Probing for Compositionality in LLMs: a literature review

Rémy Marro
Neurocomputation of Language Lab, Tilburg University, Warandelaan 2, 5037AB Tilburg, The Netherlands
r.marro@tilburguniversity.edu

The compositionality principle is considered to be a hallmark of human cognition and is said to be pivotal in human's ability to routinely understand and produce novel linguistic structures (Partee *et al.*, 1995). While Transformer-based language models seemingly appears to emulate compositional language (see Chang & Bergen, 2024 for a state-of-the-art review), the extent to which they encode compositional semantic representations to achieve this remains debated (Pavlick, 2023). In line with the field's goal of "un-blackboxing" neural networks, the probing classifier paradigm has emerged as one of the principal interpretability techniques in NLP studies. The method consists of linking the internal representations of a language model with an identifiable external property of its training sets by training a simple classifier (generally a logistic regressor or a multilayer perceptron) on its intermediate layers (Belinkov, 2022). The researcher usually starts its inquiry by selecting a relevant and annotated dataset containing the linguistic properties under scrutiny (e.g. whether a verb is in the past or in the present tense, as in Conneau *et al.*, 2018). One defines a training and a test set for the language model, trains the language model on the corresponding dataset, and extracts the embeddings of the input sentence at each intermediate layer of the network. The lightweight model will come into play and statistically segregate the embeddings according to the presence or absence of the linguistic representation, as indicated by dataset annotation. The probe will consequently flesh out either as a binary (e.g. the verb is in the present tense or not) or as a multilevel classification task (e.g. the verb is the present simple, present continuous, past simple, or past continuous form). The performance of the classifier, usually expressed in terms of F1 score or accuracy, is meant to indicate whether the scrutinized linguistic properties featured in the presented datasets can be linearly recovered from the embeddings of the network. In short, the probing paradigm proposes to query traces of symbolic computation inherited from linguistic theory among the reputedly continuous vectors of transformer-based models. Scholars have thus used probing methods to identify representations of various linguistic phenomena that serve as building blocks of compositional meaning. These include, among others, syntactic dependencies, thematic role assignment, anaphora resolution, or figures of speech. In this work, I propose to build on this decade-long body of work and draw a state-of-the-art picture of LLMs' compositional competence through the lens of probing classifiers.

However, prior to garnering the insights of probing studies, several key takeaways should be highlighted in order to sketch a careful account of LLMs' compositional competence.

From a conceptual viewpoint, the linguistic representations identified through probing do not give decisive evidence of emergent compositional behaviors (as conceptualized by Fodor & Pylyshyn, 1988). That is, the probing methodology does not give insights into the compositional pathway of transformer-based language models and thus cannot illuminate whether they have actually implemented composition-like algorithms. From a methodological viewpoint, the statistical nature of classifiers makes it impossible to guarantee that the scrutinized representations are functionally relevant to the model and are not fitting irrelevant or incidentally encoded properties (Ravichander *et al.*, 2021). In this light, negative results appear more informative: failure to identify traces of computation in a network's intermediate layers provides a strong indication that the assumed computation was likely not performed by the system. Under such constraints, the computational primitives of the probing tasks fare better than the tasks themselves in enlightening the compositional functions of LLMs. Hence, taken alone, probing studies have limited explanatory power, but when considered collectively, they provide compelling evidence against a specific formal model of LLMs' linguistic competence. Subsequently, I will, through this work, answer the following research question: how do the computational primitives of probing tasks shed light on the compositional functions of transformer-based language models?

To scaffold a comprehensive account of LLMs' compositional competence, I relied on the literature review principle, which proposes to gather all papers tied to a previously stated research question through a systematic search with key terms and inclusion/exclusion criteria. As the field of artificial intelligence is confronted with a vertiginous publication rate and a reliance on conference papers, I decided to query the literature with broad search terms[1] on the search engine Google Scholar. I then relied on inclusion and exclusion criteria to distinguish relevant from irrelevant pieces of work[2]. After extensive pruning, we identified 21 studies (as of Spring 2025) that used probing classifiers to investigate linguistic properties relevant to an inquiry of compositional meaning in LLMs. If studies usually investigated confounding semantic phenomena, each task was classified according to the semantic domain it targeted. To do so, we conceptualize meaning by adopting the structured semantic representations assumed in formal semantics frameworks. We propose to break down our posited compositional competence into four tiers of meaning: lexical semantics, the syntax-semantics interface, propositional semantics, and discourse semantics. Under such a view, each tiers operates at a distinct level of granularity and considers different primitives and combination functions, which form the basic units of our considered formal model of compositional competence. Inspired by contemporary formal account, we adopt a view of meaning that is simultaneously type-driven, model-theoretic, and truth-conditional, but remains semantically under-determined and must be enriched pragmatically (see e.g. Venhuizen *et al.*, 2022). Within our schema, theoretical choices impact the functions and computational primitives considered, thus playing a crucial role in shaping a formal model

[1] Search query was the following: "probing classifier" AND "transformer" AND semantic

[2] These included, among others, model types considered, the exclusion of theoretical paper, publication quality, etc.

of compositional competence. The results are synthesized in Table 1, which organizes the findings by semantic tiers, primitives, composition functions, and performance levels.

