The Curious Case of Analogies: Investigating Analogical Reasoning in Large Language Models

Taewhoo Lee[1,2], **Minju Song**[1], **Chanwoong Yoon**[1], **Jungwoo Park**[1,2], **Jaewoo Kang**[1,2*]
[1]Korea University
[2]AIGEN Sciences
{taewhoo, minjusong, cwyoon99, jungwoo-park, kangj}@korea.ac.kr

Abstract

Analogical reasoning is at the core of human cognition, serving as an important foundation for a variety of intellectual activities. While prior work has shown that LLMs can represent task patterns and surface-level concepts, it remains unclear whether these models can encode high-level relational concepts and apply them to novel situations through structured comparisons. In this work, we explore this fundamental aspect using proportional and story analogies, and identify three key findings. First, LLMs effectively encode the underlying relationships between analogous entities; both attributive and relational information propagate through mid-upper layers in correct cases, whereas reasoning failures reflect missing relational information within these layers. Second, unlike humans, LLMs often struggle not only when relational information is missing, but also when attempting to apply it to new entities. In such cases, strategically patching hidden representations at critical token positions can facilitate information transfer to a certain extent. Lastly, successful analogical reasoning in LLMs is marked by strong structural alignment between analogous situations, whereas failures often reflect degraded or misplaced alignment. Overall, our findings reveal that LLMs exhibit emerging but limited capabilities in encoding and applying high-level relational concepts, highlighting both parallels and gaps with human cognition.

Code — https://github.com/dmis-lab/analogical-reasoning

1 Introduction

Analogical reasoning is a fundamental aspect of human cognition, enabling humans to navigate unfamiliar situations by drawing parallels to familiar concepts (Hofstadter 2001; Holyoak, Gentner, and Kokinov 2001; Hofstadter and Sander 2013). This ability serves as the foundation for a wide range of cognitive functions, including knowledge adaptation (Keane 1996), problem solving, and creative thinking (Gentner and Markman 1996). Among various types of analogies, proportional analogies are widely used to assess one's ability to extract semantic relationships and apply them to new contexts (Brown 1989). For example, given the query "*Persuasion is to Jane Austen as 1984 is to*", one would first focus on the first pair of entities (*Persuasion*, *Jane Austen*) to identify the semantic relationship ("*author of*"), and apply it to the third entity (*1984*) to obtain the correct answer ("*George Orwell*"). Extending this rudimentary setting, the ability to draw parallels between situations can be evaluated using story analogies. For instance, despite different surface details between "*missing* a train but *encountering* a dear friend" and "*getting injured* but *coming back stronger*", we find corresponding elements binded under the same theme: that every cloud has a silver lining.

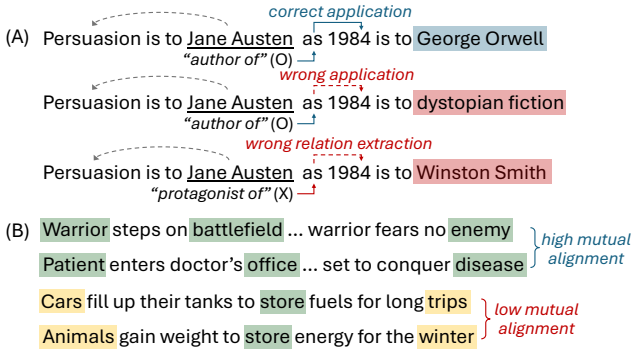


Figure 1: An overview of the mechanism behind analogical reasoning in LLMs. (A) LLMs effectively encode relational information and apply it during correct analogical reasoning, but applying the relation often remains as much a bottleneck as encoding it. (B) Identifying analogous situations is strongly associated with structural alignment, which we quantify using the Mutual Alignment Score (MAS).

Meanwhile, the advent of large language models (LLMs) and their remarkable performance on various tasks have spurred interest in the research community. Trained on massive text corpora with billions of parameters, modern LLMs have shifted the paradigm of problem-solving from task-specific fine-tuning to leveraging instructions and examples in the input prompt (Brown et al. 2020). This emergent ability has motivated researchers to explore LLMs for complex reasoning tasks in diverse domains (Kojima et al. 2022; Yao et al. 2023; Imani, Du, and Shrivastava 2023). Recently, there has been growing interest in the analogical reasoning capabilities of LLMs, focusing on evaluating (Webb, Holyoak, and Lu 2023) or advancing (Wijesiriwardene et al.

*Corresponding author.

2024) these capabilities. However, the inner mechanisms behind LLMs and their ability to perform analogical reasoning remains unexplored. How do models extract relationships and apply them to predict the correct answer? Moreover, how do they draw parallels between semantically disparate, yet analogous context?

In this work, we take a closer look at how modern LLMs perform analogical reasoning. We first examine how information that bridges different entities is extracted and applied using proportional analogies. To understand this information flow, we analyze where critical signals are processed within the input. By blocking the final token from attending to different token positions, we find that mid-upper layers within the second and third entities (e.g., "Jane Austen" and "1984" in Figure 1) carry essential information; disrupting these positions leads to noticeable drops in performance. Further analysis shows that these positions encode both attributive and relational information, with relational content showing a significant gap between correct and incorrect cases. This suggests that, much like humans, models can not only represent individual entities but also abstract the underlying relation that connects them, highlighting relational reasoning as a central mechanism in analogical understanding.

