Investigating noun-noun compound relation representations in autoregressive large language models

Saffron Kendrick Mark Ormerod Hui Wang Barry Devereux

Queen's University Belfast, Northern Ireland
{skendrick01, mormerod01, h.wang, b.devereux}@qub.ac.uk

Abstract

This paper uses autoregressive large language models to explore at which points in a given input sentence the semantic information is decodable. Using representational similarity analysis and probing, the results show that autoregressive models are capable of extracting the semantic relation information from a dataset of noun-noun compounds. When considering the effect of processing the head and modifier nouns in context, the extracted representations show greater correlation after processing both constituent nouns in the same sentence. The linguistic properties of the head nouns may influence the ability of LLMs to extract relation information when the head and modifier words are processed separately. Probing suggests that Phi-1 and LLaMA-3.2 are exposed to relation information during training, as they are able to predict the relation vectors for compounds from separate word representations to a similar degree as using compositional compound representations. However, the difference in processing condition for GPT-2 and DeepSeek-R1 indicates that these models are actively processing the contextual semantic relation information of the compound.

1 Introduction

The popularity of transformer-based large language models (LLMs) has skyrocketed since the success of Vaswani (2017) with the attention mechanism and the conception of Bidirectional Encoders from Transformers (BERT) (Devlin et al., 2019). The attention-based architecture of LLMs allows them to carry out a wide variety of natural language processing (NLP) tasks, such as classification, sentiment analysis, translation and text generation.

Despite the positive reception and widespread implementation of LLMs, the internal processes of these complex models remain a key question within the fields of interpretable and explainable AI. In particular, the notion that state-of-the-art (SoTA) LLMs can process and understand word meaning in a similar way to natural language understanding remains an ongoing discussion (Bender and Koller, 2020; Piantadosi and Hill, 2022). This has inspired research into the syntactic and semantic capabilities of language models in an attempt to unify computational processes and human language processing.

The objective of this paper is to expand on the work of Ormerod et al. (2024) to investigate whether SoTA autoregressive models are capable of representing the semantic relation information of noun-noun compounds, and where in an input sequence the semantic information is decodable. The original framework uses representational similarity analysis (RSA) to compare the extracted token representations with two datasets of English noun-noun compounds. The token representations considered are suited for the bidirectional masked language models, however, autoregressive LLMs are unidirectional, meaning that they only rely on the previous inputs. Therefore, the experiments are adapted to incorporate a continuation word, taking the final head word and modifier token representations, and the token representation of the continuation word.

The models in this paper include RoBERTa (Liu et al., 2019), BERT-Japanese, GPT-2 (Radford et al., 2019), Phi-1 (Gunasekar et al., 2023), LLaMA-3.2 (Dubey et al., 2024), and DeepSeek-R1-Distill-Qwen-1.5B (DeepSeek-AI et al., 2025). BERT-Japanese acts as a control subject because it is not trained on English and therefore should not be able to decode the English semantic relation. RoBERTa is included as the top performing encoder model.

The results show that autoregressive models are capable of decoding the semantic relation information, with strongest correlation occurring from the final word token representations. The final head noun token also holds relation information, which may be accounted for by the level of concreteness

Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, pages 253–263

of the head noun. However, the modifier word representations still contain some level of relation information, potentially reflecting frequency of relational information associated with specific modifier nouns in the training data. The models decode semantic information from the contextual representations, although Phi and LLaMA may learn information about coexisting compound relations during training as these models can predict relation information from individual word representations.

2 Background

The first transformer-based model, designed for a text translation task, consisted of six encoders and six decoders which were able to convert input sequences into output sequences. In the transformer, encoders are responsible for generating word embeddings which capture the content of an input, and positional encodings which provide information on the position of each token in the sequence. The multi-headed self-attention mechanism is a key part of the architecture which enables each encoder to focus on different parts of the input as it processes each token. Attention involves calculating the dot product of the query and key vectors, which indicates the level of emphasis that each word should place on other words. The attention weights are then passed through a softmax layer which gives a probability distribution that informs the model how much of each value representation to carry through to the next layer.

An alternative to the traditional BERT architecture, autoregressive decoder-based LLMs are unidirectional, which means that they rely on the previous input to predict the next token in the sequence. They are causal language models, consisting of stacks of decoder layers which take an input sequence and predict the next most likely term. These autoregressive models are often used for text generation and chatbots which are available to the public, thus it is crucial that their internal processes are investigated.

