UNIVERSITY OF HELSINKI

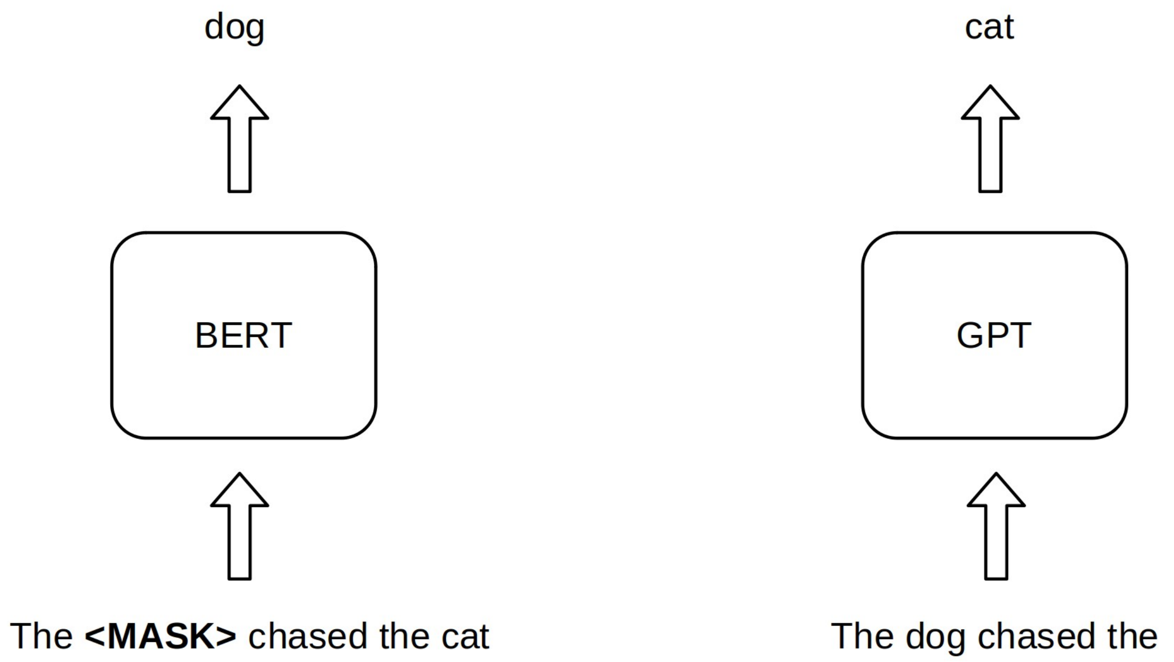
Semantics in Large Language Models

Tommi Buder-Gröndahl

How do LLMs work?

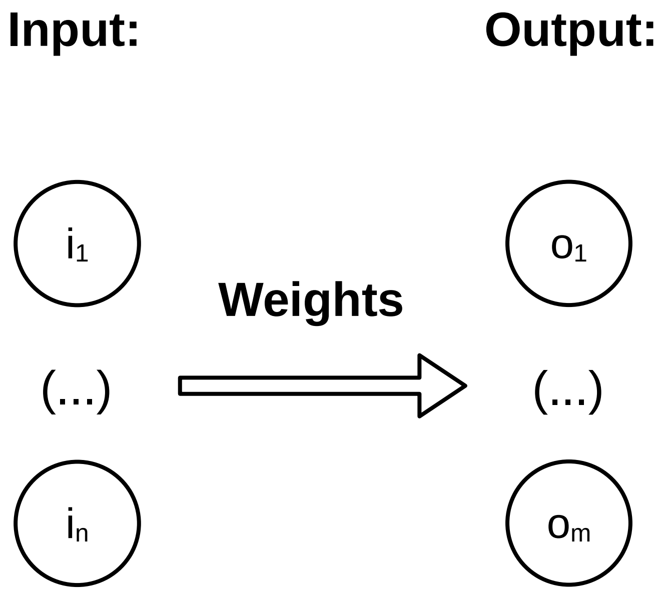
Large Language Models (LLMs)

- Large *deep neural networks* (DNNs), currently mostly *Transformers* (Vaswani et al. 2017)
- *Pre-trained* on generic linguistic tasks: e.g. predicting masked words or upcoming text
- Can be *fine-tuned* to more specific tasks on smaller training sets (*transfer learning*)
- Recently more emphasis on using pre-trained LLMs without fine-tuning (via prompting)