Extending this analysis, we find that applying relational information poses an additional challenge beyond merely identifying it. Replacing the initial entity pairs in incorrect cases with those from correct ones changes model behavior in up to 38.4% of cases, suggesting that models still struggle to transfer relational structure in the remaining cases. Building on earlier insights about the role of linking positions (e.g., "as"), we conduct patching interventions to facilitate information flow between entity pairs. These adjustments lead to successful answer revisions in up to 38.1% of the remaining cases, highlighting that failures in analogical reasoning stem not only from representational gaps but also from limitations in relational application.

To deepen our understanding of how models perform analogical reasoning, we turn to the question of structural alignment, i.e., how models identify and map high-level relational parallels between seemingly unrelated concepts. Using story analogies, we reveal that analogical structure becomes increasingly linearly separable in the middle layers, and that successful reasoning is associated with stronger token-level alignment between source and target stories, despite minimal lexical overlap. These findings suggest that, beyond encoding entity-level information, LLMs develop abstract relational representations and perform alignment operations that mirror core aspects of human analogical reasoning.

In summary, our contributions include:

- We investigate the internal mechanisms of LLMs in analogical reasoning, focusing on how models succeed (or fail) to extract and apply relational information.
- We analyze how structural alignment emerges in model representations, associating it with deeper token-level alignment between analogous situations.
- We contextualize model behavior by comparing it with human cognition, highlighting both parallels in relational abstraction and limitations in alignment and application.

2 Preliminaries

In this section, we provide an overview of prior research on analogical reasoning (see Section 2.1). We then discuss studies in mechanistic interpretability, focusing on methods used in our research (see Section 2.2). Lastly, we clarify key terminologies used throughout the paper (see Section 2.3).

2.1 Analogical Reasoning

Analogical reasoning is a cognitive process that requires identifying relational similarities to understand new situations, form abstract concepts, and draw on past experiences to tackle novel problems (Boteanu and Chernova 2015). Analogies can take several forms, including word analogies (Gladkova, Drozd, and Matsuoka 2016), proportional analogies (Mikolov, Yih, and Zweig 2013), story analogies (Jiayang et al. 2023), and long-text analogies (Sultan and Shahaf 2022). In this work, we focus on two types that best represent the cognitive requirements of analogical reasoning: proportional analogies, which require extracting and applying semantic relationships in the form "*A is to B as C is to D*"; and story analogies, which demand structural alignment between semantically distinct narratives or situations.

In the field of natural language processing (NLP), analogical reasoning has been explored through both benchmark construction (Ye et al. 2024; Jiayang et al. 2023) and behavioral evaluation (Webb, Holyoak, and Lu 2023). Others propose prompting strategies to leverage analogical capabilities more effectively, such as self-generated exemplars (Yasunaga et al. 2024) or knowledge-enhanced prompts (Wijesiriwardene et al. 2024). In a related line of work, several studies have examined how LLMs encode abstract task-level information when presented with in-context examples (Hendel, Geva, and Globerson 2023; Todd et al. 2024; Opiełka, Rosenbusch, and Stevenson 2025). These works identify task or function vectors, i.e., compact representations that reflect the operation demonstrated in ICL settings. While these studies provide evidence that models can internally represent conceptual relations, they are primarily limited to simple tasks (e.g., color matching, antonyms) and focus on detecting the presence of these vectors rather than analyzing what they represent and how they are used in more complex reasoning scenarios. In contrast, our work directly targets analogical reasoning behavior, offering a comprehensive view on how models extract, apply, and structurally align relational information.

2.2 Mechanistic Interpretability

Understanding the internal mechanisms of LLMs has been a central focus of recent research (Bereska and Gavves 2024). Among the various techniques developed to analyze intermediate activations and their causal roles in model behavior, two broad categories are particularly relevant to our work: representational analysis (nostalgebraist 2020; Belrose et al. 2023; Pal et al. 2023), which investigates what types of information are encoded in hidden states; and intervention-based methods (Vig et al. 2020; Zhang and Nanda 2024; Pochinkov et al. 2024), which manipulate internal activations to examine their functional impact on model outputs.

Our study builds on both paradigms to probe the internal computations that support analogical reasoning. Below, we introduce key methods employed in our experiments:

(1) Attention Knockout (Wang et al. 2023b; Geva et al. 2023): This method involves selectively disabling attention heads to examine their contribution to predicting outputs. By removing specific attention pathways, we can assess whether specific tokens are responsible for prediction correct outputs and identify which components are crucial for resolving relational information.

(2) Linear Probing (Alain and Bengio 2018; Belinkov 2022): This technique assesses whether specific types of information are linearly separable within a model's hidden representations. Given labeled examples, we extract activation vectors from a particular layer and train a linear classifier to predict the labels. High probe accuracy suggests that the relevant information is explicitly encoded in the representation space at that layer.

(3) Patchscopes (Ghandeharioun et al. 2024): A recent extension of activation patching that leverages the model's generative capabilities to interpret what information is encoded in its hidden representations. Specifically, a source prompt is first passed through the model, and the hidden representation of the token we wish to inspect is recorded. Next, the same model processes a *target prompt*, which is used to induce natural language descriptions regarding the representation. For example, when using the target prompt constructed by Ghandeharioun et al. (2024): "`Syria: Country in the Middle East, Leonardo DiCaprio: American actor, Samsung: South Korean multinational major appliance and consumer electronics corporation, x`", the representation in "`x`" is replaced with the previously recorded representation, resulting in the description of that representation. Throughout this work, we systematically construct diverse target prompts suitable for extracting different types of information.

2.3 Terminology

In our experiments, proportional analogies follow the structure of "e_1 is to e_2 as e_3 is to e_4". We refer to "as" as the *link*, and the underlying connection that groups entities together (e.g., "author of" in "*Persuasion is to Jane Austen*") as the *relation*. Story analogies include a source story, a target story (analogous), and a distractor story (lexically similar). For both settings, we refer to the final position of the input as the *resolution token*.