2.1 Probing LLMs

Probing is a technique commonly used within NLP interpretability to investigate whether the representations are able to capture certain information (Hewitt and Manning, 2019). Due to the complexity of blackbox models, probing methods are often extrinsic, post-hoc approaches. Classifiers are used to determine whether a model can successfully decode an abstract concept, although they do not provide causal information. For transformer models, probing often includes investigating the attention heads within the self-attention mechanism, embeddings, and token representations. Ju et al. (2024) used layer-wise probing to investigate how LLMs encode context, highlighting the emphasis that LLMs place on context knowledge in upper layers. This is supported by the work of Jawahar et al. (2019), using sentence-level probing to explore BERT's phrasal representations. They concluded that BERT encodes linguistic information including syntactic features in its middle layers and semantic features in the upper layers. Other probing studies have investigated function word comprehension, long-distance agreement, and other syntactic phenomena (Kim et al., 2019; Linzen and Baroni, 2021; Vulić et al., 2020). Probing proves to be a well-established method for exploring how LLMs are able to encode and decode semantic and syntactical information.

2.2 Conceptual combination

A major field of research that aims to bridge the gap between human language understanding and NLP focuses on the compositionality of words to form larger, meaningful phrases and sentences, a process known as conceptual combination. This process can be linked to concepts that are intersective, such as adjective-noun phrases that are overlaps of their constituent words, as well as noun-noun phrases which consist of a head noun and a modifier. A subset of noun-noun phrases can be considered lexical compounds, where they are highly idiomatised within language such that the combined meaning is not apparent from the meanings of the individual nouns themselves.

Early theories of intersective combination take inspiration from mathematical principles, proposing a fuzzy logic model that relies on a degree of overlap between two concepts. This early model led to the Selective Modification (Smith and Osherson, 1984; Smith et al., 1988) and Concept Specialisation models (Cohen and Murphy, 1984) which can be described as schema-based, where the head noun is represented by a set of empty slots and fillers, and its specialisation is determined by a modifier filling one or more of its slots. The dual-process model proposes a similar framework, however this model suggests three approaches to conceptual combination: relation-based, property-based, and a hybridisation of two concepts (Wis-

niewski, 1997). In this instance, relations are represented as slots within the schema of a head noun and when a modifier fills a slot, an appropriate relation is chosen. Building on the reasoning that relations drive conception, the Competition Among Relations in Nominals (CARIN) theory implies that the modifier representation contains knowledge of certain relations that are frequently used with the given modifier during conceptual combination, known as the relational distribution (Gagné, 2001).

2.3 Semantic properties of LLMs

Research into conceptual combination from a linguistic standpoint is far from complete, however early theories provide a starting point for probing LLMs to discern whether SoTA models extract meaningful representations about syntactic and semantic properties of language. In particular, insights into how LLMs handle complex linguistic structures can shed light on the internal mechanisms and how they relate to or deviate from human cognitive processes.

Conceptual combination in the context of NLP has primarily focused on using features to classify the relations between a head noun and a modifier word. Ó Séaghdha and Copestake (2008) adopted distributional kernels for three types of semantic classification, including the interpretation of compound nouns. For transformer-based language models, word embeddings have become an area of interest for probing the semantic capabilities. Peters et al. (2018) concluded that the complex architecture of transformers are capable of learning a hierarchy of linguistic features. Shwartz and Dagan (2019) evaluated both static and contextualised embeddings, concluding that contextualising improves performance, especially for recognising meaning shifts. Ettinger (2020) extracted word embeddings to assess phrasal similarity across layers of transformer models for two-word phrases, concluding that although models are able to represent individual word content, they struggle at representing the full compositional phrase meaning. Derby et al. (2021) investigated how the intermediate layer of long short-term memory (LSTM) models and transformers capture semantic knowledge, showing that transformers outperform LSTMs although both are able to retain semantic information after the target concept has been provided to the model.

Most recently, Coil and Shwartz (2023) investigated the interpretation and conceptualisation of noun-noun compounds on a supervised seq2seq model and GPT-3, an autoregressive LLM. They found that GPT-3 outperformed the seq2seq model when interpreting known compounds, however the LLM struggled to generalise to unseen, novel compounds. They suggested that GPT-3 relied heavily on memorisation to interpret previously seen compounds, leading to hallucinations when interpreting new compounds. Ormerod et al. (2024) focused on six encoder-based LLMs, including a multilingual model and a non-English monolingual model, to investigate whether LLMs are capable of representing the thematic relation shared between two constituent nouns within a compound. Their work highlighted the ability of BERT and RoBERTa to encode the thematic relation between the head and modifier, although they did not consider autoregressive models. Rambelli et al. (2024) also investigated the semantic relationships shared across compounds, using prompting and the Surprisal metric on a dataset of noun-noun compounds annotated with both semantic relations and concreteness ratings. Their results indicated that models identified semantic relations to varying degrees, influenced by the concreteness of a given compound. However, similar to Coil and Shwartz (2023), they found that LLMs were limited in their ability to generalise to novel compounds.

As an extension to Ormerod et al. (2024), this paper provides further support to the conclusions that autoregressive LLMs are able to extract implicit relation information after processing the full compound. Probing uses fine-grained relation information to explore the semantic information extracted from compounds at a higher level of granularity.