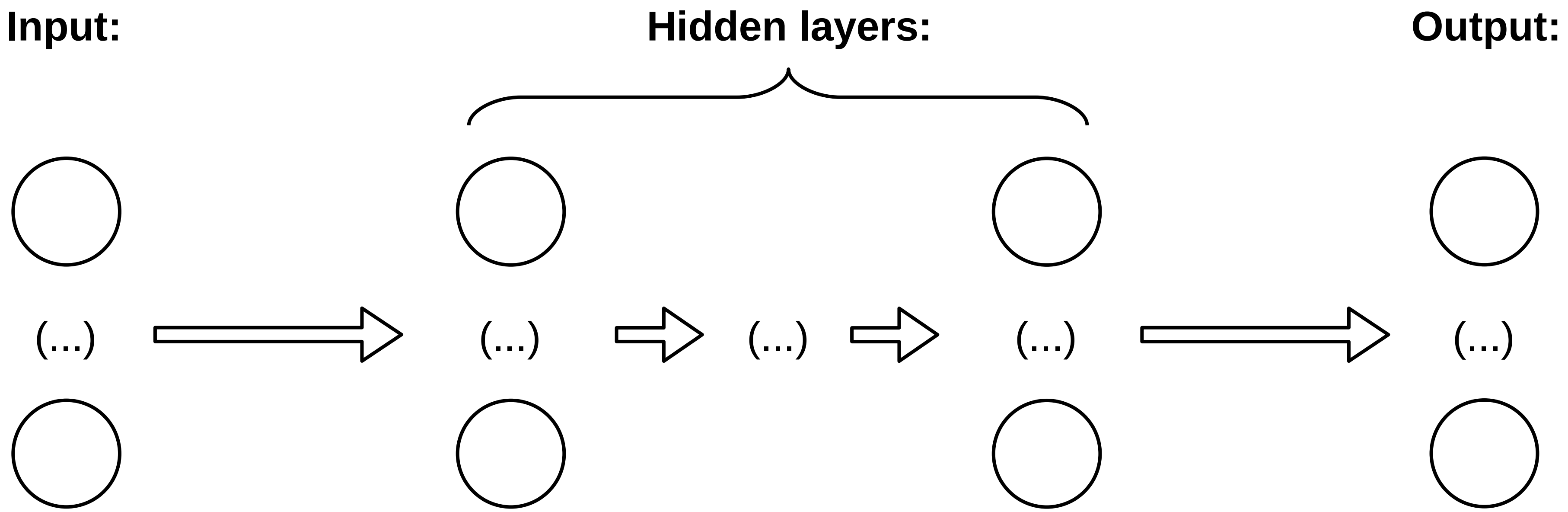
Neural network

Input vector:
i = [i_1, …, i_n]