3 Experimental Setup

3.1 Dataset Construction

For proportional analogies, we manually construct a test set that contains both correct and incorrect analogies for each model. We begin by retrieving entity pairs from AnalogyKB (Yuan et al. 2024), a million-scale analogy knowledge base that contains entity pairs of the same relation [1]. Next, to ensure a clear distinction between correct and incorrect cases for evaluation, we manually filter out relations that can lead to multiple answers (e.g., "*interested in*") or change over time (e.g., "*head of state*"). Finally, we iteratively combine different entity pairs (e_1-e_2, e_3-e_4) that share the same relation, generating a total of 50k analogies to be used for evaluation.

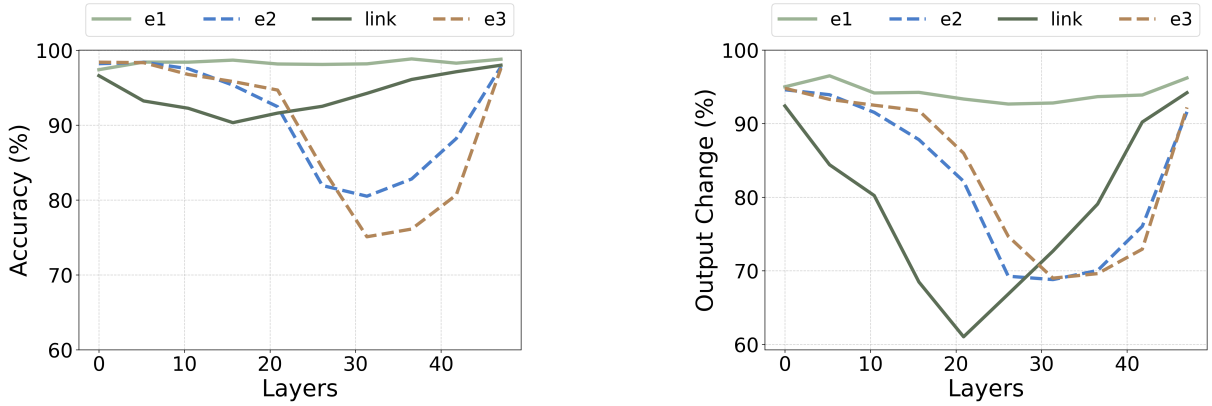
In the evaluation phase, we set up a series of additional filters to confine our experiments to analogical reasoning. First, we ensure that each model is equipped with the necessary knowledge. Formally, for each (e_i, e_j) pair, we check whether models can predict e_j given e_i and the relation. As an illustrative example, for the analogy "*Persuasion is to Jane Austen as 1984 is to George Orwell*", we construct two queries with the relation as follows: "*The author of Persuasion is*" and "*The author of 1984 is*". If a model fails to answer both queries correctly, we exclude the analogy, as we cannot determine whether the incorrect predictions stem from incorrect analogical reasoning or from a lack of prior knowledge. Second, we prevent models from relying on reasoning shortcuts (Xu et al. 2022; Wang et al. 2023a). We define reasoning shortcuts as instances where models return the correct answer without e_2 or "e_1 is to e_2". For example, we construct two queries as follows: "*Persuasion is to as 1984 is to*" and "*1984 is to*". If the model correctly predicts "*George Orwell*" in these cases, this suggests that the answer entity is strongly correlated with e_3, bypassing the need to perform analogical reasoning. In such cases, we exclude the analogy to ensure that models are genuinely engaging in relational reasoning rather than leveraging direct associations. We sample 500 analogies each from the remaining collection of correct and incorrect cases for our experiments.

For story analogies, we use the StoryAnalogy (Jiayang et al. 2023) dataset, which contains 360 multiple-choice questions. Each question involves selecting the target story that is analogous to a given source story. The incorrect options originally consist of two randomly selected stories and one distractor story with high nounal similarity to the source story. To focus our analysis on structural alignment in the presence of surface-level distractors, we discard the random options and adopt a two-option format. To minimize positional bias, we present each question twice with reversed indices and consider a response correct only if the model selects the target story in both trials. We report detailed statistics for both datasets in the Appendix.

3.2 Models

For proportional analogies, presented as a simple next-token prediction task, we investigate the following open-source models: Llama-2-13b (Touvron et al. 2023), Gemma-7B (Team et al. 2024), and Qwen2.5-14B (Yang et al. 2024). For story analogies, we use instruction-tuned models that demonstrate sufficient performance for analysis: Llama-2-13b-chat, Gemma-2-9B-it, and Qwen2.5-14B-Instruct. We mainly report results for Qwen2.5-14B models, as they exhibit representative behavior, and provide results for other models in the Appendix.

[1] We use the Wikidata subset.



(a) Correct case

(b) Incorrect case

Figure 2: Results of applying attention knockout to different positions on Qwen2.5-14B. Mid-upper layers of e_2 and e_3 are critical for answer resolution in both correct and incorrect cases. In incorrect cases, information from the link strongly influences model output, suggesting that the link may contribute to reasoning failures.

3.3 Implementation Details

For all experiments, we use two Nvidia A100 GPUs with 80GB memory. Our code is written in PyTorch (v2.3.1) and HuggingFace (v4.44.2). We report results on each model run by adopting greedy decoding to ensure reproducibility.

4 Information Flow in Analogical Reasoning

The most fundamental process of analogical reasoning involves encoding the elements of an analogy, identifying the relationship between them, and applying that relationship to a target element (French 2002; Gentner and Forbus 2011). In this section, we investigate whether this process holds true for LLMs as well using proportional analogies.

