (M = 10.73, SD = 1.15 years) performance on 872 verbal analogies from the Prowise Learn database. We applied three selection criteria when extracting the children's data (on June 19, 2021): (1) children solved at least 20 items to ensure stable ability estimates, (2) children had last played the game on or after September 1st 2020, the start of the school year and 4 months after the launch of the game, when item difficulty estimates were verified to have small standard errors and (3) children were ages 7-12 to avoid confounds in performance (i.e., younger children most likely did not have sufficient reading abilities and older children had most likely repeated a grade). This data collection was approved by the University of Amsterdam's Ethics Review Board with id FMG-3037.

Data Collection with Adults To provide an estimate of adult-level performance on this children's verbal analogy task, we collected data from 120 Dutch-speaking adults (M = 29.20, SD = 9.96 years) through Prolific's academic participant recruitment system. Each person solved 30 analogies presented in a similar format to those of children, with the aim of having each item solved by 4 adults to estimate item-level performance. Also, to test to what extent adults solve analogies by association we administered 30 additional items in C-only format (see 7.1 for a description). We applied two inclusion criteria before analyzing the adults' data. First, we included adults that solved >=50% of items correctly (i.e., achieved at least average children's performance), which led to 3 participants being excluded. Second, we included adults who explicitly stated that they did not use AI-tools to solve the analogies (1 person excluded). This data collection was approved by the University of Amsterdam's Ethics Review Board with id FMG-3105.

Item Selection The game contained three types of verbal reasoning problems; verbal analogies was one of them. From the initial set of 872 verbal analogies, we checked all items that were outliers (>1.5 SD) on the item difficulty scale and removed 17 items that were judged by two independent raters to contain errors (e.g., multiple correct solutions, requiring domain knowledge likely unfamiliar to children). This resulted in 855 items for data analysis.

4.2 Item characteristics

Relation Type Relationship type refers to how the A and B term are related. This relationship is applied to the C-term to find D. Table 2 provides a selected overview of relation types in the analogy task[1]. For analyses related to RQ2 we selected 302 items that fall into the following three categories defined by Jones et al. (2022):

- **Categorical**: one of the A:B terms defines the category and the other word is an example of this category. For example, "yellow" is part of the category "color".
- **Causal**: one of the A:B terms is the cause and the other is the effect. For example, "stumbling" will result in "falling".
- **Compositional**: one of the A:B terms is part of the other term. For example, "leaf" is part of a "tree".

Conceptual Distance Between Base and Target Domains We used three vector-based language models[2] to compute the semantic distance (1 - cosine similarity) between the A:B and the C:D pair. We used the mean value over the three vector-based models as the selected category for each item for analysis.

Distractor Salience Distractor salience was measured by the cosine similarity between C and D minus the cosine similarity between C and each incorrect answer D'. We used the same three vector-based models from Section 4.2 to compute the cosine distances between embeddings for C and each of the five D's. Then we determined distractor salience per item for each vector model and used the mean value for analysis.

4.3 Analogy completion with LLMs

Pretrained Language Models We studied how 8 transformer-based multilingual LLMs solved the same set of verbal analogies as the children.

Two of the LLMs are BERT-based masked language models. **RobBERT** (Delobelle et al., 2020) was pretrained on Dutch data only, and a multilingual variant **XLM-V** (Liang et al., 2023) was

[1]These labels were chosen and annotated by the Prowise Learn item developers.

[2]Word2Vec trained by CLIPS on different Dutch corpora (Tulkens et al., 2016), Word2Vec trained by the Nordic Language Processing Laboratory on the CoNLL17 corpus (Kutuzov et al., 2017), and FastText trained on Common Crawl and Wikipedia (Grave et al., 2018).

trained on 116 languages.[3] Identical to BERT (Devlin et al., 2018), both models contain 12 layers with 12 attention heads each.

The other LLMs are autoregressive transformer-decoder based language models. The open-source models we use are **Aya** and **Command-R**, both accessed through the Cohere API. The proprietary models we use are Anthropic's **Claude Sonnet-3.5**, Google's **Gemini-2.0-flash**, and Open AI's **GPT-3** and **GPT-4o**, each accessed through the API provided by the respective company.

Analogy completion We wanted to mimic the way the children solved the analogies in the best way possible. This was especially important because we investigate whether an associative response is more likely in the presence of a correct response. Therefore, we prompted the generative LLMs with the full analogy and asked them to choose from the five response options. For example, "tripping is to falling as picking up is to ? Choose clean, junk, mess, room, or thrift store." The response options were presented in random order.