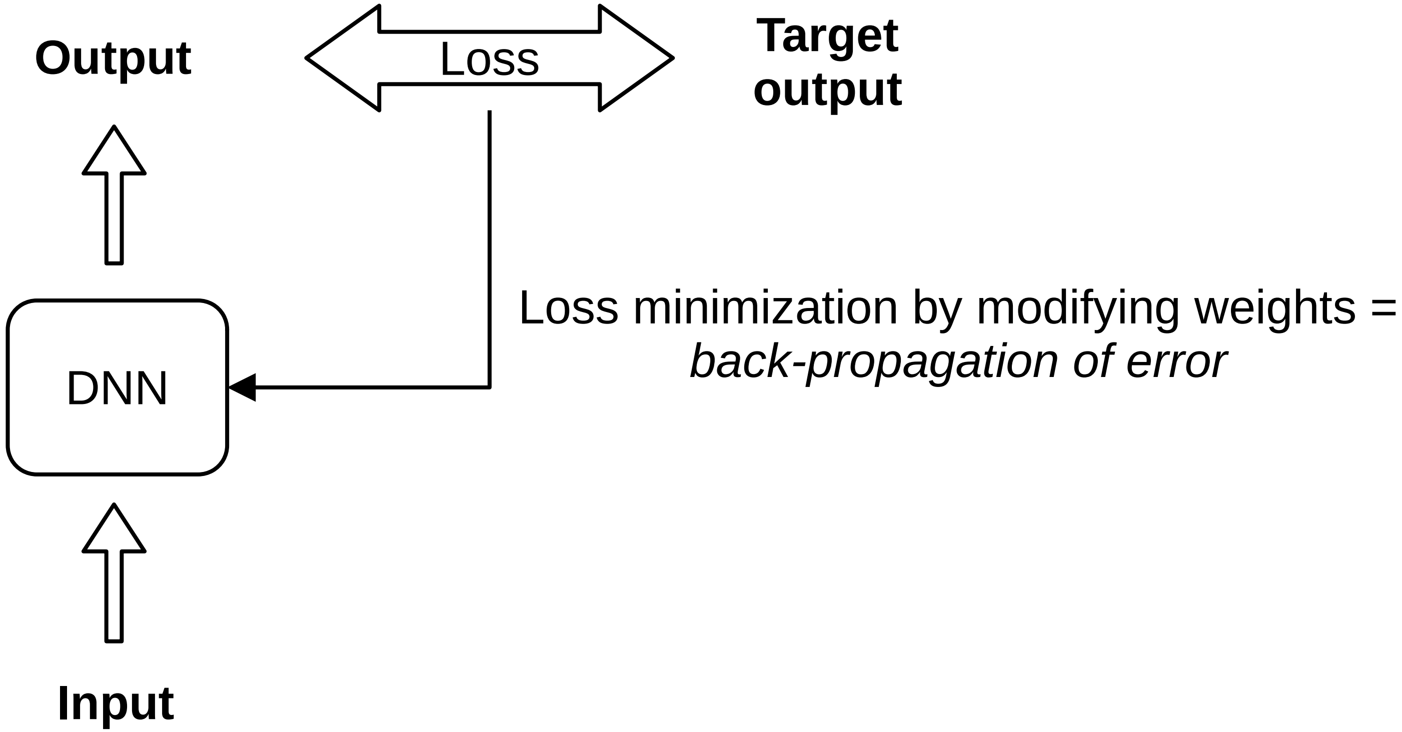


Output vector:
o = [o_1, …, o_m]

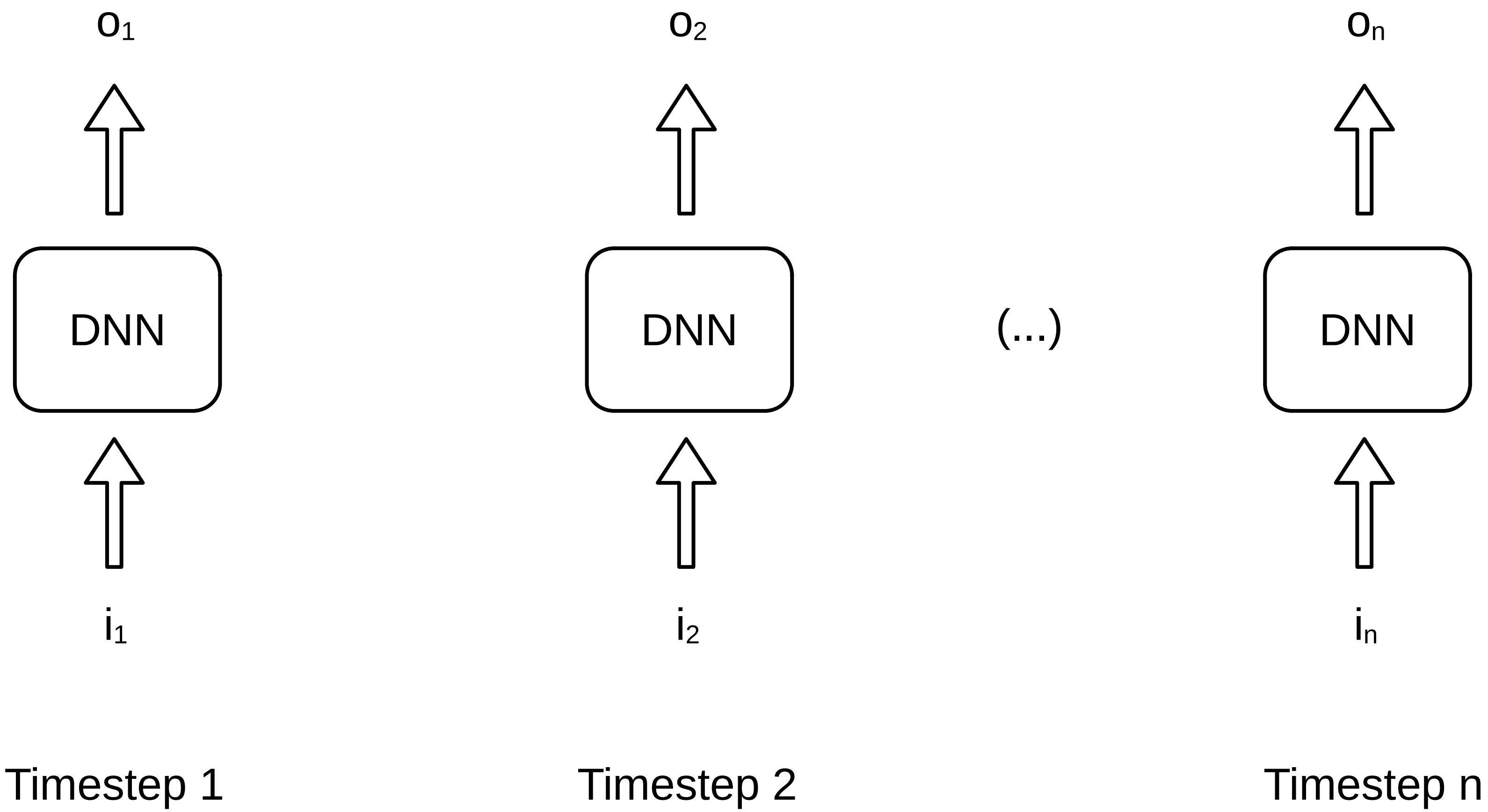
Deep Neural Network (DNN)



