Language Models Can Reason about Reference Compositionally, But It's Not Their Native Strength: The Case of the Personal Relation Task

Abstract

Do neural models, such as Large Language Models, genuinely acquire compositional abilities for interpretation of natural language? When we talk about semantic interpretation, we can distinguish two complementary aspects: establishing what an expression refers to in the world (which we call the Extensional task) and representing its sense in a structured way (which we call the Intensional task). We evaluate LLMs and humans on both tasks in the setting of the Personal Relation Task (Paperno, 2022) in which, given a universe of people and their relationships with each other, one is asked to interpret a noun phrase such as *Amber's parent's friend*. Here, for the Intensional task, the answer is the formula `friend(parent(amber))`, and for the Extensional task, the person. We find that humans and LLMs show opposite strengths: humans perform better on Extensional than Intensional tasks, and LLMs *vice versa*. Our methodology brings greater nuance to the understanding of compositional abilities in modern machine learning models. Our results support the notion that the lack of referential grounding in LLM training is a crucial missing component in mimicking human-like language understanding.

1 Introduction

Human languages can have an infinite number of meaningful sentences built from a finite set of elements (Chomsky, 1957; Montague, 1970). This is made possible by the recursive structure of sentences and the principle of *semantic compositionality*, whereby "the meaning of a whole is a function of the meanings of the parts and of the way they are syntactically combined" (Partee, 1995). Compositional structure has been argued to be not just a theoretical construct, but a property of language data that benefits deep neural networks trained on them (Galke et al., 2024).

In semantic theory, the discussion of compositionality is intertwined with the important contrast between two aspects of meaning: sense and reference (Frege, 1892). A noun phrase's *reference* (*Bedeutung*) is the actual entity it denotes in a given world, whereas its *sense* (*Sinn*) is the abstract concept or computational procedure that determines this reference. These are the objects of *Extensional* and *Intensional* semantics, respectively. Two expressions can have the same reference but different senses. If *Amber's friend* and *Bryan's enemy* both refer to the same person, we can describe their senses using two different formulae `friend(amber)` and `enemy(bryan)` but represent their reference identically as e.g. `christina`.

The linguistic fluency demonstrated by Large Language Models (LLMs) is superficially similar to human performance in many aspects (e.g. Yang et al., 2024b). A growing body of research (cf. survey Sinha et al., 2024) addresses a critical question: do AI models genuinely acquire compositional abilities for interpretation, or do their impressive linguistic outputs result primarily from sophisticated pattern matching at the form level?

Much of this work on compositional generalization of LLMs relies on variants of semantic parsing, translating natural language into a formalism; COGS (Kim and Linzen, 2020) is a representative benchmark. The referential aspect of compositional meaning and its relation to conceptual meaning (sense, which semantic parses can be seen as describing), have not received enough attention.

To address this, we adapt the ***Personal Relation Task (PRT)*** (Paperno, 2022), in which one is asked to interpret complex noun phrases such as *Felicia's parent's enemy's child*. The complexity of the phrases requires recursive compositional interpretation; it is highly unlikely that a human or an LLM would have previously learned each of such phrases as a holistic unit. Taking an example from our version of the PRT, given a question like *Who is the friend of the parent of Amber?*, the reference is the specific person the phrase points to

Complexity	Branching	Q: Who is...	A: Extensional		A: Intensional	
			English	Abstract	English	Abstract
3	L	Amber's parent's friend	Felicia	q	friend(parent(amber))	x(k(h))
3	R	the friend of the parent of Amber	Felicia	q	friend(parent(amber))	x(k(h))
4	L	Amber's parent's friend's parent	Chris	y	parent(friend(parent(amber)))	k(x(k(h)))
4	R	the parent of the friend of the parent of Amber	Chris	y	parent(friend(parent(amber)))	k(x(k(h)))

Figure 1: Variables in the Personal Relation Task (PRT): Values of complexity, branching (L and R denote Left- and Right branching, respectively), Approach (Extensional vs Intensional), and Representation type (English vs Abstract). Extensional Answers can be found using the model in Figure 6, Appendix A.

(e.g., Felicia). We call this the ***Extensional Task.*** The sense, in contrast, is the conceptual structure of the query, captured in a formal representation like `friend(parent(amber))`, which we dub the ***Intensional Task***. This distinction is critical, as a model could correctly produce the formal sense without being able to resolve its concrete reference, or vice versa.

We test human participants and five LLMs on both the Extensional task and the Intensional task. In ordinary communication, people routinely identify intended referents without resorting to formal semantic notation, so we expected humans to be better at the Extensional task; this prediction was borne out. Unlike humans, whose referential abilities are rooted in real-world interaction, LLMs operate only over text. As a result, they have limited access to grounding beyond linguistic data but excel at translation; therefore, we expected LLMs to be better at the Intensional task. This too was borne out, and the difference increased as the complexity of the noun phrases increased. §3 gives the full experimental set-up, summarized in Fig. 1, and §4 and §5 discuss the full results.

Our contributions include:

- a methodology for assessing intensional vs. extensional compositional interpretation
- data on human performance on the same tasks in maximally comparable setup to LLMs
- evaluation of several large language models and statistical analysis of model performance, highlighting different strengths of LLMs compared to human participants.

2 Related Work

When discussing how sense vs. reference relate to LLMs, some authors (Merrill et al., 2021; Allen, 2024) theorize about LLMs learning or implementing extensional vs intensional models from text, while other (Bouyamourn, 2023) connect intensionality to paraphrasing, a task which LLMs excel at since the T5 model (Raffel et al., 2020). However, when it comes to empirical evaluation of LLMs, intensional and extensional interpretation have not been contrasted explicitly.

Many of the studies in compositional learning are of immediate relevance here (Sinha et al., 2024). Work on compositional abilities of neural models highlights their limitations, e.g. Lake and Baroni (2018); Yao and Koller (2022), but there are also reports on successes (Lake and Baroni, 2023; Yao and Koller, 2024; Zhou et al., 2023). Existing benchmarks for the evaluation of compositional abilities in AI can be broadly categorized by whether they test for an extensional or an intensional understanding of language. Extensional tasks require a model to map descriptions to referents. SCAN (Lake and Baroni, 2018) can be seen as such a task: mapping a command to a concrete sequence of actions. Another example is the realistic compositional instruction following benchmark by Yang et al. (2024a). Conversely, intensional tasks, which include COGS (Kim and Linzen, 2020), CFQ (Keysers et al., 2020), GeoQUERY (Zelle and Mooney, 1996) and SPIDER (Yu et al., 2018), focus on translating a sentence into a formal, abstract representation, such as a logical formula or a database query, which can be seen as a formalization of sense. While these benchmarks have been crucial

for identifying compositional abilities and weaknesses in neural architectures, they typically test for one form of understanding or the other. A key goal of our work is to use a single, unified framework to probe both intensional and extensional interpretation in a comparable way.

Two properties might have systematically affected the difficulty of compositional tasks for language models. First, most existing tasks are intensional, whereby compositionality largely boils down to piecemeal translation from a natural language to a formal one. Shaw et al. (2021) and a follow-up study by Sun et al. (2023) found that high scores on SCAN do not guarantee good performance on datasets like GeoQUERY and SPIDER, and vice versa, which they interpret in terms of synthetic vs. natural data. However, one could also interpret the contrast in terms of the largely extensional nature of SCAN and intensional nature of other datasets. The relevance of the extensional/intensional distinction, as opposed to natural/synthetic, is further suggested by the fact that in Sun et al.'s experiment, T5 performed well on all intensional datasets, including synthetic COGS, but not on the extensional SCAN.