However, this method was not possible to implement for the BERT-based models. Therefore, for the RobBERT and XLM-V models we used the masked language model approach and fed the models '`A` is to `B`, as `C` is to `D`', replacing `D` with each possible multiple-choice solution. The `D` option with the highest probability for the completion was considered the selected response.

5 Results RQ1: How well do LLMs perform compared to children?

Figure 1 shows performance per model on the 872 items. We see that all tested models, both BERT-based and autoregressive transformer-decoder based language models, perform at or above the level of children on the multiple choice question verbal analogy task. Children already at the age of 7 perform higher than chance level (gray dashed line), with Aya , Command-R , GPT-3 , RobBERT and XLM-V around the same level as 12 years old, whereas Claude , Gemini and GPT-4o outperform all children and other models, and perform at the level of adults on this task.

We analyzed how many of the items LLMs could solve by word association and report their performance on the C:? task (Experiment 1, see also 7). Results show that for the autoregressive transformer-decoder based models, word association can explain most of their success, but also in other models a large portion of items can be solved solely by association (Figure 1, blue portion of the bars). See 7 for further details and conclusions.

6 Results RQ2: Which item factors influence analogy solving?

For RQ2, we tested the effects of solver (children, LLMs) and/or item characteristics on accuracy. 9 logistic regression models (one per each solver type) predicted the performance on each item by relation type, semantic distance between base and target domains and distractor salience. We also included by-item intercepts as random effects.

Relation Type Logistic regression analyses of children's performance revealed significant effect of relation type ($beta = -0.49, z = -2.20, p < 0.05$) such that compositional items were easier than causal items for children to solve, following a similar pattern previously found in adults (Jones et al., 2022). Relation type did not significantly influence performance in most models with the exception of Gemini that performed better on items with compositional than categorical relations ($\beta = -0.9, z = -658.5, p < 0.001$), and on items with compositional than causal relations ($\beta = -0.19, z = -144.4, p < 0.001$). RobBERT showed similar significant effect of compositional versus categorical relation items ($\beta = -0.64, z = -2.58, p < 0.001$). GPT-3 , however, performed better on categorical than compositional items ($\beta = 0.45, z = 2.06, p < 0.05$).

Semantic Distance between Base and Target Domains Logistic regression of children's performance revealed a significant effect of the semantic distance between the base and target domains. Items with shorter distance were easier for children to solve ($\beta = -0.54, t = -4.05, p < 0.001$). A similar pattern was observed in most LLMs including Aya ($\beta = -1.79, z = -2.16, p < 0.05$), Command-R ($\beta = -3.07, z = -3.90, p < 0.001$),Gemini ($\beta = -6.23, z = -4536.2, p < 0.001$), GPT-4o ($\beta = -2.84, z = -2.67, p < 0.01$), GPT-3 ($\beta = -2.64, z = -3.31, p < 0.001$), RobBERT ($\beta = --4.15, z = -4.38, p < 0.001$) and XLM-V ($\beta = -5.13, z = -4.67, p < 0.001$). Only Claude showed no significant effect of se-

[3]We found XLM-V to be more suitable than mBERT or XLM-R as it suffers less from overtokenization in Dutch and thus covers more of our test words.

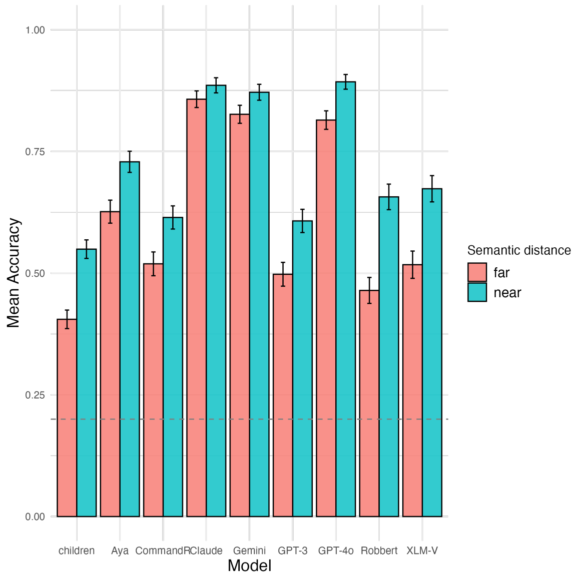


Figure 2: Near analogies are often easier to solve than far analogies for both children and LLMs. Note: for purpose of clarity, in the plot we binned semantic distance into *near* and *far* categories, where *near* is < median semantic distance and *far* is >= median semantic distance.

mantic distance on performance ($p = 0.95$)(see Figure 2).