3 Data

Two datasets are used to explore the thematic relations of noun-noun compounds. The first dataset includes 300 English noun-noun compounds that are categorised into groups of 5 compounds (Gagné, 2001). Each group consists of a target compound, a compound with the same head noun and the same relation, a compound with the same head noun but a different relation, a compound with the same modifier word and the same relation, and a compound with the same modifier and a different relation. 60 groups of five compounds are constructed and a ground-truth representational dissimilarity matrix (RDM) is constructed to reflect whether or not pairs of compounds share the same thematic re-

lation. Within each group of five compounds, there is one target compound, followed by four others that comply with the experimental conditions given in Table 1.

M	H	Experimental Condition
gas	lamp	Target
battery	lamp	Same H, same relation
cabin	lamp	Same H, different relation
gas	car	Same M, same relation
gas	hose	Same M, different relation

Table 1: Experimental conditions for each group of five noun-noun compounds used in the relation category RSA experiments, with the modifier (M) and the head noun (H).

The second dataset consists of 60 noun-noun compounds, where 34 participants were tasked with ranking the appropriateness of 18 possible relations for each compound (Devereux and Costello, 2005). This results in a dataset of 18-dimensional relation vectors. Compounds which are semantically linked, i.e. share the same thematic relation, tend to have similar relation vectors. This dataset provides a fine-grained representation of the semantic information for each compound, useful for probing the semantic capabilities of the LLMs.

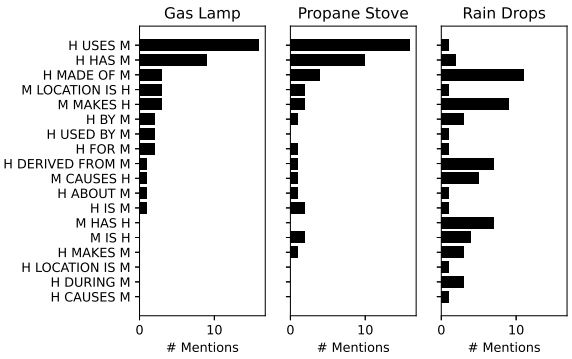

Figure 1: Sample relation vectors for three of the 60 compounds in the 60 compound dataset. Compounds GAS LAMP and PROPANE STOVE share similar relation vectors, when compared with RAIN DROPS (Devereux and Costello, 2005).

4 Experiments

The experimentation consists of two separate experiments, firstly using RSA to assess whether the token representations extracted layer-by-layer reflect the semantic relation information shared across the head and modifier of noun-noun compounds, and secondly using a linear probing classifier to discern whether the LLMs can successfully decode the thematic relation.

Experiment 1, also known as the relation category experiment, is designed to determine whether the relation between nouns influences the model's ability to distinguish between noun-noun compounds when presented in pairs. RSA, a technique commonly used in computational neuroscience, is useful for comparing disparate data sources by creating similarity matrices and analysing any shared structure, or lack of, by calculating the Pearson r correlation between the two matrices (Kriegeskorte et al., 2008). Experimental RDMs are constructed by calculating the cosine similarity of the extracted token representations, to compare with ground-truth RDMs which reflect whether two compounds share the same thematic relation (similar) or not (dissimilar), see Figure 2. The correlation indicates how strongly the extracted representations reflects the relation information represented by the ground-truth RDM, i.e. the category of relation for each compound.

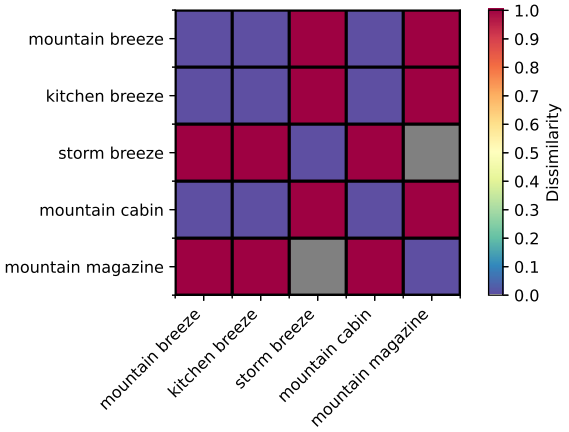

Figure 2: Ground-truth RDM, representing whether two compounds share the same thematic relation. The compounds that do not share either the same relation, head or modifier is not included in the Pearson r calculations (Ormerod et al., 2024).

This experiment also investigates the effect of considering the head and modifier in the same sentence, as opposed to considering each word separately. Higher correlation when the compound is processed together in the same sentence would indicate that the semantic relation information is represented by the models across the compound.