Table 1: Synthesis of Probing Results

	Lexical Semantics	**Syntax-Semantics Interface**	**Propositional Semantics**	**Discourse Semantics / Pragmatics**
Primitives	Concepts	Arguments, Predicates, Morphemes, Thematic Roles	Propositions, Truth-values	Implicatures, Illocutionary Force, Informational Status
Composition Functions	Lexical Relations, Cocomposition, Type Coercion	Predication, Semantic Translation	Logical Operations, Syllogisms	Inferences
Good Performance	Entity labeling (Tenney *et al.*; Zhao *et al.*)	Syntactic constituents (Tenney *et al.*; Jawahar *et al.*; Arps *et al.*), Semantic roles (Tenney *et al.*; Wang *et al.*), Predicate properties (Edge probing: Tenney *et al.*)	–	–
Moderate Performance	Lexical relations (Aspillaga *et al.*; Lin & Ng; Chen & Gao)	Tense and inflection (Jawahar *et al.*; Mikhailov *et al.*)	–	Figures of speech (Aghazadeh *et al.*; Schneidermann *et al.*; Klubička *et al.*), Informational status (Li *et al.*; Ju *et al.*)
Fair Performance	–	Abstract morphosyntactic organization (Choenni & Shutova), Grammatical functions (Alt *et al.*), Syntactic relations (Alt *et al.*), Coreference and anaphora (Tenney *et al.*; Saleh *et al.*), Predicate properties (Vertex probing: Chen & Gao)	Logical operators (Lyu *et al.*; Ryb *et al.*; Traylor *et al.*)	Speech acts (Saleh *et al.*; Chen & Gao), Discourse organization (Jawahar *et al.*; Saleh *et al.*)
Poor Performance	–	Event apprehension (Wang *et al.*), Negation (Chen & Gao)	Monotonicity (Chen & Gao), Semantic odd-man-out (Jawahar *et al.*)	Discourse macrostructure (Saleh *et al.*)

Performance levels: ***Good****: F1 > 90, no contrasting results /* ***Moderate****: 70 < F1 < 90, some contrasting results /* ***Fair****: F1 < 70, contrasting results /* ***Poor****: F1 < 60.*

As can be seen in Table 1, much more studies have been carried out in the lower tiers and focused on the syntax-semantics interface. Looking at the performance of the classifiers, a gradual drop of accuracy can be observed as probes target increasingly abstract linguistic constructs, particularly those involving larger constituents and higher-order composition, like propositions or discourse movements. Most notably, pinpointing the influence of a single word or a group of words on the logical structure of a proposition remains difficult for LLMs. By contrast, representations of concepts and basic relations appear to be consistently well

encoded. Compared to earlier neural network architectures, transformer-based models seem to represent relational information more robustly, as demonstrated by the higher prevalence of lexical and dependency-based information at the predicate level. On the other hand, a simple concatenation of static vectors can encode pragmatic features to a degree comparable to, if not higher than, transformer models. This suggests that the purported computational primitives of my formal models fail to give a comprehensive account of the seemingly compositional behavior of LLMs: while probing studies do not contradict the presence of word-level semantic information in LLMs, the evidence for propositional or discourse-level representations remains inconclusive at best. These results therefore fail to provide support for the existence of higher-order composition functions like pragmatic inferencing or logical reasoning on propositions.

What emerges from this review is an ambivalent picture of LLMs' compositional competence: they appear to encode lexical and local relational information, yet fall short of representing propositional and discourse-level structures. In that respect, transformer models do not appear to encode traces of compositional computation more robustly than other distributional models of semantics (Lenci *et al.*, 2022; Li *et al.*, 2023). This dissonance between the seemingly compositional behaviors of LLMs and the representational vacuum at the proposition and discourse-level could be interpreted (1) as evidence that LLMs bears no compositional competence as posited by the considered formal framework, or (2) that alternate pathway other than symbolic computation have emerged from transformer language model to attain composition-like behaviors. Albeit conceptually and methodologically limited in its outreach, the probing paradigm remains useful when combined with a rigorous theoretical basis. Hence, classifiers should either serve as a functional diagnosis of LLMs' limitations – as a method to systematically eliminate implausible composition functions – or as a post-hoc representational check part of a larger mechanistic schema of LLMs' competence (in the vein of Geiger *et al.*, 2025). Another promising perspective lies in combining probing with model-centered theories of meaning, considering LLMs as functionally distinct architectures shaped by statistical learning, and revealing their computational primitives in an ontology-free fashion (e.g. Cunningham *et al.*, 2023; Michael *et al.*, 2020).