Distractor Salience As can be seen in Figure 3, items with lower distractor salience were significantly easier to solve than those with high distractor salience for children ($beta = -2.65, z = -4.33, p < 0.001$) and most LLMs including Aya ($beta = -2.76, z = -4.07, p < 0.001$), Command-R ($beta = -2.63, z = -4.31, p < 0.001$), Gemini ($beta = -1.81, z = -1322.5, p < 0.001$), GPT-4o ($beta = -1.99, z = -2.61, p < 0.01$), GPT-3 ($beta = -3.39, z = -4.99, p < 0.001$), RobBERT ($beta = -2.69, z = -3.94, p < 0.001$) and XLM-V ($beta = -2.64, z = -3.63, p < 0.001$). Only Claude showed no significant effect of distractor salience on performance ($p = 0.89$).

7 Results RQ3: Do LLMs choose associative or analogical solutions?

We investigated whether LLMs choose analogical solutions to verbal analogies, after explicitly testing and controlling for associative responses.

7.1 Experiment 1: C:?

In experiment 1, we prompt the LLMs (and adult participants) with only the C-term, e.g., "C is to [MASK]". If these are solved by association as

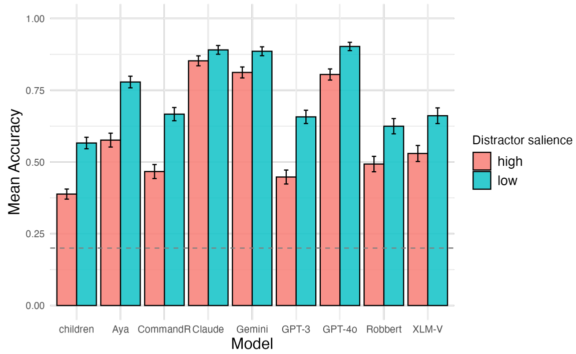


Figure 3: Analogies with low distractor salience are easier to solve than those with high distractor salience for both children and LLMs. Note: for purpose of clarity, in the plot we binned distractor salience into *low* and *high* categories, where *low* is < median distractor salience and *high* is >= median distractor salience.

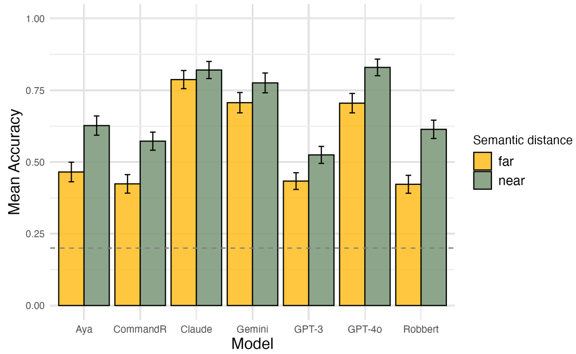


Figure 4: Near analogies are still easier than far analogies, when we control for associative responses (i.e., when filtering out the correctly solved C:? items for each model).

we expect, then LLMs should still be able to solve a substantial portion of analogies purely by association with C (Ushio et al., 2021b; Poliak et al., 2018); hypothesis 3a). This was indeed the case as can be seen in Table 1, where the generative LLMs solve up to 62% of items without being given A:B. Notably, adults also solved 57% of items with the C-only prompt, where there was 49-71% overlap between models and adults in which items could be solved by association.

7.2 Experiment 2: A:B::C:? for selected items

We removed items that each model solved correctly with C:? and reevaluated their performance along the same item factors from RQ2. This was done to test the factors affecting the models' performance on items that were not solved by word association alone. We ran logistic mixed effects models predicting the performance of each LLM by the three