Second, benchmarks like COGS reuse elements of the language inside meaning representations for it, e.g. word *cat* translates into logical constant `cat`. However, in semantic theory, forms (e.g. strings) and meanings (e.g. entities) are objects of different nature generally without an inherent relationship to each other. In our experiments, we control for the transparency of elementary meaning representations as the separate *representation* variable.

With notable exceptions such as Lake et al. (2019), compositional abilities of models are often assessed without an explicit comparison to humans. Our work aims to fill these gaps. We design an evaluation for intensional vs. extensional compositional representations in comparable conditions; we control for the transparency of elementary meaning representations and provide a comparison with human behavior on the same data.

3 Experimental Setup

To evaluate compositional reasoning, we developed a series of experiments based on a modified and extended version of the Personal Relation Task (PRT), originally designed by Paperno (2022). This task provides a controlled environment to test semantic compositionality in both Large Language Models (LLMs) and human participants.

3.1 Task Design and Variables

The core of the task is a self-contained "universe": a graph of 6 nodes (people) connected by edges representing one of four relations: `friend`, `enemy`, `parent`, or `child`. The `friend` and `enemy` relations are symmetric, while `parent` and `child` are asymmetric inverses. (See Fig. 6 in Appendix A for an example universe.) Questions were generated by recursively traversing this graph, edges used in the example were excluded. To probe different aspects of compositional reasoning, we manipulate four key variables (Fig. 1) as follows:

Approach: Extensional vs. Intensional. The Extensional task is to identify the referent of the given noun phrase, while the Intensional task is to provide a ested functional notation, such as `friend(parent(amber))`.

Representation: English vs. Abstract. The relationship between people and their names is arbitrary, as is the relationship between terms like *child* and the idea of being someone's child. To emulate this property of language, we also designed an ***Abstract*** version of each task, in which the predicates and individuals are mapped to arbitrary variables (e.g., *Amber* = `h`, *friend* = `x`). The ***English*** version uses standard names and relations (e.g., *Amber* = `Amber`, *friend* = `friend`).

This abstract format was introduced for two additional reasons. First, it acts as a control for pretraining data contamination; by using novel symbols, we can be confident that LLMs have not previously encountered this specific task format in textbook examples or other training materials. Second, it increases the difficulty of the intensional task. We hypothesized that the English intensional task might otherwise be solvable through superficial string manipulation rather than true structural understanding. Requiring models to first map abstract symbols to their meanings before constructing a formula forces a more robust form of compositional analysis.

The English Intensional task does not require the universe at all; the solution just involves rearranging the symbols in the natural language query and adding brackets. The Abstract Intensional task on the other hand requires lookup in the universe for referents for predicates and names. While everything an LLM does is, in a sense, string manipulation, the Abstract variant is less straightforward.

Branching: Left vs. Right. This variable refers to the syntactic structure of the query. Left-branching questions used a possessive structure (e.g., *Who is Amber's parent's friend?*). Right-branching questions used prepositional phrases (e.g., *Who is the friend of the parent of Amber?*).

Complexity. The complexity of a question is defined as the number of relations involved plus one. LLMs were tested on complexities 3, 4, 5, and 6, while the human experiment was limited to complexity 3 to ensure a robust and focused dataset.

3.2 Stimuli and Procedure

LLM accuracy was estimated based on a single run. Each stimulus presented to both LLMs and humans followed a standardized four-part structure: (1) a brief task introduction, (2) the complete dataset of relations, (3) a worked out example with a step-by-step solution, and (4) the question. The full dataset, including the generated universes and questions, is publicly available for reproducibility in our GitHub repository[1]. Full examples of each type of task (English/Abstract, Intensional/Extensional) are found in Appendix B. A simplified example with a four-person universe is in Fig. 2. The four main conditions' prompts differ on their universe representation and worked example. Universe items *the enemy of Amber* (right-branching) and *Amber's enemy* (left-branching) have the following representations:

English Extensional	Dana
English Intentional	`enemy(Amber)`
Abstract Extensional	s
Abstract Intensional	`w(h)`

3.2.1 Experiment 1: LLM Evaluation

We evaluated the five LLMs listed in Table 1. We included recently developed Large Reasoning Models (Guo et al., 2025, LRMs), which attempt to achieve more robust reasoning by internalizing a "thought" process, usually achieved via reinforcement learning. Open-weight models were run in under 21 hours on H100, including debugging. For each model, we generated 1,280 prompts ($40 * \text{Approach}(2) * \text{Represenation}(2) * \text{Branching}(2) * \text{Complexity}(4)$) covering every combination of experimental variables. To standardize the output for automated parsing, the instruction *"Please answer the question in the same format as the example. End with 'So the answer is [answer]'."* was appended to every prompt.

[1] `URL: anonymized`

Imagine there are four people: Amber, Bryan, Christina and Dana. They have the following relationships to each other:

the enemy of Amber	= Dana
the friend of Amber	= Christina
the enemy of Bryan	= Christina
the friend of Bryan	= Dana
the enemy of Christina	= Bryan
the friend of Christina	= Amber
the enemy of Dana	= Amber
the friend of Dana	= Bryan

Here is how to figure out the answer to the following question:

Who is the enemy of the friend of Amber?

Here are the steps to arrive at the answer:
1. the friend of Amber = Christina
2. the enemy of Christina = Bryan

So the answer is Bryan

Now answer the following question:
Who is the friend of the enemy of Amber?

Figure 2: Illustrative example of the experimental stimuli. This simplified version uses a four-person universe (instead of six) and only includes the *friend* and *enemy* relations (excluding *parent*/*child*). The example shown here is Complexity 3, Extensional, English, Right-branching condition. Otherwise, the simplified example follows the same standardized four-part structure as full stimuli: a brief introduction, the dataset of relations, a worked example, and an actual test question.

3.2.2 Experiment 2: Human Evaluation

Anonymized data was collected from 40 participants via the crowd-sourcing platform Prolific (2014–2025). The final analysis used a sample of 32 after excluding participants who did not achieve 100% accuracy on four control questions consisting of a single relationship (e.g. *Amber's friend*).

The experiment was implemented in PCIbex (J and F, 2018). Participants completed 8 experimental trials and 4 control trials. Due to budgetary constraints, humans were only tested on complexity 3. For extensional questions, participants selected an answer from an alphabetized dropdown menu; for intensional questions, they typed their response into a text field.

Model Name	Model Alias	Reasoning	Open	Reference
OpenAI GPT 4.1	gpt-4.1	No	No	OpenAI 2025a
OpenAI o3mini	o3-mini	Yes	No	OpenAI 2025b
Qwen2.5-32B	qwen-2.5	No	Yes	Qwen Team 2024
DeepSeek-R1-Distill-Qwen-32B	deepseek-distill	Yes	Yes	DeepSeek-AI 2025
Llama-3.3-70B-Instruct	llama-3.3	No	Yes	Meta 2024

Table 1: Selected Large Language Models (LLMs).

3.3 Analysis

The primary performance **metric** measures exact match with the correct answer, except a wrong number of closing brackets in intensional answers is not penalized. This allows the analysis to focus on semantic understanding over syntactic precision. To statistically evaluate the factors influencing accuracy, we fitted a series of Generalized Linear Mixed Models (GLMMs) with a binomial family using the `lme4` package (Bates et al., 2015) in R (R Core Team, 2021). For more details on the statistical analysis procedure, see Appendix C.