The second experiment, known as the compositional probe, applies a linear classifier to probe

whether context is required in order to decode fine-grained relation information. The original framework set out by Mitchell and Lapata (2008) has been adapted for this probe experiment, where the 2-vs-2 testing framework is used to determine whether the extracted representations of pairs of compounds align more with ground-truth or with each other. Using the fine-grained 60-compound dataset, for each possible pair of compounds (out of a possible 1770), a linear regression model is trained on the mean-pooled representations for RoBERTa and BERT-Japanese, and the final word token representations of the remaining compounds for the autoregressive models to predict the 18-dimension relation vectors.

The regression model generates predictions $Y^{\tilde{i}}$ and $Y^{\tilde{j}}$ from Y^i and Y^j. A test is considered successful if it satisfies

$$\begin{aligned} &\text{dist}(Y^{\tilde{i}}, Y^i) + \text{dist}(Y^{\tilde{j}}, Y^j) < \\ &\text{dist}(Y^{\tilde{i}}, Y^j) + \text{dist}(Y^{\tilde{j}}, Y^i), \end{aligned} \tag{1}$$

where the distances are calculated as mean squared errors. Therefore, if the predicted vectors for i and j are closer to the true relation vectors for i and j, rather than the other way around, the test is marked as successful. The probing experiment considers two processing conditions, where the head and modifier word are processed together in context and where they are processed separately before being averaged. If the proportion of successful tests is high for the contextual processing condition, this would suggest that the models are actively processing the contextual composition of each compound, rather than relying on previously learned association information.

Four publicly available autoregressive models are used, including GPT-2-Small, Phi-1, LLaMA-3.2-3B, and DeepSeek-R1-Distill-Qwen-1.5B. All four models adopt different tokenisers to break down the input sequences into sub-units (tokens). DeepSeek is of particular interest as it adopts Mixture of Experts (MoE) approach, allowing it to predict multiple potential outputs simultaneously. For RoBERTa and BERT-Japanese, the input sequences are taken as the Gloss sentences from the dataset, e.g. "It is a MOUNTAIN STREAM." However, as the autoregressive models read from left to right, the input sequences include an additional continuation word so that the token representation can be extracted from the final word, e.g. "It is a MOUNTAIN STREAM *[and/but/that]*", where the average of applying each word is taken. When taking the processing condition into account, two separate sentences are used, each employing the head noun and modifier word, e.g. "It is a MOUNTAIN." and "It is a STREAM." Once again, for GPT, Phi, DeepSeek and LLaMA, the sentences include a continuation word.

4.1 Relation category RSA

Three experimental RDMs are constructed for each layer within the models by calculating the pairwise cosine similarity for the mean-pooled token representations, the head noun representations, and the modifier word representations for the BERT models. For the autoregressive models, the final head, final modifier, and final word representations are used. The Pearson r correlation between each experimental RDM and the ground-truth RDM is then plotted to show how well the extracted representations reflect the thematic relation of each compound, see Figure 3.

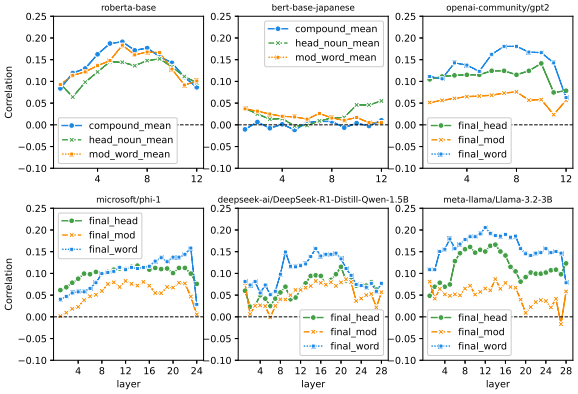

Figure 3: Results for the relation category RSA experiment (Section 4.1). The Pearson r correlation is calculated between the same thematic relation ground-truth RDM and experimental RDMs across all layers of six transformer models.

Figure 3 shows that all models, excluding BERT-Japanese, are capable of capturing the semantic relation information. As expected, the non-English model performs poorly, although the head and modifier representations produce higher correlation than the mean-pooled compound representations, suggesting that the separate representations may contribute to the thematic relation information.

For the autoregressive models, the final word representation produces the highest correlation to ground truth, followed by the final head noun representations. The correlation for the modifier word representation is non-zero, which may reflect statis-

tical and semantic information about how modifier words are associated with thematic relations in the language data that the models have been exposed to during pre-training. The results for DeepSeek follow a similar trend to LLaMA, which could be due to the fact that DeepSeek-R1 is partly based on the LLaMA model. For all four autoregressive models, the correlation drops for the final word representations at the final layer.

4.2 Relation category RSA with processing condition

This is an extension of the previous experiment that considers whether context affects how well the models extract the thematic relation. Two experimental RDMs are constructed by comparing the cosine similarity of the mean-pooled token representations for BERT models, and the final word token representations for the four autoregressive models, of the compounds presented in context (the "Together" condition) and in separate sentences (the "Separate" condition). The results are presented in Figure 4.