model	Exp 0 A:B::C:? LLMs N items	Acc (SD)	Exp 1 C:? LLMs Acc (SD)	Exp 2 filtered A:B::C:? LLMs N items	Acc (SD)	Children 7-yrs M (SD)	8-yrs M (SD)	9-yrs M (SD)	10-yrs M (SD)	11-yrs M (SD)	12-yrs M (SD)
Aya	855	.67 (.47)	.49 (.50)	435	.54 (.50)	.21 (.37)	.25 (.38)	.30 (.38)	.35 (.38)	.42 (.39)	.49 (.39)
Command-R	855	.57 (.50)	.42 (.49)	494	.50 (.50)	.23 (.39)	.28 (.39)	.34 (.39)	.39 (.39)	.47 (.38)	.53 (.39)
Claude	855	.86 (.34)	.60 (.49)	343	.80 (.40)	.15 (.32)	.19 (.33)	.25 (.34)	.30 (.35)	.37 (.36)	.44 (.36)
Gemini	855	.84 (.36)	.62 (.48)	321	.73 (.45)	.10 (.27)	.14 (.28)	.19 (.29)	.25 (.31)	.33 (.33)	.40 (.35)
GPT-4o	855	.85 (.36)	.58 (.49)	359	.76 (.43)	.15 (.32)	.19 (.33)	.24 (.34)	.30 (.35)	.37 (.36)	.44 (.37)
GPT-3	855	.55 (.36)	.31 (.49)	359	.47(.50)	.24 (.39)	.27 (.40)	.33 (.39)	.38 (.39)	.45 (.39)	.52 (.38)
RobBERT	680	.56 (.50)	.29 (.45)	484	.51 (.50)	.25 (.40)	.30 (.40)	.35 (.40)	.40 (.39)	.48 (.39)	.54 (.38)
XLM-V	622	.59 (.49)	.28 (.45)	447	.51 (.50)	.24 (.39)	.28 (.39)	.34 (.39)	.41 (.39)	.48 (.38)	.55 (.38)

Table 1: LLM Performance on Experiment 0 (original set of A:B::C:? items), Experiments 1 (C:?) and 2 (selection of A:B::C:? after filtering out correct C:? items by each model). Children's mean proportion correct (by age group) on the same selection of items per LLM from Experiment 2.

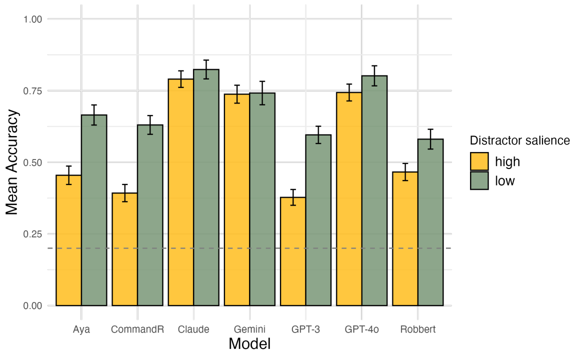


Figure 5: Analogies with low distractor salience are still easier for LLMs, when we control for associative responses.

item characteristics we tested (relation type, semantic distance between base and target domains and distractor salience) on the selected items after filtering out items that were also solved by association only for each model [4]. The models also included by-item intercepts as random effects. Results show that LLMs' performance is still affected by the semantic distance between the base and the target words (see 4) and by distractor salience (see 5). The only exceptions were for GPT-4o and XLM-V , where there were no longer significant effects of distractor salience ($p = 0.43$ and $p = 0.22$, respectively), and for Gemini where there were no significant effects of both semantic distance ($p = 0.57$) and distractor salience ($p = 0.88$) (see Appendix C for the full report of results).

Table 1 shows an overview of model versus children's performance where all items solved correctly with the C:? prompt had been filtered out. We see that BERT-based models solve nearly 30% of analogies correctly when prompted with only "C:?", so without any information about the relation A:B to be mapped. The autoregressive transformer-decoder models solved even greater portions correctly ($40-60\%$) with the C-only prompt. Notably, for the youngest children in our dataset, 7-8-year olds, performance dropped to below chance level on the filtered items sets.

7.3 RQ4: Do LLMs choose the same distractors as children do?

In this exploratory analysis we compared LLM errors to those of children and pilot data from adults. For each of the tested models, we looked at the subset of items it answered incorrectly and compared the distractor it chose to the one chosen by most children and piloted adults. We computed Cohen's Kappa coefficient (Cohen, 1960) to test the agreement of distractor choice between each pair of models and between each model and the children (see Figure 6). As can be expected, the Bert-based models, RobBERT and XLM-V , show similar error patterns, while having low agreement with the autoregressive transformer-decoder models. Notably, neither type of model architecture nor adult pilot data showed similar error patterns to those of children. These results suggest that the high performance of LLMs in this task is not driven by the same process as children. However, top-performing models -Gemini and GPT-4o - had similar error patterns to small sample of adults.

8 Discussion

The main goal of this paper was to investigate whether LLMs rely on association to solve verbal analogies, similar to young children. Direct performance comparisons showed that some LLMs

[4]This means that the set of A:B::C: selected items, as well as their total number, differ across models.