4.1 Methods

We first apply attention knockout to identify positions that are critical for resolving the answer. We focus on four positions that precede the resolution token: e_1, e_2, link, and e_3. For correct cases, we report the accuracy of the generated response. For incorrect cases, we check whether the knockout results in a change in the generated text to assess the impact of blocked layers (Biran et al. 2024). We keep a window of k layers around each layer to account for information that propagates across multiple layers (Geva et al. 2023), where k is set to one-fifth of the total number of layers.

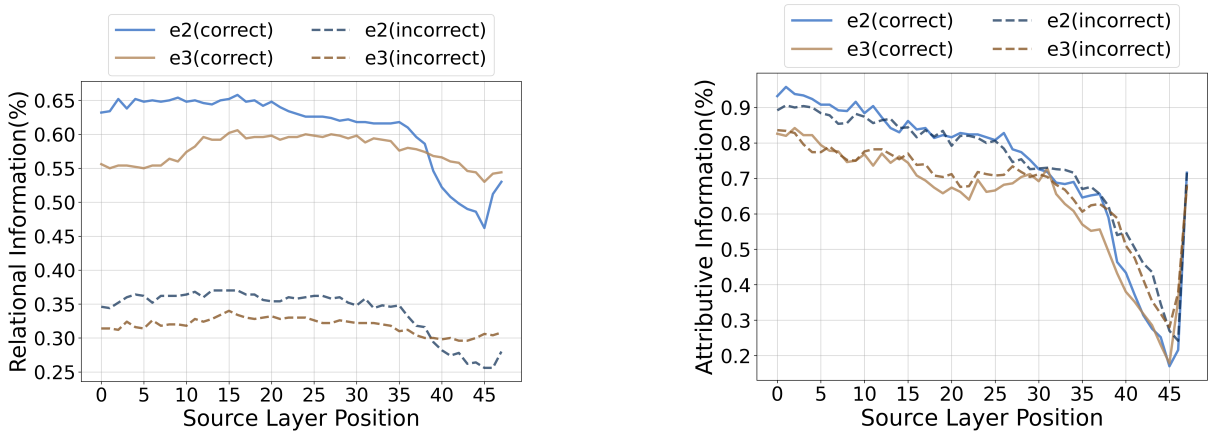
Next, to analyze what information is encoded in the hidden representations of these positions, we first categorize information into two types: *attributive* and *relational* information. Attributive information reflects how well the representation captures the inherent attributes of an entity, while relational information indicates whether the representation encodes the relation. To analyze attributive information within each entity, we employ Patchscopes and use the same target prompt used in Section 2.2 to obtain descriptions of hidden representations in natural language. Next, to check whether each description involves the correct attributes, we take inspiration from Geva et al. (2023) and construct a set of tokens highly related to the entity of interest. Specifically, for each entity, we retrieve 100 paragraphs from Wikipedia [2] using BM25 (Robertson et al. 1994), and extract related entities using en_core_web_trf (Honnibal and Montani 2017). We consider a hidden representation to have encoded attributive information if the corresponding description contains one or more entities related to the entity of interest.

For relational information, our goal is to inspect whether a specific entity encodes the correct relation. To achieve this, we design three target prompts for each entity, encouraging models to explicitly output the encoded relation while considering their respective positions. For e_2, we use the following prompt: "`Japan is to Tokyo: capital of, Theory of Evolution is to Charles Darwin: founder of, Peace is to olive branch: symbol of, {} is to x`", where curly brackets are replaced with e_1. Similarly, for e_3, we use the same exemplars but replace the final phrase with "`x is to {}`", where curly brackets are replaced with e_4. Finally, for the resolution token ("to"), we use "`{} is x`" for the final phrase. Note that we use two prompts, where curly brackets are replaced with either e_3 or e_4. We consider a hidden representation to have encoded relational information if the corresponding description contains the correct relation. We provide example descriptions generated from each custom prompt in the Appendix.

4.2 Results

Figure 2 shows the results of applying attention knockout to different positions preceding the resolution token, from which we identify three notable patterns. First, for both correct and incorrect cases, blocking attention edges from the resolution token to e_1 has little impact on model performance or generation. This indicates that e_1 plays a limited role in retaining information that is essential within the first pair. Second, blocking attention edges to either e_2 or e_3 re-

[2] We use the dump from December 20, 2024.



(a) Relational information across layers

(b) Attributive information across layers

Figure 3: Proportion of cases where relational or attributive information is successfully decoded using Patchscopes. Attributive information persists across mid-upper layers regardless of correctness, while relational information shows a sharp decline in incorrect cases. This underscores the critical role of relational information in accurate answer resolution.

sults in noticeable performance drops or fluctuations in generation, mainly around the mid-upper layers. This suggests that information propagating directly from e_2 and e_3 has a strong influence on model behavior, with information from e_2 propagating in slightly earlier layers than that from e_3. Third, information propagating from the link heavily affects model generations in incorrect cases, particularly in the early to middle layers. This either indicates an incorrect encoding of information passed to the link, or a failure of the link to transfer information to the target element. Based on this observation, we conduct further experiments to better understand incorrect cases in Section 5.

Figure 3 displays the proportion of cases where relational and attributive information is successfully decoded from each source layer. We see that attributive information is consistently encoded within e_2 and e_3, persisting until the mid-upper layers before declining sharply in the upper layers. Given that we ensure models are equipped with the necessary knowledge (Section 3.1), we confirm that attributive information remains intact for e_2 and e_3 in both correct and incorrect cases. However, a significant gap is observed between these cases in terms of relational information. This suggests that relational information encoded in e_2 and e_3 serves as a key factor in answer resolution. Moreover, while both types of information follow a similar trend for e_2, relational information in e_3 remains consistent up to the upper layers, implying its role in answer resolution at these layers.