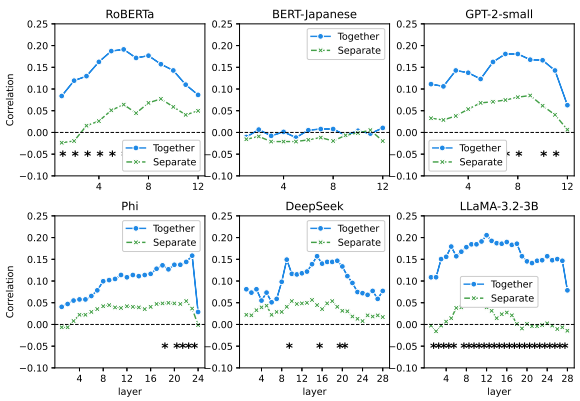

Figure 4: Results for the relation category RSA experiment with the processing condition (Section 4.2). The Pearson r correlation is taken between the same relation ground-truth RDM and the experimental RDMs of the mean-pooled and final word representations for the "Together" and "Separate" processing conditions. Asterisks reflect $p < 0.05$, i.e. under a paired t-test, the difference in correlation across processing conditions is statistically significant.

The "Together" processing condition consistently produces greater correlations when compared with the "Separate" condition, suggesting that GPT, Phi, DeepSeek and LLaMA represent the semantic relation information across the compound. In particular, after performing a paired t-test, the results are statistically significant across almost all layers for LLaMA, with statistically significant differences for GPT and DeepSeek across the middle layers, and Phi towards the later layers.

Processing the head and modifier words separately still alludes to some ability for the models to capture the relation information, which could be the result of the models taking into account the frequency of modifiers and head nouns coexisting with a particular relation during training.

For all models except BERT-Japanese, the gap between the correlations of each processing condition is most defined through the middle layers, supporting the results above that semantic information shared across the compound is encoded in the middle layers. Correlation for the "Separate" case for the autoregressive models falls almost to zero, implying that the token representations reflect very little semantic information in the final layers.

4.3 Compositional probe

The 2-vs-2 tests are performed pairwise to investigate whether context is required for decoding fine-grained semantic information.

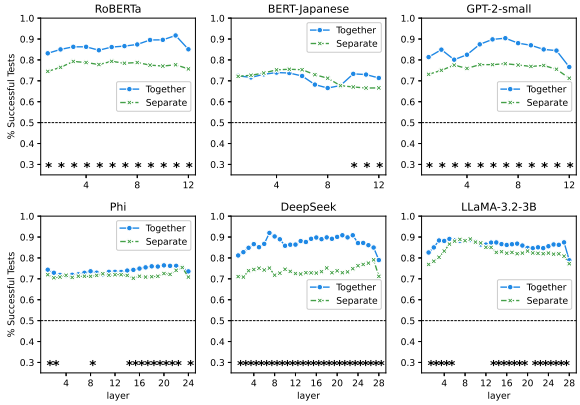

Figure 5: Results for the compositional probe experiment (Section 4.3). Proportion of successful tests out of 1770 tests, i.e. the models decode the thematic relation, using both the "Together" and "Separate" processing conditions. Asterisks signify tests for false-discovery, where $p < 0.05$, i.e. there is a significant difference between the number of successes across both processing conditions.

In Figure 5, the results for DeepSeek and GPT suggest that the models capture relation information in context, and contextual combination is responsible for creating representations that reflect the semantic information. However, there is little difference between the processing conditions for Phi and LLaMA, implying that the models are already aware of particular thematic relations and

can identify these from the individual words.

In addition to the probing experiment, a false-discovery procedure is performed in order to account for statistical dependencies. These occurrences are marked as asterisks. The results for DeepSeek in particular show a large number of these tests across all layers, despite such a high proportion of successful tests.

5 Representations across layers

The experiments show that these models are capable of extracting semantic information, but in order to identify the points in the input sentences at which each model is able to decode the relation information, the correlation is plotted layer-wise against the input sentence.

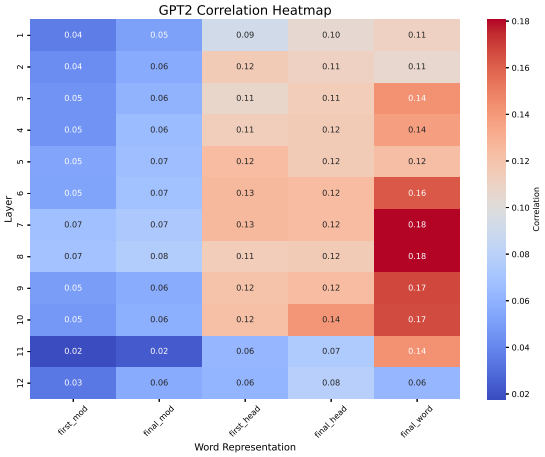

Figure 6: Correlation between the final word token representation RDM and the ground truth RDM across layers for GPT.

