Appendix for: Do large language models solve verbal analogies like children do?

Anonymous ACL submission

1 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 (**?**). 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 **?**). 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 **??**, 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.

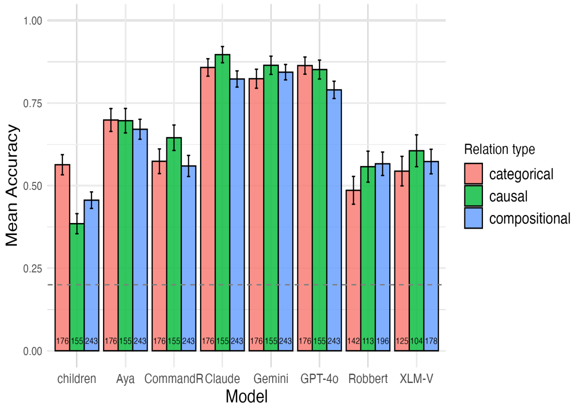

Figure 1: In children (as with adults) categorical relations are easiest, followed by compositional relations and causal are most difficult. LLM performance does not follow this pattern and differs per model.

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

2.1 Examples for each Relation Type

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 1: * Mapping of selected relations in verbal analogies game to those examined in **?**.