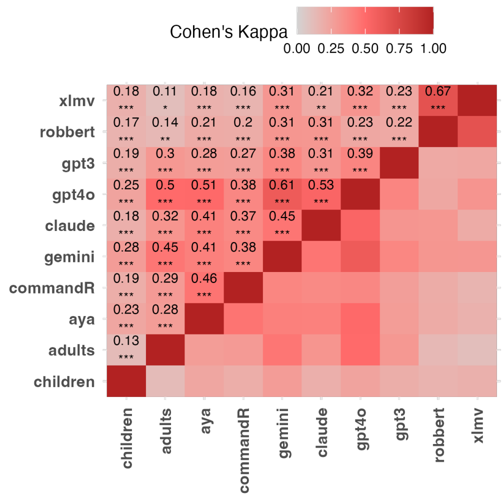


Figure 6: Inter-solver agreement in distractor choice measured using Cohen's Kappa. Values closer to 1 indicate higher agreement. LLMs and children show different error patterns in solving verbal analogies.

perform at the level of 12-year-olds, while top-performing models surpass children and reach adult-level performance. All LLMs appeared to rely heavily on association, though they made different errors than children, and association alone doesn't fully explain their superior performance on this children's analogy task.

To examine whether LLMs are influenced by the same analogy item characteristics as children, we tested the effects of distractor salience, semantic distance between base and target domains, and relation type. Both distractor salience and semantic distance affected LLM performance similarly to children, especially in smaller models. These effects persisted even when association didn't fully account for the reasoning. Relation type, however, did not influence most LLMs in the same way it does children.

A notable finding was that LLMs solved 28%–62% of analogies when prompted with only "C:?", without any information about the relation A:B to be mapped. This experimental manipulation is similar to Ushio et al. (2021b) who found that RoBERTa and BERT only dropped 10 to 15 percentage points in accuracy, still achieving accuracies of 30% or higher on the SAT analogies. In our case, LLMs also dropped around 10 percentage points after filtering out items solved correctly with C:? only. Interestingly, 7-8 year-olds performance often dropped to below chance level on the filtered item sets, which is what was expected as association is the most utilized strategy in this age-group (see Table 1; Thibaut and French (2016); Stevenson and Hickendorff (2018)). A small sample of adults tested on same items also solved 56% of the items when prompted with the C term only.

Our error analysis provides further insight into the similarities in verbal analogical reasoning between children and LLMs. While LLMs exhibit comparable error patterns—particularly among models with the same architecture—their mistakes only loosely align with those made by children. This suggests that there are differences in the way LLMs and children solve verbal analogies. We collected pilot data to examine whether LLM error patterns better coincided with adults to determine whether LLMs resemble more advanced human reasoning or rely on fundamentally different processes. Preliminary results show that the errors of top-performing LLMs, Gemini and GPT-4o - but not Claude , are somewhat similar to those of adults. However, each separate item was solved by only four adults, so future work must determine the reliability of these results.

Our study relies on behavioral methods to evaluate the model's performance on analogy tasks, which, while comparable to the methods used to investigate analogical reasoning in humans, do not allow us to draw definitive conclusions about the underlying mechanisms the model uses to solve these items. In future work, we intend to address this gap by employing mechanistic interpretability techniques—such as visualizing attention patterns—to more directly investigate the process by which LLMs solve analogies.

9 Conclusion

In sum, LLMs perform at or above the level of children on our verbal analogical reasoning task. While word association plays a significant role in LLMs' success, they are able to solve analogies also when this strategy is absent. While LLMs share some similarity to children in the factors that affect performance, the errors they make suggest a different mechanism. Future work can contrast adult-like relational mapping with other possible mechanisms children have been postulated to use such as relational priming (Leech et al., 2008) or partial analogical reasoning (Stevenson and Hickendorff, 2018) to further examine how LLMs solve verbal analogies.

References

P. Alexander and J. Kulikowisch. 1991. Domain knowledge and analogical reasoning ability as predicators of expository text. *Journal of Reading Behavior*, 23(2):165–190.

Joan Bresnan. 2007. Is syntactic knowledge probabilistic? experiments with the english dative alternation. *Roots: Linguistics in search of its evidential base*, 96:77–96.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Alexander Clark. 2014. Distributional learning as a theory of language acquisition. In *Proceedings of the 5th Workshop on Cognitive Aspects of Computational Language Learning (CogACLL)*.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Tamara Czinczoll, Helen Yannakoudakis, Pushkar Mishra, and Ekaterina Shutova. 2022. Scientific and creative analogies in pretrained language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*.