3.4 Hypotheses

To summarize, we had the following hypotheses about the effects of experimental variables for LLMs and human participants, all of which are supported by our experiments:

1. LLMs will be better at Intensional than Extensional tasks.
2. Conversely, humans will be better at Extensional than Intensional.
3. Both humans and LLMs will be better at English than Abstract.
4. Unlike older models, LLMs will not show a difference between Left- and Right-branching.
5. LLMs' performance will decay as complexity goes up.

4 Results

This section presents the comparative performance of Large Language Models (LLMs) and humans on the Personal Relation Task. We analyze the main effects of task approach (Intensional/Extensional), representation (English/Abstract), and complexity (3-6), supported by descriptive statistics and Generalized Linear Mixed Models (GLMMs). Details of the statistical models are provided in Appendix C and a full table of results are in Appendix D. Overall accuracy for humans was 76.95% and for LLMs was 87.4%.

4.1 Effect of Task Approach: Intensional vs. Extensional

Our analysis reveals a significant interaction between the participant group (Human vs. LLM) and the task approach, as illustrated in Figure 4 and 5. Humans performed better on the Extensional task (82.8% accuracy) than the Intensional task (71.1%). The GLMM for human performance showed a significant negative effect for the Intensional approach ($\beta = -1.08, p < 0.01$).

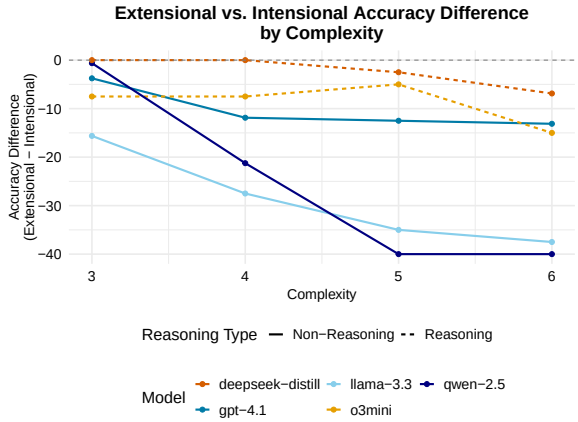

Figure 3: Extensional vs. Intensional performance for different complexities. Data shown is from the LLMs tested. The difference is calculated by following formula: Extensional accuracy − intensional accuracy

Conversely, LLMs achieved higher accuracy on the Intensional task (95.0%) than on the Extensional task (79.8%). This was reflected in the LLM GLMM as a large positive effect for the Intensional approach ($\beta = 1.29, p < 0.001$). This contrast between the human and LLM pattern can be seen in Fig. 5. As seen in Fig. 3, for non-reasoning models, this performance gap widens significantly as complexity increases, whereas reasoning-enhanced models maintain a much flatter and more consistent performance difference across all complexity levels.

A GLMM comparing Humans to LLMs analyz-

ing complexity-3 tasks confirmed this difference. It showed a negative effect of intensionallity for humans $\beta = -2.05, p < 0.01$ while having a strong positive interaction for LLMs with intensionallity ($\beta = 3.17, p < 0.001$).

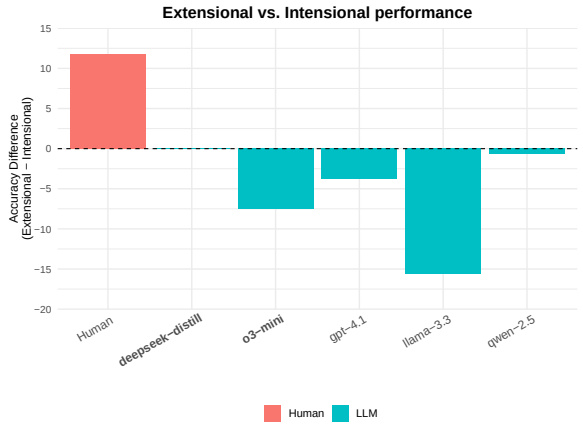

Figure 4: Extensional vs. Intensional performance for Humans and LLMs (complexity 3). Reasoning models are printed in bold. The difference is calculated as Extensional accuracy − Intensional accuracy

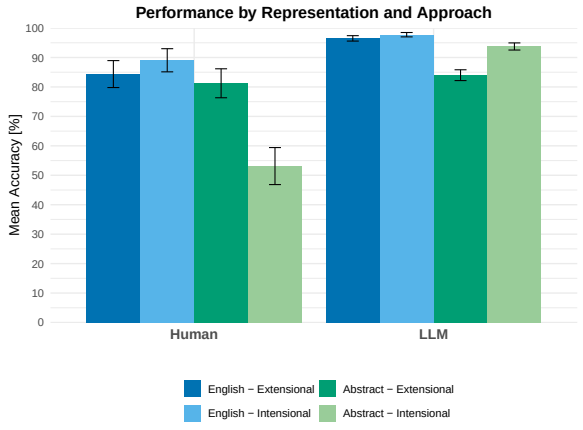

Figure 5: Mean accuracy (%) for Humans and Large Language Models (LLMs) by Representation (English vs. Abstract) and Approach (Extensional vs. Intensional) at complexity 3. The abstract cases (green bars) demonstrate an inverse performance pattern between humans and LLMs: humans found the Extensional approach (dark green) easier than Intensional (light green), while LLMs found Intensional (light green) easier than Extensional (dark green). This divergence becomes clearer at higher complexities.

4.2 Effect of Task Representation: English vs. Abstract

Both LLMs and humans found tasks with abstract representations more difficult than those presented in English. Human accuracy dropped from 85.0% in the English condition to 74.0% in the abstract condition. The human-specific GLMM showed a significant positive effect for the English representation ($\beta = 1.29, p < 0.001$).

Similarly, the LLMs' average accuracy was higher for English tasks (95.1%) than for abstract ones (79.7%), with a significant positive effect for English representation in the LLM GLMM ($\beta = 0.84, p < 0.05$). The combined GLMM did not find a significant interaction between participant group and representation ($p = 0.852$). The lowest accuracy for LLMs (66.1%) was observed in the abstract-extensional condition.

4.3 Effect of Task Complexity on LLMs

For LLMs, task complexity had a significant negative impact on performance. As shown in Table 2, the average accuracy across all models decreased from 93.0% at complexity 3 to 82.4% at complexity 6. The LLM GLMM confirmed this trend with a significant negative main effect for complexity ($\beta = -0.43, p < 0.001$). Individual models showed different degradation curves, with `gpt-4.1` being the most robust. (Humans were tested only on complexity 3.)

Model	*3*	*4*	*5*	*6*	Av.
qwen-2.5	92.2	82.5	79.4	76.9	82.8
deepseek-distill	91.9	**95.0**	90.0	89.1	91.5
gpt-4.1	**98.1**	94.1	**93.1**	**92.8**	**94.5**
llama-3.3	88.4	75.6	74.4	68.8	76.8
o3-mini	94.4	93.8	**93.1**	84.4	91.4
Average	93.0	88.2	86.0	82.4	87.4

Table 2: Model accuracy across different question complexities for LLMs, aggregated across all versions of the PRT.

4.4 Branching Effects and Interaction Analyses

Branching direction did not yield a significant main effect in the human GLMM ($p = 0.326$) or show a consistent pattern across LLMs, though the (non-reasoning) model *qwen-2.5* did struggle slightly more with right-branching ($\beta = -0.94, p < 0.05$), especially in the Extensional task ($\beta = 2.00, p < 0.05$).

A few additional higher-order interactions were statistically significant, but precisely what they mean is unclear; see Appendix C for full GLMM models.