References

AGHAZADEH E., FAYYAZ M. & YAGHOOBZADEH Y. (2022). Metaphors in Pre-Trained Language Models: Probing and Generalization Across Datasets and Languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, p. 2037–2050, Dublin, Ireland: Association for Computational Linguistics. DOI : 10.18653/v1/2022.acl-long.144.

ALT C., GABRYSZAK A. & HENNIG L. (2020). Probing Linguistic Features of Sentence-Level Representations in Relation Extraction. In *Proceedings of the 58th Annual Meeting*

of the Association for Computational Linguistics, p. 1534–1545, Online: Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.140.

ARPS D., SAMIH Y., KALLMEYER L. & SAJJAD H. (2022). Probing for Constituency Structure in Neural Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, p. 6738–6757, Abu Dhabi, United Arab Emirates: Association for Computational Linguistics. DOI : 10.18653/v1/2022.findings-emnlp.502.

ASPILLAGA C., MENDOZA M. & SOTO A. (2021). Inspecting the concept knowledge graph encoded by modern language models. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, p. 2984–3000, Online: Association for Computational Linguistics. DOI : 10.18653/v1/2021.findings-acl.263.

BELINKOV Y. (2022). Probing Classifiers: Promises, Shortcomings, and Advances. *Computational Linguistics*, **48**(1), 207–219. DOI : 10.1162/coli_a_00422.

CHANG T. A. & BERGEN B. K. (2024). Language model behavior: A comprehensive survey. *Computational Linguistics*, **50**(1), 293–350.

CHEN Z. & GAO Q. (2022). Probing Linguistic Information for Logical Inference in Pre-trained Language Models. *Proceedings of the AAAI Conference on Artificial Intelligence*, **36**(10), 10509–10517. DOI : 10.1609/aaai.v36i10.21294.

CHOENNI R. & SHUTOVA E. (2022). Investigating Language Relationships in Multilingual Sentence Encoders Through the Lens of Linguistic Typology. *Computational Linguistics*, **48**(3), 635–672. DOI : 10.1162/coli_a_00444.

CONNEAU A., KRUSZEWSKI G., LAMPLE G., BARRAULT L. & BARONI M. (2018). What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In I. GUREVYCH & Y. MIYAO, Éds., *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, p. 2126–2136, Melbourne, Australia: Association for Computational Linguistics. DOI : 10.18653/v1/P18-1198.

CUNNINGHAM H., EWART A., RIGGS L., HUBEN R. & SHARKEY L. (2023). Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*.

FODOR J. A. & PYLYSHYN Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, **28**(1-2), 3–71.

GEIGER A., IBELING D., ZUR A., CHAUDHARY M., CHAUHAN S., HUANG J., ARORA A., WU Z., GOODMAN N., POTTS C. & ICARD T. (2025). Causal abstraction: A theoretical foundation for mechanistic interpretability. *Journal of Machine Learning Research*, **26**(83), 1–64.

JAWAHAR G., SAGOT B. & SEDDAH D. (2019). What Does BERT Learn about the Structure of Language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, p. 3651–3657, Florence, Italy: Association for Computational Linguistics. DOI : 10.18653/v1/P19-1356.

JU T., SUN W., DU W., YUAN X., REN Z. & LIU G. (2024). How large language models encode context knowledge? a layer-wise probing study. In N. CALZOLARI, M.-Y. KAN, V. HOSTE, A. LENCI, S. SAKTI & N. XUE, Éds., *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, p. 8235–8246, Torino, Italia: ELRA and ICCL.

KLUBIČKA F., NEDUMPOZHIMANA V. & KELLEHER J. (2023). Idioms, Probing and Dangerous Things: Towards Structural Probing for Idiomaticity in Vector Space. In *Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023)*, p. 45–57, Dubrovnik, Croatia: Association for Computational Linguistics. DOI : 10.18653/v1/2023.mwe-1.8.

LENCI A., SAHLGREN M., JEUNIAUX P., CUBA GYLLENSTEN A. & MILIANI M. (2022). A comparative evaluation and analysis of three generations of distributional semantic models. *Language resources and evaluation*, **56**(4), 1269–1313.

LI B. Z., NYE M. & ANDREAS J. (2021). Implicit Representations of Meaning in Neural Language Models. 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)*, p. 1813–1827, Online: Association for Computational Linguistics. DOI : 10.18653/v1/2021.acl-long.143.