5 Application as a Hurdle

For humans, the primary difficulty in solving analogies lies in extracting the underlying relation; once retrieved or cued, mapping it onto a new context is relatively straightforward (Kubricht, Lu, and Holyoak 2017). In the previous section, we have identified two potential explanations for model failures: incorrect encoding of information passed to the link, or ineffective transfer of information through the link itself. In this section, we aim to deepen our understanding of how models fail at analogical reasoning, focusing on the observed influence of the link in incorrect generations and the pivotal role of e_2, which encodes both attributive and relational information. We begin by re-evaluating model performance when provided with the correct first pair. For cases where the model still fails, we then intervene by patching the representations of e_2 into the linking position to better facilitate the propagation of critical relational information.

Model	Exp 1	Exp 2	Overall
Llama-2-13B	+32.3%	+25.9%	+49.8%
Gemma-7B	+38.4%	+38.1%	+61.9%
Qwen2.5-14B	+35.6%	+30.5%	+55.3%

Table 1: Results from error analysis experiments. "Exp 1" indicates setting where we evaluate models using correct first pairs. "Exp 2" indicates setting where we patch representations for the remaining incorrect cases.

5.1 Methods

For the first experiment, we replace the first pairs of incorrect cases with those from correct cases. To ensure a sufficient number of samples for replacement, we select three representative relations from our test set: "*official language of*", "*author of*", and "*composer of*". For each incorrect input analogy, we randomly choose a correct analogy from the same relation and swap their first pairs. We evaluate models using this newly constructed test set. For the second experiment, we patch the hidden representations of each layer in e_2 to each layer in the link to see if models can benefit from directly injecting critical information encoded in e_2. We re-

port the performance improvement from the combination of layers that yields the highest gain.

5.2 Results

Table 1 shows the performance gains observed from each experiment. We find that model responses can be rectified by replacing the first pair in up to 38.4% of incorrect cases. This indicates that a non-negligible portion of model errors stem from insufficient extraction of information within the first pair. This also highlights the importance of information encoded in e_2, as we have previously confirmed that the resolution token strongly attends to e_2 for answer resolution.

Interestingly, for cases where replacing the first pairs did not result in correct answers, we observe that patching the representations of e_2 to the link leads to noticeable performance gains up to 38.1%. This indicates that even if the model correctly extracts the necessary information from the first pair, the extent to which the link effectively conveys that information to subsequent positions can significantly impact model generation. Moreover, we inspect the generation results across different layers for both e_2 and the link. For e_2, we find that patching representations up to the middle layers is mainly effective in rectifying model responses. Given that both relational and attributive information is strongly formed up to the mid-upper layers of e_2 in correct cases, we see that injecting information encoded from these regions into the link assists in propagating these information to subsequent positions. For the link, where patching is performed, applying the patched representation to the early layers proves to be effective, suggesting that the representation need to pass through a certain number of layers to be properly contextualized with relational information.

6 Structural Alignment in Analogies

A crucial aspect of analogical reasoning is the concept of structural alignment, i.e., the process of establishing a one-to-one correspondence between elements of two situations in a way that maximizes relational similarity (Markman and Gentner 1993; Gentner and Forbus 2011). This ability goes beyond recognizing lexically similar positions in context, and involves identifying parallels between seemingly unrelated, high-level concepts. In this section, we first analyze internal representations to determine whether the model distinguishes analogical context from lexically similar context. We then examine how structural alignment emerges across layers when the task is explicitly posed, and how this progression influences model behavior.

6.1 Methods

For the first experiment, we extract the source, target, and distractor stories from each sample in the StoryAnalogy dataset. We construct a probing dataset by pairing each source story with both the target and distractor stories. For each input pair, we extract the activation at the final token from every attention head in each layer, yielding a probing dataset $\{(x_i^{(h,\ell)}, y_i)\}_{i=1}^{N}$, where $x_i^{(h,\ell)}$ denotes the activation from head h in layer ℓ for the i-th input pair. We train a binary linear classifier on these representations to assess whether analogical structure is linearly separable from lexical similarity in the model's internal representations. To ensure robust performance estimates and mitigate overfitting, we apply 5-fold cross-validation, reporting the average validation accuracy across folds as the final probe accuracy. For the second experiment, we assess whether structural alignment is reflected in the model's internal geometry during analogical reasoning, and how it diverges between correct and incorrect cases in the presence of distractor stories. To this end, we define the *Mutual Alignment Score* (MAS) as the proportion of mutual best matches between contextualized token representations from the source and candidate spans (Algorithm 1), computed at each layer for both the target and distractor stories. A token pair (s_i, c_j) forms a mutual best match if each is the other's most similar token based on cosine similarity between their layer-specific representations. By computing MAS across layers, we trace the emergence of structural alignment and examine its relationship with successful analogical reasoning.