4.5 Focused Analysis on a Reasoning Model Pair.

To isolate the impact of reasoning-specific fine-tuning, we compared the standard *qwen-2.5* model with its reasoning-enhanced *deepseek-distill* variant. The analysis revealed several significant interactions. A strong positive interaction was found between the standard model (*qwen-2.5*) and the Intensional approach ($\beta = -2.27, p < 0.001$), showing the performance gap between Extensional and Intensional tasks was larger for the standard model. Furthermore, significant interactions were identified between model type and both complexity ($\beta = 0.31, p < 0.05$) and branching direction ($\beta = 0.81, p < 0.01$), indicating that the reasoning model compensates for the weaknesses of the standard model when complexity increases (as seen in 3), and in right-branching noun phrases.

5 Discussion

The broad question of interest in this work is the nature of the "reasoning" that LLMs perform when asked to semantically interpret natural language. The particular task is to interpret noun phrases in the context of a universe of people and their relationships to one another.

LLMs are highly proficient at text-based tasks, particularly translation; for example, one of successes of GPT-3 was achieving SOTA BLEU in translation from French and German to English (Brown et al., 2020). The intensional interpretation task strongly resembles translation; for English noun phrases, it can often be reduced to a syntactic rearrangement of elements. For example, converting *Amber's parent's friend* to `friend(parent(Amber))` is a structural transformation. Given their proficiency in similar tasks, we predicted that LLMs would find this intensional task of rearranging a noun phrase into a formula easier than the extensional task of navigating the complex universe to find the correct referent (Hypothesis 1). This hypothesis is related to Bender and Koller's (2020) argument that the lack of grounding in the world outside text limits semantic capabilities of language models. Even though the experimental setup presents a textual representation of a "world" of individuals and their relationships, models trained on text may be less predisposed to use such world information while excelling at translation-like tasks, while humans, whose language use is fundamentally grounded in interactions with objects in the outside world, would exhibit the opposite tendency (Hypothesis 2.

Hypotheses 1 and 2 are well supported by the results. On average, LLMs performed 15.2% worse on the Extensional tasks than the Intensional tasks. Conversely, humans performed 11.7% better on Extensional than Intensional (Fig. 4). This interaction is statistically significant ($\beta = 3.17, p < 0.001$).

While LLMs demonstrated strong performance on the tasks, maxing out near 100% and averaging 87.4% accuracy overall, they show a markedly different pattern from humans. English Intensional is a particularly easy translation task, with an English vocabulary but a different syntax, and here LLMs score highest (95%). In the full LLM experiment, both Intensional ($\beta = 1.29, p < 0.001$) and English ($\beta = 0.84, p < 0.05$) made the task significantly easier for the models. Similarly, Abstract Extensional is the farthest from being a straightforward translation operation – the names must be translated into variables and the universe much be navigated – and this task is by far the hardest for all LLMs, with an average score of 66%, versus an average of 99.4% for the other three conditions in the full LLM experiment.

As expected (Hypothesis 3), both humans and LLMs performed better on English than Abstract (11% and 15% better respectively). That said, LLMs performed quite well on the Abstract tasks (79%). The arbitrariness of the sign is not too much of a hindrance.

For humans, we found the expected results of Extensional (identifying the person, 83%) being easier than Intensional (creating a formula, 71%) ($\beta = -1.08, p < 0.01$). Unexpectedly, our participants did not find the English Intensional task significantly harder than the English Extensional task[2]. Switching from English to Abstract made the Intensional task extremely hard, however, with participants scoring only 53% when they needed to write formulae with variables such as `x(y(z))`. As expected, English was overall much (29%) easier for people than Abstract ($\beta = 1.29, p < 0.001$). This supports the idea that humans excel when they can simply tap into their linguistic abilities.

Reasoning Models: While a broad comparison of reasoning versus non-reasoning models was inconclusive due to model heterogeneity, Figure 3

[2] 89% vs 84% accuracy. The best statistical model did not include the interaction, so this difference could be statistical noise. Additionally, one pilot showed the opposite trend, with English Extensional easier than English Intensional.

demonstrates that the performance gap between difficult extensional and easier intensional tasks remains relatively small and stable for reasoning models as complexity increases, unlike standard models where the gap widens considerably. Furthermore, a focused analysis of our single minimal pair —*qwen-2.5* and its reasoning-distilled variant *deepseek-distill* — offers a meaningful insight. This targeted comparison revealed that the reasoning model significantly narrows the performance gap between the difficult referential (Extensional) task and the easier symbolic (Intensional) task. Though a conclusion based on a single minimal pair of models, this finding supports the idea that the extensional reasoning weakness we identified in LLMs is a specific, addressable challenge rather than an inherent architectural limitation.

Branching (Hypothesis 4): We hypothesised that unlike for earlier results for recurrent neural networks (Paperno, 2022), modern Transformer LLMs would not show a branching bias, as the processing is not strictly sequential. This is indeed what we found, with no main effect of branching and only slight and hard-to-interpret interaction effects for one model (*qwen-2.5*).

Complexity (Hypothesis 5): LLMs show the expected decay in performance as complexity increases ($\beta = -0.43, p < 0.0001$). Task complexity correlates with both recursion depth of the English expression and the number of steps – i.e. opportunities for mistakes – required to complete the task, so this can be an issue of language, task complexity, or both.

6 Conclusion

In this work, we formalized the task of representing compositional meanings inspired by formal semantics and analytic philosophy. Both the intensional and the extensional versions of the task follow this tradition. The abstract version further follows the spirit of the analytical tradition, although in practice formal semantic representations often resemble our English version of the task, motivated by human readability. The Personal Relations Task allowed us to test all these kinds of semantic representations in a way that is accessible both to laypeople (our experimental participants) and to language models, so we could compare the two.

In contrast to previous generations of neural models, the LLMs in our study show a remarkable progress. Even the weakest LLMs we tested give accurate responses most of the time, unlike smaller models tested previously (Bezema, 2019; Monster, 2021; de Wolf, 2023). Compared to recurrent architectures tested in (Paperno, 2022), LLMs do not suffer from systematic structural biases, and don't require extensive task-specific training for reasonable accuracy, as in Paperno's or Yao et al.'s 2025 experiments. And while the extensional version of the task remains relatively difficult for LLMs, adding reasoning seems to reduce the gap.

We deliberately made semantic compositionality accessible to language models by representing meaning through text; sticking to English words makes the task even easier. More realistic interpretation of referring expressions will require going beyond textual inputs and outputs.

In a possible future version of the task, not only could extensions (referential meanings) be presented e.g. visually, but also the outputs could include not just symbolic labels but pointers to a perceptual space, e.g. to image regions.

Another interesting future direction involves mechanistic interpretability. In the current paper, we treated compositional meaning representation as an explicit task. One may wonder whether model internals also include similar compositional mechanisms, as suggested in the literature (Lindsey et al., 2025; Yao et al., 2025), and to what extent such mechanisms are critical for the downstream performance of the model.

In sum, LLMs can be very good at different versions of the compositional Personal Relations Tasks. However, they do not behave like humans, who are better at identifying referents, while LLMs are better at writing formulae representing meanings. This parallels findings in the planning domain, where GPT-4 was found to be better at representations of planning problems than at solving them directly (Liu et al., 2023). Whatever LLMs are doing – arguably, leaning on their translation capabilities – it is not what humans do; LLMs might be achieving similar or better outcomes through a different underlying process. Arbitrary reference still provides a noticeable challenge even for frontier models, supporting the argument (Bender and Koller, 2020; Boleda, 2020; Xu et al., 2025) that referential grounding may be an important missing component to integrate in AI systems.