LI R., ZHAO X. & MOENS M.-F. (2023). A Brief Overview of Universal Sentence Representation Methods: A Linguistic View. *ACM Computing Surveys*, **55**(3), 1–42. DOI : 10.1145/3482853.

LIN R. & NG H. T. (2022). Does BERT Know that the IS-A Relation Is Transitive? In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, p. 94–99, Dublin, Ireland: Association for Computational Linguistics. DOI : 10.18653/v1/2022.acl-short.11.

LYU Q., HUA Z., LI D., ZHANG L., APIDIANAKI M. & CALLISON-BURCH C. (2022). Is "My Favorite New Movie" My Favorite Movie? Probing the Understanding of Recursive Noun Phrases. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, p. 5286–5302, Seattle, United States: Association for Computational Linguistics. DOI : 10.18653/v1/2022.naacl-main.388.

MICHAEL J., BOTHA J. A. & TENNEY I. (2020). Asking without telling: Exploring latent ontologies in contextual representations. In B. WEBBER, T. COHN, Y. HE & Y. LIU, Éds., *Proceedings of the 2020 Conference on Empirical Methods in Natural Language*

Processing (EMNLP), p. 6792–6812, Online: Association for Computational Linguistics. DOI : 10.18653/v1/2020.emnlp-main.552.

MIKHAILOV V., SERIKOV O. & ARTEMOVA E. (2021). Morph Call: Probing Morphosyntactic Content of Multilingual Transformers. In *Proceedings of the Third Workshop on Computational Typology and Multilingual NLP*, p. 97–121, Online: Association for Computational Linguistics. DOI : 10.18653/v1/2021.sigtyp-1.10.

PARTEE B. *et al.* (1995). Lexical semantics and compositionality. *An invitation to cognitive science: Language*, **1**, 311–360.

PAVLICK E. (2023). Symbols and grounding in large language models. *Philosophical Transactions of the Royal Society A*, **381**(2251), 20220041.

RAVICHANDER A., BELINKOV Y. & HOVY E. (2021). Probing the Probing Paradigm: Does Probing Accuracy Entail Task Relevance? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, p. 3363–3377, Online: Association for Computational Linguistics. DOI : 10.18653/v1/2021.eacl-main.295.

RYB S., GIULIANELLI M., SINCLAIR A. & FERNÁNDEZ R. (2022). AnaLog: Testing Analytical and Deductive Logic Learnability in Language Models. In *Proceedings of the 11th Joint Conference on Lexical and Computational Semantics*, p. 55–68, Seattle, Washington: Association for Computational Linguistics. DOI : 10.18653/v1/2022.starsem-1.5.

SALEH A., DEUTSCH T., CASPER S., BELINKOV Y. & SHIEBER S. (2020). Probing Neural Dialog Models for Conversational Understanding. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, p. 132–143, Online: Association for Computational Linguistics. DOI : 10.18653/v1/2020.nlp4convai-1.15.

SCHNEIDERMANN N., HERSHCOVICH D. & PEDERSEN B. (2023). Probing for Hyperbole in Pre-Trained Language Models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, p. 200–211, Toronto, Canada: Association for Computational Linguistics. DOI : 10.18653/v1/2023.acl-srw.30.

TENNEY I., XIA P., CHEN B., WANG A., POLIAK A., MCCOY R. T., KIM N., VAN DURME B., BOWMAN S. R., DAS D. & PAVLICK E. (2019). What do you learn from context? Probing for sentence structure in contextualized word representations. Version Number: 1, DOI : 10.48550/ARXIV.1905.06316.

TRAYLOR A., FEIMAN R. & PAVLICK E. (2021). AND does not mean OR: Using Formal Languages to Study Language Models' Representations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint*

Conference on Natural Language Processing (Volume 2: Short Papers), p. 158–167, Online: Association for Computational Linguistics. DOI : 10.18653/v1/2021.acl-short.21.

VENHUIZEN N. J., HENDRIKS P., CROCKER M. W. & BROUWER H. (2022). Distributional formal semantics. *Information and Computation*, **287**, 104763.

WANG B., DU X. & CARDIE C. (2023). Probing Representations for Document-level Event Extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, p. 12675–12683, Singapore: Association for Computational Linguistics. DOI : 10.18653/v1/2023.findings-emnlp.844.

ZHAO M., DUFTER P., YAGHOOBZADEH Y. & SCHÜTZE H. (2020). Quantifying the Contextualization of Word Representations with Semantic Class Probing. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, p. 1219–1234, Online: Association for Computational Linguistics. DOI : 10.18653/v1/2020.findings-emnlp.109.