Pieter Delobelle, Thomas Winters, and Bettina Berendt. 2020. Robbert: a dutch roberta-based language model. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3255–3265.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019*, page 4171–4186.

Leonidas A. A. Doumas, Robert G. Morrison, and Lindsey E. Richland. 2018. Individual differences in relational learning and analogical reasoning: A computational model of longitudinal change. *Frontiers in Psychology*, 9:1235.

J. R. Gentile, L. Tedesco-Stratton, E. Davis, N. J. Lund, and B. C. Agunanne. 1977. Associative responding versus analogical reasoning by children. *Intelligence*, 1(4):369–380.

D. Gentner. 1988. Metaphor as structure mapping: The relational shift. *Child Development*, 59(1):47–59.

Dedre Gentner and Kenneth D Forbus. 2011. Computational models of analogy. *Cognitive science*, 2(3):266–276.

Dedre Gentner and Christian Hoyos. 2017. Analogy and abstraction. *Topics in cognitive science*, 9(3):672–693.

Nina Gierasimczuk, Han LJ van der Maas, and Maartje EJ Raijmakers. 2013. An analytic tableaux model for deductive mastermind empirically tested with a massively used online learning system. *Journal of Logic, Language and Information*, 22(3):297–314.

Anna Gladkova, Aleksandr Drozd, and Satoshi Matsuoka. 2016. Analogy-based detection of morphological and semantic relations with word embeddings: what works and what doesn't. In *Proceedings of the NAACL Student Research Workshop*, pages 8–15.

Y. Glady, R. M. French, and J.-P. Thibaut. 2017. Children's failure in analogical reasoning tasks: A problem of focus of attention and information integration? *Frontiers in Psychology*, 8:707.

M. K. Goddu, T. Lombrozo, and A. Gopnik. 2020. Transformations and transfer: Preschool children understand abstract relations and reason analogically in a causal task. *Child Development*, 91(6):1898–1915.

U. Goswami. 1991. Analogical reasoning: What develops? a review of research and theory. *Child Development*, 62(1):1–22.

Usha Goswami and Ann L Brown. 1990. Melting chocolate and melting snowmen: Analogical reasoning and causal relations. *Cognition*, 35(1):69–95.

E. Grave, P. Bojanowski, P. Gupta, A. Joulin, and T. Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Douglas R Hofstadter. 1997. *Fluid concepts and creative analogies: computer models of the fundamental mechanisms of thought*. Allen Lane, The Penguin Press.

Keith J Holyoak. 2012. Analogy and relational reasoning. *The Oxford handbook of thinking and reasoning*, pages 234–259.

Nicholas Ichien, Hongjing Lu, and Keith J Holyoak. 2020. Verbal analogy problem sets: An inventory of testing materials. *Behavior Research Methods*, 52(5):1803–1816.

Laura L Jones, Matt J Kmiecik, John L Irwin, and Robert G Morrison. 2022. Differential effects of semantic distance, distractor salience, and relations in verbal analogy. *Psychonomic Bulletin & Review*.

Sylvia Klinkenberg, Marthe Straatemeier, and Han L van der Maas. 2011. Computer adaptive practice of maths ability using a new item response model for on the fly ability and difficulty estimation. *Computers & Education*, 57(2):1813–1824.

Hubert Kucwaj, Michał Ociepka, and Adam Chuderski. 2022. Various sources of distraction during analogical reasoning. *Memory & Cognition*, 50(7):1614–1628.

Andrey Kutuzov, Murhaf Fares, Stephan Oepen, and Erik Velldal. 2017. Word vectors, reuse, and replicability: Towards a community repository of large-text resources. In *Proceedings of the 21st Nordic Conference on Computational Linguistics*.

Robert Leech, Denis Mareschal, and Richard P Cooper. 2008. Analogy as relational priming: A developmental and computational perspective on the origins of a complex cognitive skill. *Behavioral and Brain Sciences*, 31(4):357–378.

Davis Liang, Hila Gonen, Yuning Mao, Rui Hou, Naman Goyal, Marjan Ghazvininejad, Luke Zettlemoyer, and Madian Khabsa. 2023. XLM-V: Overcoming the Vocabulary Bottleneck in Multilingual Masked Language Models. *arXiv e-prints*, page arXiv:2301.10472.