Limitations

Pursuing the goal of comparing compositional interpretation in human participants vs. LLMs, we created a setup for both that was as comparable as possible, but also somewhat artificial. Results of the human experiment are therefore context specific; Prolific participants performing the task of referential interpretation from detailed textual instructions under time pressure should not be taken as representative of human behavior in all contexts. Due to budget limitations, we were unable to test humans on higher complexities, and overall, the amount of human data is lower than ideal, which we suspect is a contributor to the amount of statistical noise in this experiment.

LLMs, in their turn, were evaluated in favorable conditions, with an explicit textual encoding of the compositional task in a one-shot setting with a fully worked out reasoning example. In more natural settings, LLM performance in tasks involving compositional reasoning may be worse.

We leave a broader comparison of basic instruction-tuned vs. reasoning enhanced models for future research. Here, we report results for only one minimal pair.

As many empirical studies of LLMs, our findings may be limited to the kinds of current models that we examine. Future models may exhibit somewhat different properties.

Ethical Considerations

The human data collection protocol has been approved by `anonymized` Ethics Committee. Participants were provided an information form at `anonymized link`. Participants received £2.50 for completing the questionnaire. Compensation level was estimated on the basis of the hourly rate of £10 and based on experiment duration confirmed in a pilot study.

We used ChatGPT when revising the text of this paper, mainly to paraphrase individual sentences.

References

Bradley Allen. 2024. Carnap's robot redux: LLMs, intensional semantics, and the implementation problem in conceptual engineering.

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.

Emily M Bender and Alexander Koller. 2020. Climbing towards nlu: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 5185–5198.

Daniel Bezema. 2019. Investigating the generalization ability of convolutional neural networks for interpreted languages. BSc AI Thesis.

Gemma Boleda. 2020. Distributional semantics and linguistic theory. *Annual Review of Linguistics*, 6(1):213–234.

Adam Bouyamourn. 2023. Why LLMs hallucinate, and how to get (evidential) closure: Perceptual, intensional, and extensional learning for faithful natural language generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3181–3193.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.

Noam Chomsky. 1957. *Syntactic Structures*. Mouton, The Hague.

Lars de Wolf. 2023. Exploring the generalization capabilities of a generative pre-trained transformer model. BSc AI Thesis.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning math-500 mmlu swe-bench verified deepseek-r1 openai-o1-1217 deepseek-r1-32b openai-o1-mini deepseek-v3.

Gottlob Frege. 1892. Uber Sinn und Bedeutung. *Zeitschrift für Philosophie Und Philosophische Kritik*, 100(1):25–50.

Lukas Galke, Yoav Ram, and Limor Raviv. 2024. Deep neural networks and humans both benefit from compositional language structure. *Nature Communications*, 15(1):10816.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. *arXiv preprint arXiv:2501.12948*.

Florian Hartig. 2024. *DHARMa: Residual Diagnostics for Hierarchical (Multi Level / Mixed) Regression Models*. R package version 0.4.7.

Zehr J and Schwarz F. 2018. Penncontroller for internet based experiments (ibex). Accessed: 2025-02-19.

Daniel Keysers, Nathanael Schärli, Nathan Scales, Hylke Buisman, Daniel Furrer, Sergii Kashubin, Nikola Momchev, Danila Sinopalnikov, Lukasz Stafiniak, Tibor Tihon, Dmitry Tsarkov, Xiao Wang, Marc van Zee, and Olivier Bousquet. 2020. Measuring compositional generalization: A comprehensive method on realistic data. In *International Conference on Learning Representations (ICLR)*.

Najoung Kim and Tal Linzen. 2020. COGS: A compositional generalization challenge based on semantic interpretation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9087–9105. Association for Computational Linguistics.

Brenden Lake and Marco Baroni. 2018. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*.

Brenden M Lake and Marco Baroni. 2023. Human-like systematic generalization through a meta-learning neural network. *Nature*, 623(7985):115–121.

Brenden M. Lake, Tal Linzen, and Marco Baroni. 2019. Human few-shot learning of compositional instructions. *Preprint*, arXiv:1901.04587.

Jack Lindsey, Wes Gurnee, Emmanuel Ameisen, Brian Chen, Adam Pearce, Nicholas L. Turner, Craig Citro, David Abrahams, Shan Carter, Basil Hosmer, Jonathan Marcus, Michael Sklar, Adly Templeton, Trenton Bricken, Callum McDougall, Hoagy Cunningham, Thomas Henighan, Adam Jermyn, Andy Jones, and 8 others. 2025. On the biology of a large language model. *Transformer Circuits Thread*.

Bo Liu, Yuqian Jiang, Xiaohan Zhang, Qiang Liu, Shiqi Zhang, Joydeep Biswas, and Peter Stone. 2023. Llm+ p: Empowering large language models with optimal planning proficiency. *arXiv preprint arXiv:2304.11477*.

William Merrill, Yoav Goldberg, Roy Schwartz, and Noah A Smith. 2021. Provable limitations of acquiring meaning from ungrounded form: What will future language models understand? *Transactions of the Association for Computational Linguistics*, 9:1047–1060.

Meta. 2024. Llama 3.3 70b instruct. Accessed: 2025-02-27.

Joris Monster. 2021. Generalizing relations: using a transformer neural network to generalize compositional semantics. BSc AI Thesis.

Richard Montague. 1970. Universal grammar. *Theoria*, 36(3):373–398.

OpenAI. 2025a. Introducing GPT-4.1 in the API. Accessed 06.10.2025. Version: gpt-4.1-2025-04-14.

OpenAI. 2025b. Openai o3-mini. Accessed: 2025-02-27. Version: o3-mini-2025-01-31.

Denis Paperno. 2022. On learning interpreted languages with recurrent models. *Computational Linguistics*, 48(2):477–485.

Barbara H. Partee. 1995. Lexical semantics and compositionality. In Lila R. Gleitman and Mark Liberman, editors, *An Invitation to Cognitive Science, Vol 1: Language*, pages 311–360. MIT Press, Cambridge, MA.

Prolific. 2014–2025. Prolific. `https://www.prolific.com`. Version: May 2025. Copyright © 2025, Prolific Academic Ltd., London, UK.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Peter Shaw, Ming-Wei Chang, Panupong Pasupat, and Kristina Toutanova. 2021. Compositional generalization and natural language variation: Can a semantic parsing approach handle both? In *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)*, pages 922–938. Association for Computational Linguistics.

Sania Sinha, Tanawan Premsri, and Parisa Kordjamshidi. 2024. A survey on compositional learning of ai models: Theoretical and experimental practices. *Preprint*, arXiv:2406.08787.

Kaiser Sun, Adina Williams, and Dieuwke Hupkes. 2023. [re] a replication study of compositional generalization works on semantic parsing. *ReScience C*, 9(2).

Qihui Xu, Yingying Peng, Samuel A Nastase, Martin Chodorow, Minghua Wu, and Ping Li. 2025. Large language models without grounding recover non-sensorimotor but not sensorimotor features of human concepts. *Nature human behaviour*, pages 1–16.

Haoran Yang, Hongyuan Lu, Wai Lam, and Deng Cai. 2024a. Exploring compositional generalization of large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop)*, pages 16–24.

Qiyuan Yang, Pengda Wang, Luke D. Plonsky, Frederick L. Oswald, and Hanjie Chen. 2024b. From babbling to fluency: Evaluating the evolution of language models in terms of human language acquisition. *arXiv preprint arXiv:2410.13259*.

Yuekun Yao, Yupei Du, Dawei Zhu, Michael Hahn, and Alexander Koller. 2025. Language models can learn implicit multi-hop reasoning, but only if they have lots of training data. *Preprint*, arXiv:2505.17923.