Algorithm 1: Mutual Alignment Score

Input: Source token representations $S = \{s_1, \ldots, s_m\}$, Candidate token representations $C = \{c_1, \ldots, c_n\}$
Output: Mutual alignment score M

Normalize each vector in S and C to unit norm
Compute similarity matrix $M_{ij} = \cos(s_i, c_j)$ for all $i \in [1, m]$, $j \in [1, n]$
Initialize counter `mutual_matches` $\leftarrow 0$
for $i = 1$ to m **do**
  $j^* \leftarrow \arg\max_j M_{ij}$ // Best-matching in C for s_i
  $i^* \leftarrow \arg\max_{i'} M_{i'j^*}$ // Best-matching in S for c_{j^*}
  if $i^* = i$ **then**
    `mutual_matches` $\leftarrow$ `mutual_matches` $+ 1$
  end if
end for
$M \leftarrow$ `mutual_matches` / $\min(m, n)$
return M

6.2 Results

Figure 4 presents linear probe accuracies for distinguishing analogical from lexically similar stories across all layers of the model. The heatmap reveals a clear progression in representational quality across depth. Early to middle layers begin to show accuracies above chance, suggesting an initial emergence of analogical structure at relatively shallow depths. A marked increase in probe accuracy extends through the middle layers, with layers 20 through 30 showing an average accuracy of 82.9%. This pattern indicates that analogical distinctions are not immediately encoded at the input level but instead develop gradually across layers, reaching maximal discriminability in the middle layers of the model. These findings imply that models develop an internal representation of analogical structure that becomes linearly separable from lexical similarity as processing deepens.

Figure 5 shows the relative MAS between the source and target stories versus the source and distractor stories, measured as the difference in MAS across layers. For correct

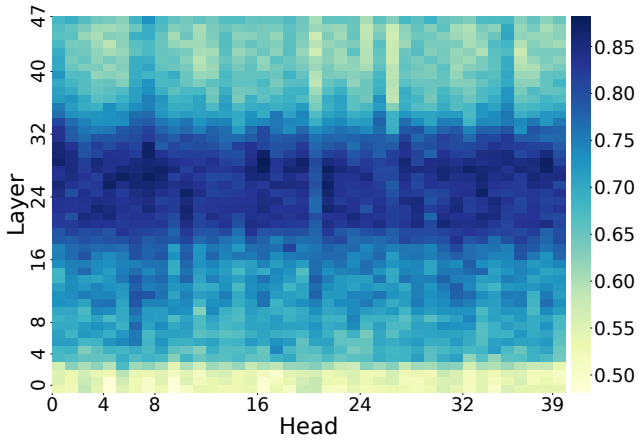


Figure 4: Linear probe accuracy across layers. High accuracy in the middle layers indicates the internal representation of analogical structure in these regions.

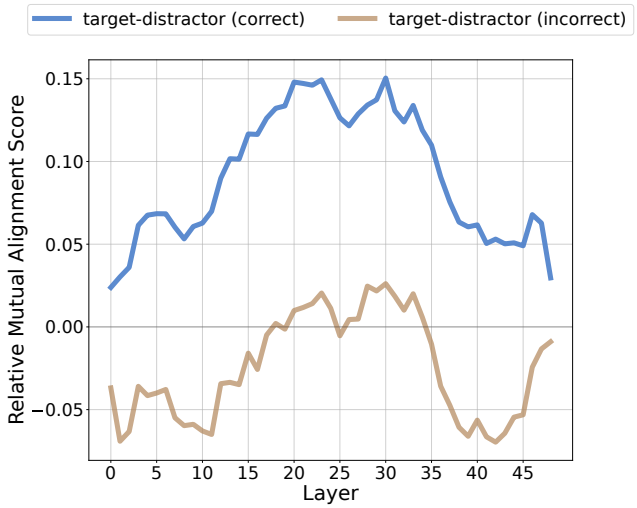


Figure 5: Relative Mutual Alignment Score (MAS) across layers, computed as the difference between the MAS of source-target pairs and source-distractor pairs.

cases, the MAS between source and target stories consistently exceeds that between source and distractor stories, suggesting that models encode deeper structural alignment beyond surface-level lexical cues. This is especially notable given that target stories are designed to have minimal entity overlap with the source, indicating that models are capturing underlying relational structure, similar to how humans seek one-to-one alignments that maximize relational similarity during analogical reasoning.

The relative gap peaks in the middle layers, suggesting that structural alignment between correct analogical pairs is strongest at these depths. This aligns with our probing results, which show that analogical distinctions become most linearly separable in these layers. In contrast, for incorrect cases, the model constructs stronger alignment between

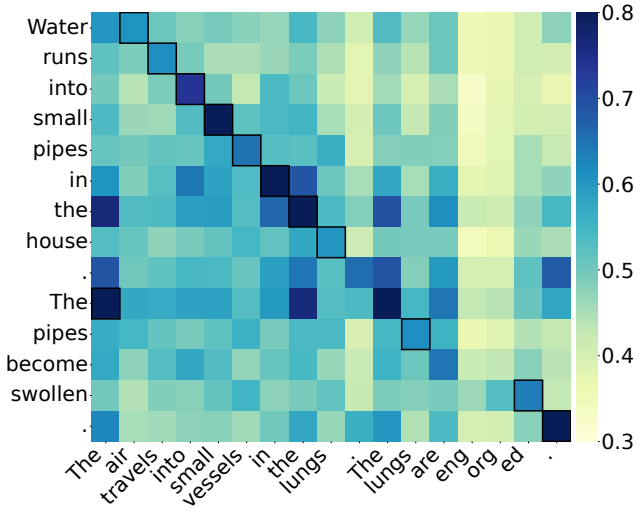


Figure 6: Sample heatmap of average similarity scores across layers between source and target stories. Black boxes indicate mutual best matches. Analogous token pairs (e.g., *Water-air*, *house-lungs*) form mutual best matches with high similarity scores, despite surface-level disparities.

source and distractor stories across most layers, with only a slight preference for source–target alignment in the middle layers. The overall gap is much less pronounced than in correct cases, indicating that the model fails to reliably identify the intended analogical structure.