Bryce Linford, Nicholas Ichien, Keith Holyoak, and Hongjing Lu. 2022. Impact of semantic representations on analogical mapping with transitive relations. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 44.

Hongjing Lu, Nicolas Ichien, and Keith J Holyoak. 2022. Probabilistic analogical mapping with semantic relation networks. *Psychological Review*, 129:1078–1103.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013a. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

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

Melanie Mitchell. 2021. Abstraction and analogy-making in artificial intelligence. *Annals of the New York Academy of Sciences*, 1505(1):79–101.

Sam Musker, Alex Duchnowski, Raphaël Millière, and Ellie Pavlick. 2024. Semantic structure-mapping in llm and human analogical reasoning. *arXiv preprint arXiv:2406.13803*.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Molly Petersen and Lonneke van der Plas. 2023. Can language models learn analogical reasoning? investigating training objectives and comparisons to human performance. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16414–16425, Singapore. Association for Computational Linguistics.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In *Proceedings of the 7th Joint Conference on Lexical and Computational Semantics*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Lindsey E. Richland, Robert G. Morrison, and Keith J. Holyoak. 2006. Children's development of analogical reasoning: Insights from scene analogy problems. *Journal of Experimental Child Psychology*, 94(3):249–273.

Anna Rogers, Olga Kovaleva, Doug Downey, and Anna Rumshisky. 2020. Getting closer to ai complete question answering: A set of prerequisite real tasks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8722–8731.

Jenny R Saffran, Richard N Aslin, and Elissa L Newport. 1996. Statistical learning by 8-month-old infants. *science*, 274(5294):1926–1928.

Robert J Sternberg. 1977. Component processes in analogical reasoning. *Psychological Review*, 84(4):353–378.

Robert J Sternberg and Georgia Nigro. 1980. Developmental patterns in the solution of verbal analogies. *Child Development*, 51:27–38.

Claire E Stevenson. 2017. Role of working memory and strategy-use in feedback effects on children's progression in analogy solving: An explanatory item response theory account. *International Journal of Artificial Intelligence in Education*, 27:393–418.

Claire E Stevenson, Willem J Heiser, and Wilma CM Resing. 2013. Working memory as a moderator of training and transfer of analogical reasoning in children. *Contemporary Educational Psychology*, 38(3):159–169.

Claire E Stevenson and Marian Hickendorff. 2018. Learning to solve figural matrix analogies: The paths children take. *Learning and Individual Differences*, 66:16–28.

Jean-Pierre Thibaut and Robert M French. 2016. Analogical reasoning, control and executive functions: a developmental investigation with eye-tracking. *Cognitive Development*, 38:10–26.

Stéphan Tulkens, Chris Emmery, and Walter Daelemans. 2016. Evaluating unsupervised Dutch word embeddings as a linguistic resource. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4130–4136, Portorož, Slovenia. European Language Resources Association (ELRA).

Peter D Turney, Michael L Littman, Jeffrey Bigham, and Victor Shnayder. 2003. Combining independent modules to solve multiple-choice synonym and analogy problems. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP-03)*.

Akihiro Ushio, Jose Camacho-Collados, and Steven Schockaert. 2021a. Distilling relation embeddings from pre-trained language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9044–9062. Association for Computational Linguistics.

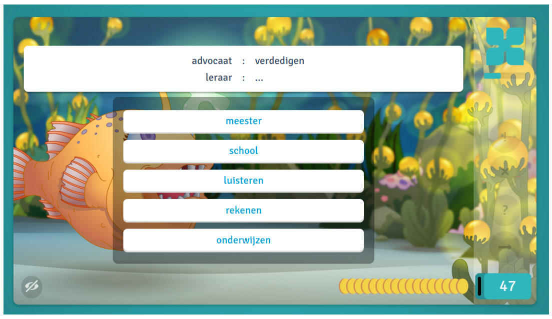
Akihiro Ushio, Luis Espinosa-Anke, Steven Schockaert, and Jose Camacho-Collados. 2021b. Bert is to nlp what alexnet is to cv: Can pre-trained language models identify analogies? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 3609–3624. Association for Computational Linguistics.

Sanne HG van der Ven, Marthe Straatemeier, Brenda RJ Jansen, Sharon Klinkenberg, and Han LJ van der Maas. 2015. Learning multiplication: An integrated analysis of the multiplication ability of primary school children and the difficulty of single digit and multidigit multiplication problems. *Learning and Individual Differences*, 43:48–62.

Taylor Webb, Keith J Holyoak, and Hongjing Lu. 2023. Emergent analogical reasoning in large language models. *Nature Human Behaviour*, 7:1526–1541.