Yuekun Yao and Alexander Koller. 2022. Structural generalization is hard for sequence-to-sequence models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5048–5062, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yuekun Yao and Alexander Koller. 2024. Simple and effective data augmentation for compositional generalization. *Preprint*, arXiv:2401.09815.

Tao Yu, Rui Zhang, Kai Yang, Michihiro Yasunaga, Dongxu Wang, Zifan Li, James Ma, Irene Li, Qingning Yao, Shanelle Roman, and 1 others. 2018. Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-sql task. *arXiv preprint arXiv:1809.08887*.

John M Zelle and Raymond J Mooney. 1996. Learning to parse database queries using inductive logic programming. In *Proceedings of the national conference on artificial intelligence*, pages 1050–1055.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, and Ed Chi. 2023. Least-to-most prompting enables complex reasoning in large language models. In *International Conference on Learning Representations (ICLR)*. ArXiv preprint from 2022.

A Relation graph

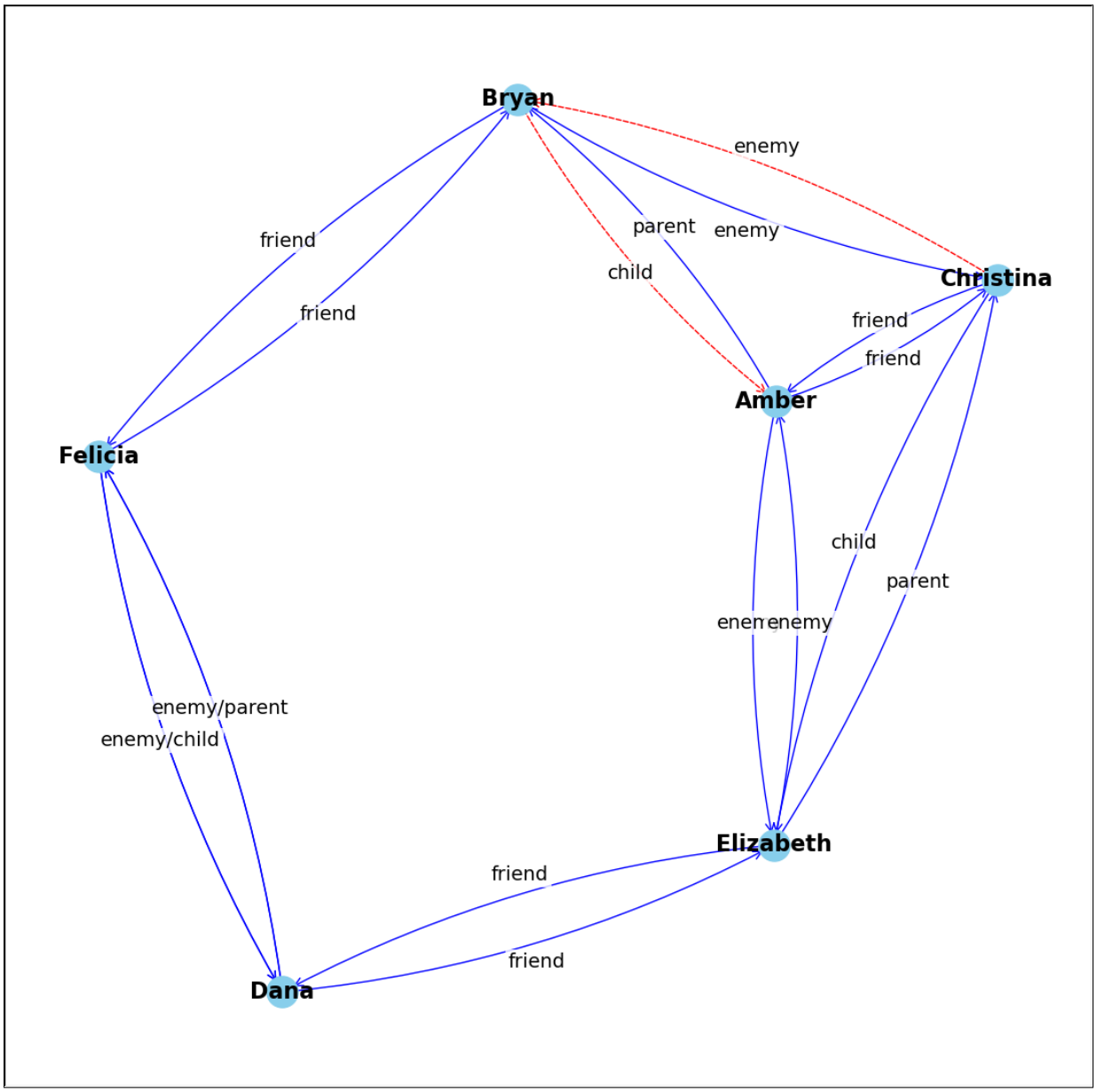

Figure 6: Graph of persons and their relations to each other used to sample paths for in the stimuli. The paths used in the example are given in red. They were excluded from sampling in the questions.

B Example stimuli

Abstract Intensional
English phrases can be translated into formulae. Use the following translations to answer the question.

Amber	= h	the friend of Christina	= $x(y)$
Bryan	= t	the friend of Amber	= $x(h)$
Christina	= y	the enemy of Dana	= $w(s)$
Dana	= s	the enemy of Bryan	= $w(t)$
Elizabeth	= v	the parent of Elizabeth	= $k(v)$
Felicia	= q	the child of Felicia	= $t(q)$

Here is an example of how to get a formula for a phrase.
What is the formula for "the enemy of the parent of Amber"?

1. the parent of Amber = $k(h)$
2. the enemy of the parent of Amber = $w(k(h))$

So the answer is $w(k(h))$.

Now answer the following question: What is the formula for "the friend of the parent of Amber"?

Figure 7: Example stimulus for the **Abstract Intensional** condition.

Abstract Extensional
Imagine there are six people: Amber, Bryan, Christina, Dana, Elizabeth and Felicia. They have the following relationships:

Amber	= h	the enemy of Amber	= v	the friend of Amber	= y
Bryan	= t	the enemy of Bryan	= y	the friend of Bryan	= q
Christina	= y	the enemy of Christina	= t	the friend of Christina	= h
Dana	= s	the enemy of Dana	= q	the friend of Dana	= v
Elizabeth	= v	the enemy of Elizabeth	= h	the friend of Elizabeth	= s
Felicia	= q	the enemy of Felicia	= s		

the parent of Amber	= t	the child of Bryan	= h
the parent of Dana	= q	the child of Christina	= v
the parent of Elizabeth	= y	the child of Felicia	= s

Here is how to answer the following question:
Who is the enemy of the parent of Amber?

1. the parent of Amber = t
2. the enemy of Bryan = y

So the answer is y.

Now answer the following question: Who is the friend of the parent of Amber?

Figure 8: Example stimulus for the **Abstract Extensional** condition.

English Intensional
English phrases can be translated into formulae. Use the following translations:

the friend of Christina	= `friend(christina)`
the friend of Amber	= `friend(amber)`
the enemy of Dana	= `enemy(dana)`
the enemy of Bryan	= `enemy(bryan)`
the parent of Elizabeth	= `parent(elizabeth)`
the child of Felicia	= `child(felicia)`

Example: What is the formula for "the enemy of the parent of Amber"?

1. the parent of Amber = `parent(amber)`
2. the enemy of the parent of Amber = `enemy(parent(amber))`

So the answer is `enemy(parent(amber))`.

Now answer the following question: What is the formula for "the friend of the parent of Amber"?

Figure 9: Example stimulus for the **English Intensional** condition.