Overall, these results indicate that successful analogical reasoning in the model is strongly associated with higher token-level structural alignment between source and target stories. In contrast, incorrect cases exhibit a much smaller alignment gap, with distractors often receiving greater alignment, suggesting that the model fails to clearly differentiate the intended relational structure. This highlights structural alignment as a key internal signal for analogical success and reveals the model's vulnerability to surface-level interference when the relational mapping is not robustly encoded.

7 Conclusion

In this work, we study the internal mechanisms of LLMs in analogical reasoning. Using proportional analogies, we find that correct reasoning is associated with the encoding of abstract relational information in the mid-upper layers. While models are capable of abstracting these relations, we find that applying them remains a major bottleneck. By patching the representation of the second entity into the link, we uncover the link's role in transferring relational information to downstream positions, and show that failure at this stage leads to incorrect generations. Finally, our analysis of story analogies shows that successful reasoning aligns with strong structural mapping between source and target stories, while failures often reflect weak or distractor-biased alignment. Overall, our work paves the way for future research into understanding and improving the analogical reasoning capabilities of LLMs.

Acknowledgments

We thank Minbyul Jeong, Hyeon Hwang, and Yein Park for their invaluable feedback on this work. This research was supported by the National Research Foundation of Korea (NRF-2023R1A2C3004176), the Ministry of Health & Welfare, Republic of Korea (HR20C002103), the Ministry of Science and ICT (MSIT) (RS-2023-00262002), the ICT Creative Consilience program through the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the MSIT (IITP-2025-RS-2020-II201819), and the Culture, Sports and Tourism R&D Program through the Korea Creative Content Agency (KOCCA) grant funded by the Ministry of Culture, Sports and Tourism (MCST) in 2023 (Project Name: Development of storytelling AI technology for cultural heritage tailored to the various interests of users, Project Number: RS-2023-00220195, Contribution Rate: 100%).

References

Alain, G.; and Bengio, Y. 2018. Understanding intermediate layers using linear classifier probes. arXiv:1610.01644.

Belinkov, Y. 2022. Probing Classifiers: Promises, Shortcomings, and Advances. *Computational Linguistics*, 48(1): 207–219.

Belrose, N.; Furman, Z.; Smith, L.; Halawi, D.; Ostrovsky, I.; McKinney, L.; Biderman, S.; and Steinhardt, J. 2023. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*.

Bereska, L.; and Gavves, E. 2024. Mechanistic Interpretability for AI Safety–A Review. *arXiv preprint arXiv:2404.14082*.

Biran, E.; Gottesman, D.; Yang, S.; Geva, M.; and Globerson, A. 2024. Hopping Too Late: Exploring the Limitations of Large Language Models on Multi-Hop Queries. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 14113–14130. Miami, Florida, USA: Association for Computational Linguistics.

Boteanu, A.; and Chernova, S. 2015. Solving and explaining analogy questions using semantic networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Brown, W. R. 1989. Two traditions of analogy. *Informal Logic*, 11(3).

French, R. M. 2002. The computational modeling of analogy-making. *Trends in cognitive Sciences*, 6(5): 200–205.

Gentner, D.; and Forbus, K. D. 2011. Computational models of analogy. *Wiley interdisciplinary reviews: cognitive science*, 2(3): 266–276.

Gentner, D.; and Markman, A. B. 1996. Keith J. Holyoak and Paul Thagard, Mental Leaps: Analogy in Creative Thought. *Pragmatics & Cognition*, 4(2): 407–409.

Geva, M.; Bastings, J.; Filippova, K.; and Globerson, A. 2023. Dissecting Recall of Factual Associations in Auto-Regressive Language Models. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 12216–12235. Singapore: Association for Computational Linguistics.

Ghandeharioun, A.; Caciularu, A.; Pearce, A.; Dixon, L.; and Geva, M. 2024. Patchscopes: a unifying framework for inspecting hidden representations of language models. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Gladkova, A.; Drozd, A.; and Matsuoka, S. 2016. Analogy-based detection of morphological and semantic relations with word embeddings: what works and what doesn't. In Andreas, J.; Choi, E.; and Lazaridou, A., eds., *Proceedings of the NAACL Student Research Workshop*, 8–15. San Diego, California: Association for Computational Linguistics.

Hendel, R.; Geva, M.; and Globerson, A. 2023. In-Context Learning Creates Task Vectors. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Findings of the Association for Computational Linguistics: EMNLP 2023*, 9318–9333. Singapore: Association for Computational Linguistics.

Hofstadter, D. R. 2001. Epilogue: Analogy as the core of cognition.

Hofstadter, D. R.; and Sander, E. 2013. *Surfaces and essences: Analogy as the fuel and fire of thinking*. Basic books.

Holyoak, K.; Gentner, D.; and Kokinov, B. 2001. The place of analogy in cognition. *The analogical mind: Perspectives from cognitive science*, 119.

Honnibal, M.; and Montani, I. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. *To appear*.

Imani, S.; Du, L.; and Shrivastava, H. 2023. Mathprompter: Mathematical reasoning using large language models. *arXiv preprint arXiv:2303.05398*.

Jiayang, C.; Qiu, L.; Chan, T.; Fang, T.; Wang, W.; Chan, C.; Ru, D.; Guo, Q.; Zhang, H.; Song, Y.; Zhang, Y.; and Zhang, Z. 2023. StoryAnalogy: Deriving Story-level Analogies from Large Language Models to Unlock Analogical Understanding. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 11518–11537. Singapore: Association for Computational Linguistics.

Keane, M. T. 1996. On Adaptation in Analogy: Tests of Pragmatic Importance and Adaptability in Analogical Problem Solving. *The Quarterly Journal of Experimental Psychology Section A*, 49(4): 1062–1085.