Xunjie Zhu and Gerard de Melo. 2020. Sentence analogies: Linguistic regularities in sentence embeddings. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3389–3400.

Figure 7: Example analogy "lawyer : defending :: teacher : educating"



A Prowise Learn Verbal Analogies Data

Prowise Learn games are adaptive, so that children solve items that are neither too difficult nor too easy, presenting children with items that they have a 65-85% chance of solving correctly, using response time to improve ability estimates (Klinkenberg et al., 2011). Each time a child solves an item his/her ability score on the game is updated according to an algorithm similar to the adaptive ELO rating system used for chess players (for details see Klinkenberg et al., 2011). At the same time the item's difficulty level is adapted according to the same algorithm. In this way item difficulty is on the same scale as the children's ability, and, as such item difficulties can be used to study children's abilities (see van der Ven et al., 2015; Gierasimczuk et al., 2013, for examples in math and logical reasoning). The ELO algorithm is based on the one-parameter logistic function from item response theory where we estimate the probability a child will solve an item correctly given the child's ability score θ and the item's difficulty level β as shown in Equation 1.

$$P(X = 1|\theta, \beta) = \frac{e^{(\theta-\beta)}}{1 + e^{(\theta-\beta)}} \tag{1}$$

Information extracted per item The following information was extracted per item: question text, answer options, item difficulty rating, standard error of item difficulty rating, type of analogy relation, number of times the item was solved, proportion of times each response option was selected.

B Effect of Relation Type on Children's and LLMs' Performance

B.1 Examples for each Relation Type

See Table 2.

Prowise Learn relations	N	relations*	example
action-result	36	causal	parasol : shadow :: sun : warmth
cause-effect	11	causal	falling : broken :: heating : hot
problem-solution	6	causal	noisy : earplugs :: illness : medicine
same category	28	categorical	lion : tiger :: dog : wolf
classification	51	categorical	lego : toys :: sock : clothes
item-characteristic	45	compositional	skyscraper : high :: lead : heavy
object-function	34	compositional	pan : cooking :: pen : writing
part-whole	51	compositional	gate : city :: door : house
share characteristic	25	compositional	giant : mountain :: dwarf : mouse

Table 2: * Mapping of selected relations in verbal analogies game to those examined in Jones et al. (2022).

C Results for item characteristics on items not solved by word association

Results from Aya show effect of semantic distance($beta = -3.94, z = -3.22, p < 0.01$) and distractor salience ($beta = -3.23, z = -3.34, p < 0.001$). Results from linear model of Command-R show effect of semantic distance ($beta = -4.14, z = -3.97, p < 0.001$) and distractor salience ($beta = -3.31, z = -3.99, p < 0.001$). No significant effect of semantic distance and distractor salience was found for Claude ($p = 0.27$ and $p = 0.73$, respectively). No significant effect of semantic distance and distractor salience was found for Gemini ($p = 0.57$ and $p = 0.88$, respectively). A significant effect of semantic distance was found for GPT-4o ($beta = -3.73, z = -2.55, p < 0.05$) but no significant effect of distractor salience ($p = 0.43$). Results from GPT-3 show effect of distractor salience($beta = -2.90, z = -3.97, p < 0.001$) with no significant effect of semantic distance($p = 0.12$). Results from linear model of RobBERT show effect of semantic distance($beta = -4.28, z = -4.08, p < 0.001$) and distractor salience ($beta = -1.86, z = -2.37, p < 0.05$). A significant effect of semantic distance was found for XLM-V ($beta = -5.01, z = -3.96, p < 0.001$) but no significant effect of distractor salience ($p = 0.22$).

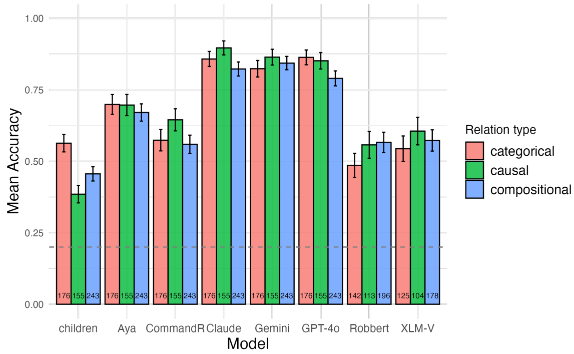


Figure 8: In children (as with adults) compositional relations are easier than causal. Pattern in LLM performance differs per model.