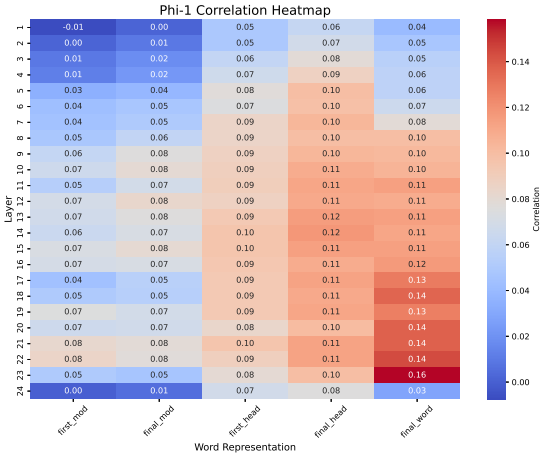

Figure 7: Correlation between the final word token representation RDM and the ground truth RDM across layers for Phi.

For GPT, Figure 6 shows that positive correlation begins at the first head token position. This implies that, once the model has been exposed to the full modifier word, it is capable of extracting a level of semantic relation information. The correlation is strongest in the middle and later layers for the final word representations, which aligns with the results from the relation category RSA experiment (Section 4.1).

Whilst there is a clear distinction between the modifier and head token correlations for GPT, the heatmap for Phi in Figure 7 shows that the modifier tokens do correlate to some extent with ground-truth, although this is not high at 0.08. Once again, the correlation is stronger towards the end of the input sequence, with the greatest correlation produced by the final word token representations.

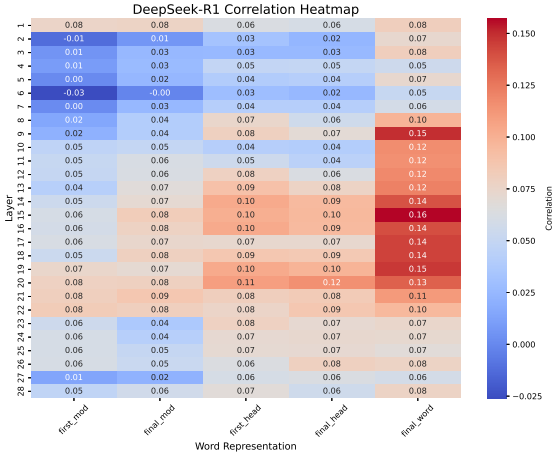

Figure 8: Correlation between the final word token representation RDM and the ground truth RDM across layers for DeepSeek.

For DeepSeek, the heatmap in Figure 8 shows that the modifier tokens may reflect some of the semantic relation information of the compounds. The correlation is most pronounced across the middle layers of the final word representation, before decreasing.

Similar to GPT, Figure 9 shows that the correlation for LLaMA begins to increase after the model has been presented with the full modifier. LLaMA achieves the highest correlation of 0.21 after processing the compound.

6 Discussion

The experimental results support the conclusions that transformer-based LLMs can retrieve the semantic relation information of noun-noun compounds. Intuitively, the averaged token representa-

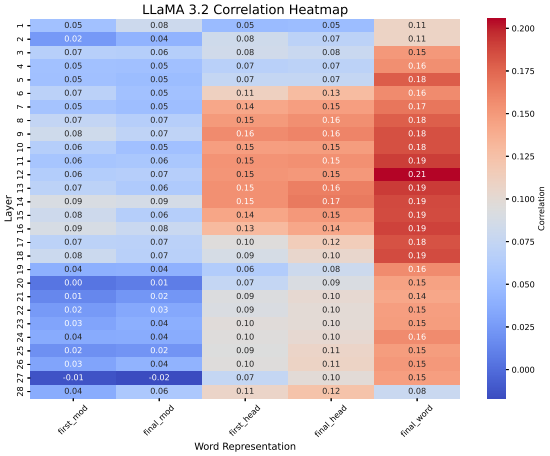

Figure 9: Correlation between the final word token representation RDM and the ground truth RDM across layers for LLaMA.

tion for the continuation words led to the greater correlation, which is to be expected as the autoregressive models will have processed the full compound. This result contradicts Ettinger (2020), who concluded that representations of two-word phrase embeddings do not reflect the semantic phrasal composition.