Kojima, T.; Gu, S. S.; Reid, M.; Matsuo, Y.; and Iwasawa, Y. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35: 22199–22213.

Kubricht, J. R.; Lu, H.; and Holyoak, K. J. 2017. Individual differences in spontaneous analogical transfer. *Memory & Cognition*, 45(4): 576–588.

Markman, A.; and Gentner, D. 1993. Structural Alignment during Similarity Comparisons. *Cognitive Psychology*, 25(4): 431–467.

Mikolov, T.; Yih, W.-t.; and Zweig, G. 2013. Linguistic Regularities in Continuous Space Word Representations. In Vanderwende, L.; Daumé III, H.; and Kirchhoff, K., eds., *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 746–751. Atlanta, Georgia: Association for Computational Linguistics.

nostalgebraist. 2020. Interpreting GPT: The Logit Lens. *LessWrong*.

Opiełka, G.; Rosenbusch, H.; and Stevenson, C. E. 2025. Analogical reasoning inside large language models: Concept vectors and the limits of abstraction. *arXiv preprint arXiv:2503.03666*.

Pal, K.; Sun, J.; Yuan, A.; Wallace, B.; and Bau, D. 2023. Future Lens: Anticipating Subsequent Tokens from a Single Hidden State. In Jiang, J.; Reitter, D.; and Deng, S., eds., *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*. Singapore: Association for Computational Linguistics.

Pochinkov, N.; Benoit, A.; Agarwal, L.; Majid, Z. A.; and Ter-Minassian, L. 2024. Extracting Paragraphs from LLM Token Activations. In *MINT: Foundation Model Interventions*.

Robertson, S.; Walker, S.; Jones, S.; Hancock-Beaulieu, M.; and Gatford, M. 1994. Okapi at TREC-3. 0–.

Sultan, O.; and Shahaf, D. 2022. Life is a Circus and We are the Clowns: Automatically Finding Analogies between Situations and Processes. In Goldberg, Y.; Kozareva, Z.; and Zhang, Y., eds., *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 3547–3562. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

Team, G.; Mesnard, T.; Hardin, C.; Dadashi, R.; Bhupatiraju, S.; Pathak, S.; Sifre, L.; Rivière, M.; Kale, M. S.; Love, J.; et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Todd, E.; Li, M.; Sharma, A. S.; Mueller, A.; Wallace, B. C.; and Bau, D. 2024. Function Vectors in Large Language Models. In *The Twelfth International Conference on Learning Representations*.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Vig, J.; Gehrmann, S.; Belinkov, Y.; Qian, S.; Nevo, D.; Singer, Y.; and Shieber, S. 2020. Investigating Gender Bias in Language Models Using Causal Mediation Analysis. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 12388–12401. Curran Associates, Inc.

Wang, F.; Mo, W.; Wang, Y.; Zhou, W.; and Chen, M. 2023a. A Causal View of Entity Bias in (Large) Language Models. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Findings of the Association for Computational Linguistics: EMNLP 2023*, 15173–15184. Singapore: Association for Computational Linguistics.

Wang, K. R.; Variengien, A.; Conmy, A.; Shlegeris, B.; and Steinhardt, J. 2023b. Interpretability in the Wild: a Circuit for Indirect Object Identification in GPT-2 Small. In *The Eleventh International Conference on Learning Representations*.

Webb, T.; Holyoak, K. J.; and Lu, H. 2023. Emergent analogical reasoning in large language models. *Nature Human Behaviour*, 7(9): 1526–1541.

Wijesiriwardene, T.; Wickramarachchi, R.; Vennam, S.; Jain, V.; Chadha, A.; Das, A.; Kumaraguru, P.; and Sheth, A. 2024. Exploring the Abilities of Large Language Models to Solve Proportional Analogies via Knowledge-Enhanced Prompting. *arXiv preprint arXiv:2412.00869*.

Xu, N.; Wang, F.; Li, B.; Dong, M.; and Chen, M. 2022. Does Your Model Classify Entities Reasonably? Diagnosing and Mitigating Spurious Correlations in Entity Typing. In Goldberg, Y.; Kozareva, Z.; and Zhang, Y., eds., *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 8642–8658. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

Yang, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Li, C.; Liu, D.; Huang, F.; Wei, H.; et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Yao, S.; Zhao, J.; Yu, D.; Du, N.; Shafran, I.; Narasimhan, K.; and Cao, Y. 2023. ReAct: Synergizing Reasoning and Acting in Language Models. In *International Conference on Learning Representations (ICLR)*.

Yasunaga, M.; Chen, X.; Li, Y.; Pasupat, P.; Leskovec, J.; Liang, P.; Chi, E. H.; and Zhou, D. 2024. Large Language Models as Analogical Reasoners. In *The Twelfth International Conference on Learning Representations*.

Ye, X.; Wang, A.; Choi, J.; Lu, Y.; Sharma, S.; Shen, L.; Tiyyala, V. M.; Andrews, N.; and Khashabi, D. 2024. AnaloBench: Benchmarking the Identification of Abstract and Long-context Analogies. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 13060–13082. Miami, Florida, USA: Association for Computational Linguistics.

Yuan, S.; Chen, J.; Sun, C.; Liang, J.; Xiao, Y.; and Yang, D. 2024. ANALOGYKB: Unlocking Analogical Reasoning of Language Models with A Million-scale Knowledge Base. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1249–1265. Bangkok, Thailand: Association for Computational Linguistics.

Zhang, F.; and Nanda, N. 2024. Towards Best Practices of Activation Patching in Language Models: Metrics and Methods. In *The Twelfth International Conference on Learning Representations*.