English Extensional
Imagine there are six people: Amber, Bryan, Christina, Dana, Elizabeth and Felicia. Their relationships:

the enemy of Amber	= Elizabeth	the friend of Amber	= Christina
the enemy of Bryan	= Christina	the friend of Bryan	= Felicia
the enemy of Christina	= Bryan	the friend of Christina	= Amber
the enemy of Dana	= Felicia	the friend of Dana	= Elizabeth
the enemy of Elizabeth	= Amber	the friend of Elizabeth	= Dana
the enemy of Felicia	= Dana	the friend of Felicia	= Bryan
the parent of Amber	= Bryan	the child of Bryan	= Amber
the parent of Dana	= Felicia	the child of Christina	= Elizabeth
the parent of Elizabeth	= Christina	the child of Felicia	= Dana

Worked question: Who is the enemy of the parent of Amber?

1. the parent of Amber = Bryan
2. the enemy of Bryan = Christina

So the answer is Christina.

Now answer the following question: Who is the friend of the parent of Amber?

Figure 10: Example stimulus for the **English Extensional** condition.

C Statistical Analysis and Model Selection

To evaluate the factors influencing accuracy across our experiments, we employed Generalized Linear Mixed Models (GLMMs) with a binomial family and a logit link function. These models were implemented using the lme4 package (Bates et al., 2015) in R (R Core Team, 2021). The primary performance metric, Correct_Forgiving, was used as the binary outcome, which tolerates minor syntactic deviations (e.g., mismatched closing brackets in intensional answers) to focus on semantic understanding rather than strict output format adherence.

For each experiment (LLM performance, Human Experiment, as well as the combined Human-LLM comparison), a systematic model selection process was followed:

1. Baseline Model: We began by fitting a baseline GLMM including all main effects of the experimental variables relevant to the specific dataset (e.g., Model, Complexity, Branching, Approach, Representation for LLM analysis; Experiment, Branching, Approach, Representation for the combined analysis). Random intercepts were included for ParticipantId (in human experiments and combined models) and PathId (in LLM experiments, and explored for humans).

2. Interaction Terms: We incrementally added interaction terms, starting with two-way interactions, then three-way, and finally four-way interactions where appropriate. The inclusion of interaction terms was guided by initial descriptive analyses and theoretical expectations about how variables might jointly influence performance.

3. Model Comparison: Likelihood Ratio Tests (LRTs) were used to compare nested models. A more complex model was selected over a simpler one if it provided a statistically significant improvement in fit (typically with $p < 0.05$). Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) were also considered to balance model fit with parsimony.

4. Diagnostic Checks: After selecting the final model for each analysis, DHARMa diagnostic plots (Hartig, 2024) were generated to assess the model's distributional assumptions (residuals vs. predicted values, dispersion, and inflation). These checks ensured that the chosen GLMM structure adequately described the data and its error characteristics, as detailed in Appendix B.

For the specific GLMM on data from just LLMs (Tables 3 and 4), the reference model for the Model factor was o3-mini due to its consistent high performance. The gpt-4.1 model was excluded from this GLMM analysis due to its near-ceiling performance, which resulted in limited data variability and convergence issues for the statistical model.

For the combined GLMM comparing Human and LLM performance (Table 5), the data for LLMs was subsetted to only include complexity level 3, matching the human experiment data to ensure a fair comparison. The reference levels for this model were Experiment (Human), Branching (Left), Approach (Extensional), and Representation (Abstract).

This rigorous approach to model selection ensures that the reported fixed effects and interaction terms provide the most statistically robust explanation for the observed patterns in compositional reasoning performance.

Factor	Estimate	Std. Error	z value	Pr(> |z|)
(Intercept)	3.53463	0.53312	6.630	3.36e-11 ***
Model (deepseek-distill)	-1.41619	0.70175	-2.018	0.043582 *
Model (llama-3.3)	-1.50096	0.63392	-2.368	0.017898 *
Model (qwen-2.5)	-0.78518	0.68257	-1.150	0.250011
Complexity	-0.43322	0.10221	-4.239	2.25e-05 ***
Branching (Right)	0.18347	0.30149	0.609	0.542824
Approach (Intensional)	1.29185	0.38854	3.325	0.000885 ***
Representation (English)	0.83973	0.34390	2.442	0.014614 *
Two-Way Interactions:				
Branching (Right) : Approach (Intensional)	0.05752	0.57403	0.100	0.920188
Branching (Right) : Representation (English)	0.06740	0.50658	0.133	0.894155
Approach (Intensional) : Representation (English)	0.13048	0.69102	0.189	0.850235
Model (deepseek-distill) : Complexity	0.24162	0.13625	1.773	0.076175 .
Model (llama-3.3) : Complexity	-0.05561	0.12280	-0.453	0.650674
Model (qwen-2.5) : Complexity	-0.15224	0.13413	-1.135	0.256380
Model (deepseek-distill) : Branching (Right)	0.77078	0.44013	1.751	0.079901 .
Model (llama-3.3) : Branching (Right)	-0.64162	0.38530	-1.665	0.095867 .
Model (qwen-2.5) : Branching (Right)	-0.93841	0.38769	-2.421	0.015499 *
Model (deepseek-distill) : Approach (Intensional)	-0.63755	0.49045	-1.300	0.193627
Model (llama-3.3) : Approach (Intensional)	1.32294	0.50575	2.616	0.008903 **
Model (qwen-2.5) : Approach (Intensional)	1.93563	0.57763	3.351	0.000805 ***
Model (deepseek-distill) : Representation (English)	1.66534	0.63930	2.605	0.009190 **
Model (llama-3.3) : Representation (English)	0.80306	0.43297	1.855	0.063628 .
Model (qwen-2.5) : Representation (English)	2.55215	0.56983	4.479	7.51e-06 ***

*Note: Significance codes: *** p < 0.001, ** p < 0.01, * p < 0.05, . p < 0.1*

Table 3: Fixed effects from GLMM on data from just LLMs. All main effects are included and up to three-way interactions are included between Model, Branching, Approach, and Representation. Formula: `Correct_forgiving Model * (Complexity + (Branching * Approach * Representation))`. **Part 1**: Main Effects and Two-Way Interactions.

Factor	Estimate	Std. Error	z value	Pr(> |z|)
Three-Way Interactions:				
Branching (Right) : Approach (Intensional) : Representation (English)	-0.54003	0.97298	-0.555	0.578878
Model (deepseek-distill) : Branching (Right) : Approach (Intensional)	-0.89940	0.73700	-1.220	0.222334
Model (llama-3.3) : Branching (Right) : Approach (Intensional)	-0.21014	0.71516	-0.294	0.768884
Model (qwen-2.5) : Branching (Right) : Approach (Intensional)	2.00128	1.02000	1.962	0.049758 *
Model (deepseek-distill) : Branching (Right) : Representation (English)	-1.02109	0.93086	-1.097	0.272674
Model (llama-3.3) : Branching (Right) : Representation (English)	1.02568	0.64342	1.594	0.110916
Model (qwen-2.5) : Branching (Right) : Representation (English)	0.37938	0.78954	0.481	0.630865
Model (deepseek-distill) : Approach (Intensional) : Representation (English)	0.62596	1.34386	0.466	0.641367
Model (llama-3.3) : Approach (Intensional) : Representation (English)	-1.07692	0.86801	-1.241	0.214727
Model (qwen-2.5) : Approach (Intensional) : Representation (English)	-3.79422	0.97642	-3.886	0.000102 ***
Four-Way Interactions:				
Model (deepseek-distill) : Branching (Right) : Approach (Intensional) : Representation (English)	-0.02908	1.70129	-0.017	0.986363
Model (llama-3.3) : Branching (Right) : Approach (Intensional) : Representation (English)	0.05765	1.22076	0.047	0.962332
Model (qwen-2.5) : Branching (Right) : Approach (Intensional) : Representation (English)	-1.43200	1.47396	-0.972	0.331284

*Note: Significance codes: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, . $p < 0.1$*

Table 4: Fixed effects from GLMM on data from just LLMs (continued). **Part 2**: Three-Way and Four-Way Interactions.