The head noun representations also correlate with the ground truth RDMs, suggesting some capability for the head nouns to store relational information. A possible explanation is the intrinsic semantic properties of the constituent nouns in the compounds, where head nouns reflect concrete concepts. From the fine-grained 60-compound dataset, the top mentioned relations include "H FOR M", "H MADE OF M", and "H USES M". These relations are more likely to be shared by at least one concrete noun (BREAKFAST SUGAR, HORSE STABLES, RAIN DROPS), compared with compounds that share less common relations such as "H DERIVED FROM M" or "M CAUSES H" (JOB ANXIETY, TAX PRESSURE, THERMAL TORTURE). Compounds which share the top mentioned relations are more likely to be tangible concepts in the real world, and thus require less context and inference than abstract themes. Conclusions from Rambelli et al. (2024) suggest that the linguistic property of concreteness may be, to some extent, responsible for the variation in success of LLMs interpreting compounds. The results here suggest that LLMs are leveraging the properties of the head nouns in order to represent relational information. When taking into account the processing condition of the head and modifier nouns (Section 4.2), the results for the "Separate" condition suggest that the constituent nouns may hold some level of relation information. In conjunction with the results that the head noun representations produce higher correlations, it is plausible that the properties of the head nouns contribute to the models' ability to reflect semantic information. The modifier token representations show slightly positive correlations, where the models are only exposed to the modifier word instead of the full compound. This result may support the CARIN model of conceptual combination which argues that modifiers contain a relational distribution and therefore more frequent modifiers can provide relational information. Models may learn relational distribution information from compounds processed during pre-training.

The results from Section 4.1 suggest that the four autoregressive models extract semantic information towards the middle and later layers, similar to the results that show BERT models encode the relation information in the middle layers. The correlation for LLaMA spikes in the early layers, whilst also producing peaks around the middle layers. After probing the layer-wise representations of LLaMA-2, Liu et al. (2024) found that the lower layers of the model are responsible for extracting lexical semantic information, and higher layers are better suited for predictive tasks. This resonates with the results from the relation category RSA experiment, where LLaMA produced high correlations for both the final head token and final word token representations between layers 6-18.

The decrease in correlation for autoregressive models in the final layers may be explained by Ethayarajh (2019), who concluded that GPT-2, as opposed to BERT, does not represent word meanings in the final layer. The vector space of embeddings appears to flatten, such that semantics, syntax and other linguistic properties of language are not reflected in the token representations extracted from the final layer. As a result, further fine-tuning for specific semantic tasks may be effective when focused on the middle layers of the transformer models.

DeepSeek is of particular interest as this model is a distilled LLM that uses MoE to generate predictions efficiently. The traditional transformer architecture is adapted by replacing feedforward networks (FFNs) with MoE layers (Dai et al., 2024). Each MoE is similar to a FFN in structure, and a number of experts are activated in parallel throughout the transformer when an input is being pro-

cessed. This complex internal mechanism means that there may be subnetworks within the model that each contribute to the model's overall understanding. Contributions from a number of "experts" may result in the semantic relation information being decodable at early points in the input sequence, i.e. from the modifier representations. The distilled DeepSeek-R1 model is based on both LLaMA and Qwen, which could explain the similar correlation patterns for the relation category experiments.

7 Limitations and Future Work

The models included in this paper are all trained on the same scale of parameters (1-3B). Larger models may be explored to investigate the effect of the scale of pre-training, and whether a larger number of hidden dimension enhances or inhibits the extraction of meaningful representations. The datasets are also limited in size, with only 60 possible relation vectors available for probing. Whilst suitable for the size of the models being tested in this paper, investigating LLMs trained on billions of parameters and fine-tuned models would require larger datasets to account for potential noise and model sensitivity. Additionally, expanding the probing experiment to consider novel compounds would explore the generalisability of the findings.

These models were tested using their base configurations in order to explore their intrinsic semantic capabilities. For models such as Phi and LLaMA where little context required for decoding the relation in the probe experiment, fine-tuning could reveal contextual composition where the models can no longer rely on the relation information embedded during training.

Exploring the MoE architecture of DeepSeek could also reveal whether there are particular "experts" that are activated to extract semantic information, and whether these vary across layers or vary according to context. Understanding how DeepSeek dynamically selects experts during the processing of compounds may lead to further insights on subnetworks that exist within the network, and how they contribute to the success of the model interpreting relation information.

8 Conclusion

The main research question concerns whether autoregressive language models consisting of decoder-only layers are able to reflect the semantic relation information of noun-noun compounds, and which parts of the input sequences make the particular relation decodable. The RSA and probing results indicate that the LLMs successfully retrieve semantic information, with meaningful representations extracted after the models have been exposed to the full compound in context. Head noun token representations also reflect information about the thematic relation, which may be the result of the intrinsic concrete properties of the nouns. The modifier nouns show the potential for embedding relational information, however this may be explained by LLMs being exposed to compounds during training. For GPT and DeepSeek, probing reveals that they are actively processing the information stored across the compound in order to accurately predict the appropriate relation, whereas Phi and LLaMA appear to predict the relation just as well from the individual word representations.

References

E. M. Bender and A. 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.

B. Cohen and G. L. Murphy. 1984. Models of concepts. *Cognitive science*, 8(1):27–58.

A. Coil and V. Shwartz. 2023. From chocolate bunny to chocolate crocodile: Do language models understand noun compounds? In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2698–2710, Toronto, Canada. Association for Computational Linguistics.

D. Dai, C. Deng, C. Zhao, R. X. Xu, H. Gao, D. Chen, J. Li, W. Zeng, X. Yu, Y. Wu, et al. 2024. Deepseekmoe: Towards ultimate expert specialization in mixture-of-experts language models. *arXiv preprint arXiv:2401.06066*.

DeepSeek-AI, D. Guo, D. Yang, H. Zhang, and et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

S. Derby, P. Miller, and B. Devereux. 2021. Representation and pre-activation of lexical-semantic knowledge in neural language models. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 211–221. Association for Computational Linguistics.

B. Devereux and F. Costello. 2005. Investigating the relations used in conceptual combination. *Artificial Intelligence Review*, 24:489–515.

J. Devlin, M. Chang, K. Lee, and K. Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

K. Ethayarajh. 2019. How contextual are contextualized word representations? comparing the geometry of BERT, ELMo, and GPT-2 embeddings. *CoRR*, abs/1909.00512.

A. Ettinger. 2020. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

C. L. Gagné. 2001. Relation and lexical priming during the interpretation of noun–noun combinations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 27(1):236.

S. Gunasekar, Y. Zhang, J. Aneja, C. C. T. Mendes, A. Del Giorno, S. Gopi, M. Javaheripi, P. Kauffmann, G. de Rosa, O. Saarikivi, et al. 2023. Textbooks are all you need. *arXiv preprint arXiv:2306.11644*.

J. Hewitt and C. D. Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138. Association for Computational Linguistics.

G. Jawahar, B. Sagot, and D. Seddah. 2019. What does BERT learn about the structure of language? In *ACL 2019-57th Annual Meeting of the Association for Computational Linguistics*, pages 3651—-3657. Association for Computational Linguistics.

T. Ju, W. Sun, W. Du, X. Yuan, Z. Ren, and G. Liu. 2024. How large language models encode context knowledge? a layer-wise probing study. *arXiv preprint arXiv:2402.16061*.

N. Kim, R. Patel, A. Poliak, A. Wang, P. Xia, T. R. McCoy, I. Tenney, A. Ross, T. Linzen, and B. Van Durme. 2019. Probing what different NLP tasks teach machines about function word comprehension. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (SEM 2019)*, pages 235—-249. Association for Computational Linguistics.

N. Kriegeskorte, M. Mur, and P. A. Bandettini. 2008. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2:249.

T. Linzen and M. Baroni. 2021. Syntactic structure from deep learning. *Annual Review of Linguistics*, 7(1):195–212.

Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Z. Liu, C. Kong, Y. Liu, and M. Sun. 2024. Fantastic semantics and where to find them: Investigating which layers of generative llms reflect lexical semantics. *arXiv preprint arXiv:2403.01509*.

J. Mitchell and M. Lapata. 2008. Vector-based models of semantic composition. In *proceedings of ACL-08: HLT*, pages 236–244. Association for Computational Linguistics.

D. Ó Séaghdha and A. Copestake. 2008. Semantic classification with distributional kernels. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 649–656. Association for Computational Linguistics.

M. Ormerod, J. M. del Rincón, and B. Devereux. 2024. How is a "kitchen chair" like a "farm horse"? Exploring the representation of noun-noun compound semantics in transformer-based language models. *Computational Linguistics*, 50(1):49–81.

M. E. Peters, M. Neumann, L. Zettlemoyer, and W.T. Yih. 2018. Dissecting contextual word embeddings: Architecture and representation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1499–1509. Association for Computational Linguistics.

S. T. Piantadosi and F. Hill. 2022. Meaning without reference in large language models. *arXiv preprint arXiv:2208.02957*.

A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI*.

G. Rambelli, E. Chersoni, C. Collacciani, and M. Bolognesi. 2024. Can large language models interpret noun-noun compounds? a linguistically-motivated study on lexicalized and novel compounds. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11823–11835. Association for Computational Linguistics.

V. Shwartz and I. Dagan. 2019. Still a pain in the neck: Evaluating text representations on lexical composition. *Transactions of the Association for Computational Linguistics*, 7:403–419.

E. E. Smith and D. N. Osherson. 1984. Conceptual combination with prototype concepts. *Cognitive science*, 8(4):337–361.

E. E. Smith, D. N. Osherson, L. J. Rips, and M. Keane. 1988. Combining prototypes: A selective modification model. *Cognitive science*, 12(4):485–527.

A. Vaswani. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6000–6010.

I. Vulić, E. M. Ponti, R. Litschko, G. Glavaš, and A. Korhonen. 2020. Probing pretrained language models for lexical semantics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7222–7240. Association for Computational Linguistics.

E. J. Wisniewski. 1997. When concepts combine. *Psychonomic bulletin & review*, 4:167–183.