Factor	Estimate	Std. Error	z value	Pr(> |z|)
(Intercept)	1.92381	0.64130	3.000	0.002701 **
ExperimentLLM	-0.10828	1.32646	-0.082	0.934937
Branching (Right)	0.53510	0.78788	0.679	0.497033
Approach (Intensional)	-2.04905	0.72574	-2.823	0.004752 **
Representation (English)	0.57004	0.80507	0.708	0.478904
Two-Way Interactions:				
ExperimentLLM : Branching (Right)	-1.33605	0.88753	-1.505	0.132234
ExperimentLLM : Approach (Intensional)	3.17485	0.92316	3.439	0.000584 ***
Branching (Right) : Approach (Intensional)	-0.08093	1.01198	-0.080	0.936261
ExperimentLLM : Representation (English)	0.17760	0.95440	0.186	0.852378
Branching (Right) : Representation (English)	-0.51145	1.15625	-0.442	0.658247
Approach (Intensional) : Representation (English)	2.33362	1.14334	2.041	0.041245 *
Three-Way Interactions:				
ExperimentLLM : Branching (Right) : Approach (Intensional)	1.12568	1.29613	0.868	0.385124
ExperimentLLM : Branching (Right) : Representation (English)	3.38204	1.63937	2.063	0.039112 *
ExperimentLLM : Approach (Intensional) : Representation (English)	-1.38978	1.67654	-0.829	0.407127
Branching (Right) : Approach (Intensional) : Representation (English)	0.56641	1.62993	0.348	0.728213
Four-Way Interactions:				
ExperimentLLM : Branching (Right) : Approach (Intensional) : Representation (English)	-5.90502	2.38078	-2.480	0.013127 *

*Note: Significance codes: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$*

Table 5: Fixed effects from the Generalized Linear Mixed Model (GLMM) comparing Human and LLM performance on the Personal Relation Task. The model includes all main effects and interaction terms for Experiment (Human vs. LLM), Branching (Left vs. Right), Approach (Extensional vs. Intensional), and representation (English vs. Abstract), with random intercepts for RelationId and ParticipantId. The reference levels are Experiment (Human), Branching (Left), Approach (Extensional), and Representation (Abstract).

Factor	Estimate	Std. Error	z value	Pr($>\|z\|$)
(Intercept)	1.2209	0.4711	2.591	0.009559 **
Branching (Right)	0.3250	0.3307	0.983	0.325751
Approach (Intensional)	-1.0759	0.3451	-3.118	0.001821 **
Representation (English)	1.2871	0.3519	3.658	0.000254 ***

*Note: Significance codes: *** p < 0.001, ** p < 0.01, * p < 0.05, . p < 0.1*

Table 6: GLMM Fixed Effects for Human Performance. The reference levels are Branching (Left), Approach (Extensional), and representation (Abstract). Significance codes: *** p < 0.001, ** p < 0.01.

Factor	Estimate	Std. Error	z value	Pr($>\|z\|$)
(Intercept)	4.83591	0.48292	10.014	$<$ 2e-16 ***
Model (deepseek-distill)	-0.38780	0.71260	-0.544	0.58630
Complexity	-0.49894	0.08661	-5.761	8.38e-09 ***
Branching (Right)	-0.39780	0.17948	-2.216	0.02666 *
Approach (Intensional)	2.59844	0.22957	11.319	$<$ 2e-16 ***
Representation (Abstract)	-2.40062	0.21880	-10.972	$<$ 2e-16 ***
Two-Way Interactions::				
Model (deepseek-distill) : Complexity	0.31116	0.12597	2.470	0.01351 *
Model (deepseek-distill) : Branching (Right)	0.81159	0.27671	2.933	0.00336 **
Model (deepseek-distill) : Approach (Intensional)	-2.27228	0.31043	-7.320	2.48e-13 ***
Model (deepseek-distill) : Representation (Abstract)	0.23241	0.37212	0.625	0.53225

*Note: Significance codes: *** p < 0.001, ** p < 0.01, * p < 0.05, . p < 0.1*

Table 7: Fixed effects from GLMM comparing the minimal pair of reasoning models (i.e. *qwen-2.5* and *deepseek-distill*). *Deepseek-distill* is the reference level. All main effects are included. Formula: $Correct\_forgiving \sim Model * (Complexity + Branching + Approach + Representation)$.

D Accuracies table

Complexity	Model	Extensional Abstract L	Ext. Abstract R	Ext. English L	Ext. English R	Intensional Abstract L	Int. Abstract R	Int. English L	Int. English R
3	*Human*	78.1%	84.4%	84.4%	84.4%	50.0%	56.3%	87.5%	90.6%
3	**deepseek-distill**	80.0%	90.0%	100.0%	97.5%	77.5%	90.0%	100.0%	100.0%
3	**o3-mini**	95.0%	80.0%	92.5%	95.0%	97.5%	95.0%	100.0%	100.0%
3	llama-3.3	70.0%	65.0%	87.5%	100.0%	95.0%	92.5%	97.5%	100.0%
3	gpt-4.1	87.5%	97.5%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
3	qwen-2.5	97.5%	77.5%	95.0%	97.5%	92.5%	97.5%	100.0%	80.0%
4	**deepseek-distill**	87.5%	92.5%	100.0%	100.0%	87.5%	92.5%	100.0%	100.0%
4	**o3-mini**	77.5%	90.0%	95.0%	97.5%	92.5%	100.0%	100.0%	97.5%
4	llama-3.3	47.5%	30.0%	70.0%	100.0%	92.5%	80.0%	87.5%	97.5%
4	gpt-4.1	75.0%	77.5%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
4	qwen-2.5	55.0%	35.0%	97.5%	100.0%	95.0%	100.0%	77.5%	100.0%
5	**deepseek-distill**	77.5%	82.5%	97.5%	97.5%	90.0%	82.5%	97.5%	95.0%
5	**o3-mini**	87.5%	90.0%	95.0%	90.0%	100.0%	87.5%	97.5%	97.5%
5	llama-3.3	30.0%	22.5%	90.0%	85.0%	90.0%	85.0%	100.0%	92.5%
5	gpt-4.1	70.0%	80.0%	100.0%	97.5%	100.0%	100.0%	100.0%	97.5%
5	qwen-2.5	25.0%	15.0%	100.0%	97.5%	100.0%	100.0%	100.0%	97.5%
6	**deepseek-distill**	62.5%	92.5%	92.5%	95.0%	90.0%	85.0%	100.0%	95.0%
6	**o3-mini**	65.0%	75.0%	80.0%	87.5%	85.0%	97.5%	92.5%	92.5%
6	llama-3.3	37.5%	27.5%	70.0%	65.0%	85.0%	80.0%	95.0%	90.0%
6	gpt-4.1	75.0%	72.5%	100.0%	97.5%	100.0%	100.0%	97.5%	100.0%
6	qwen-2.5	32.5%	17.5%	92.5%	85.0%	95.0%	97.5%	100.0%	95.0%

Table 8: Accuracies for humans and LLMs across all variables. Reasoning models are bolded, and the human data is italicized. Human data was only collected for complexity 3, while LLM data spans complexities 3–6.