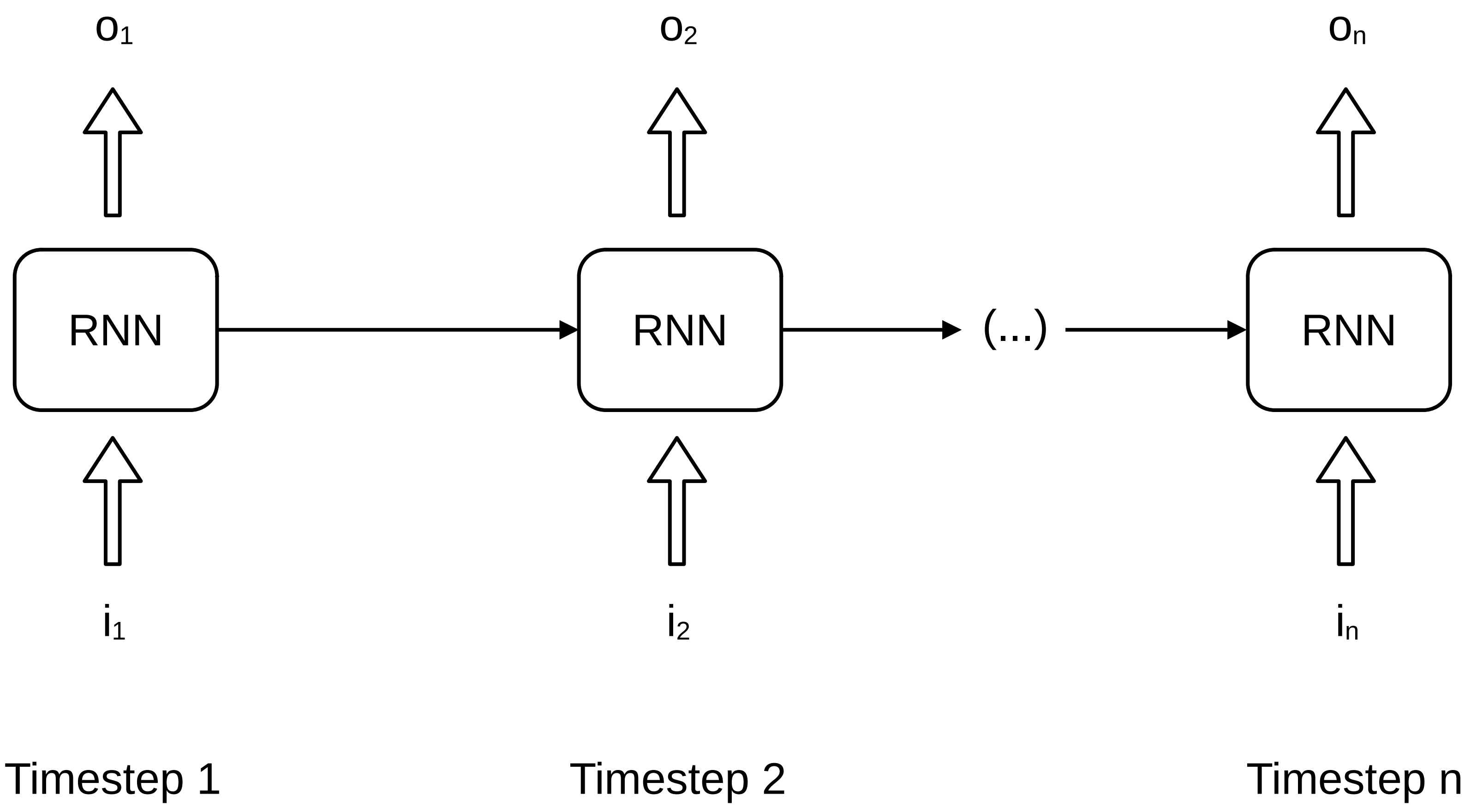
Training a DNN



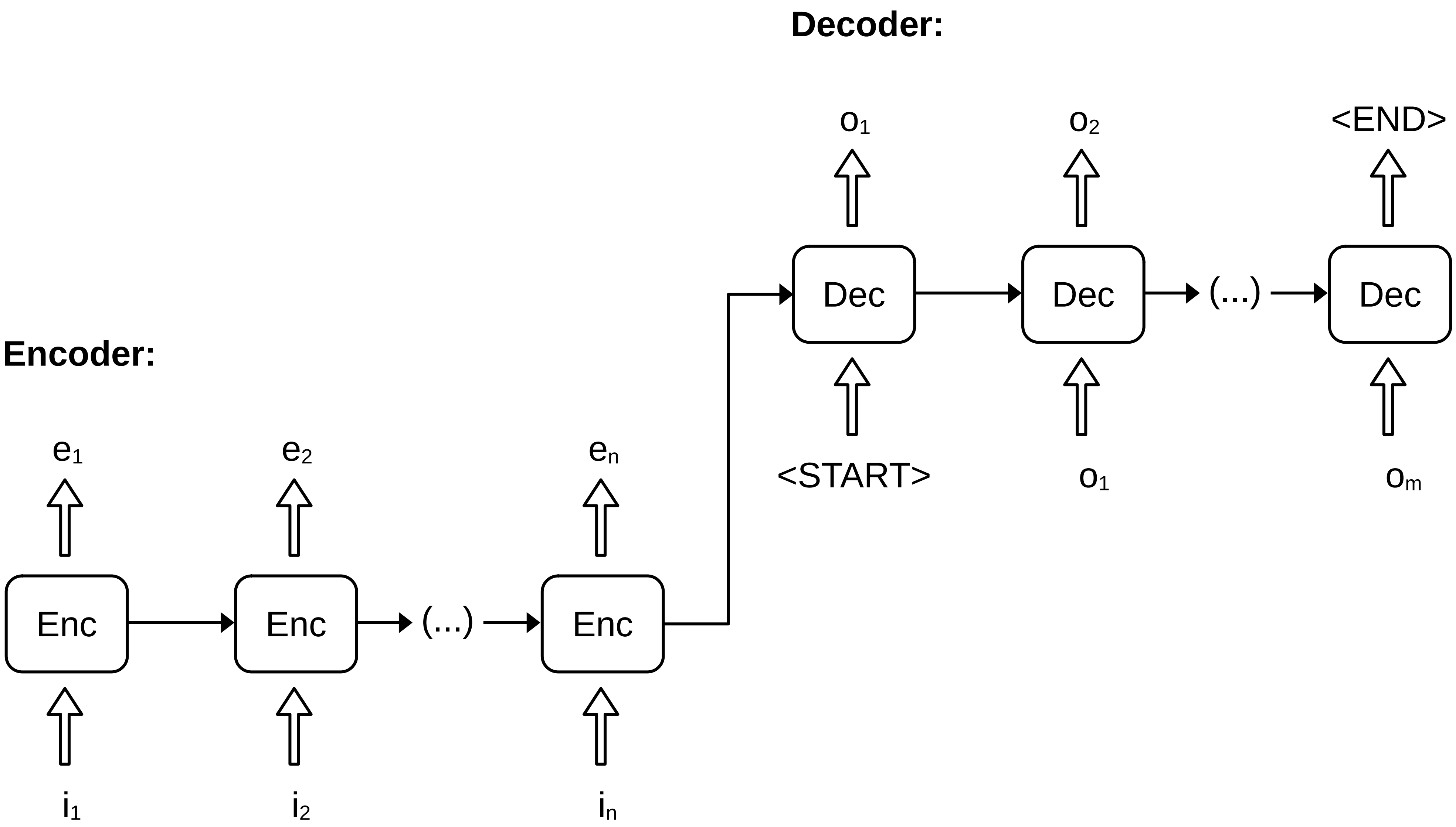
Sequential data



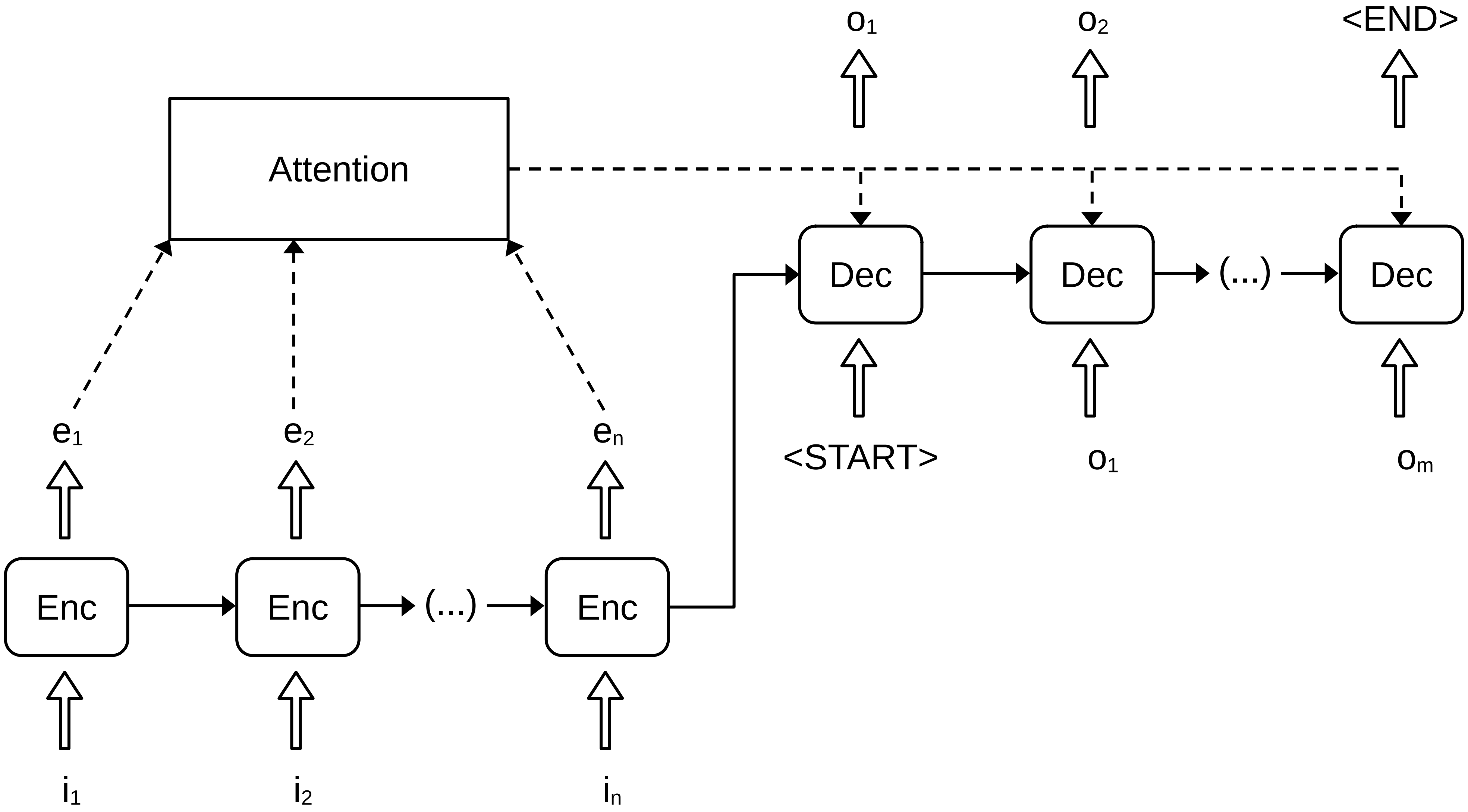
Recurrent Neural Network (RNN)



Encoder-decoder RNN



Encoder-decoder RNN + Attention



Transformer

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com

Noam Shazeer* Google Brain noam@google.com

Niki Parmar* Google Research nikip@google.com

Jakob Uszkoreit* Google Research usz@google.com

Llion Jones* Google Research llion@google.com

Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu

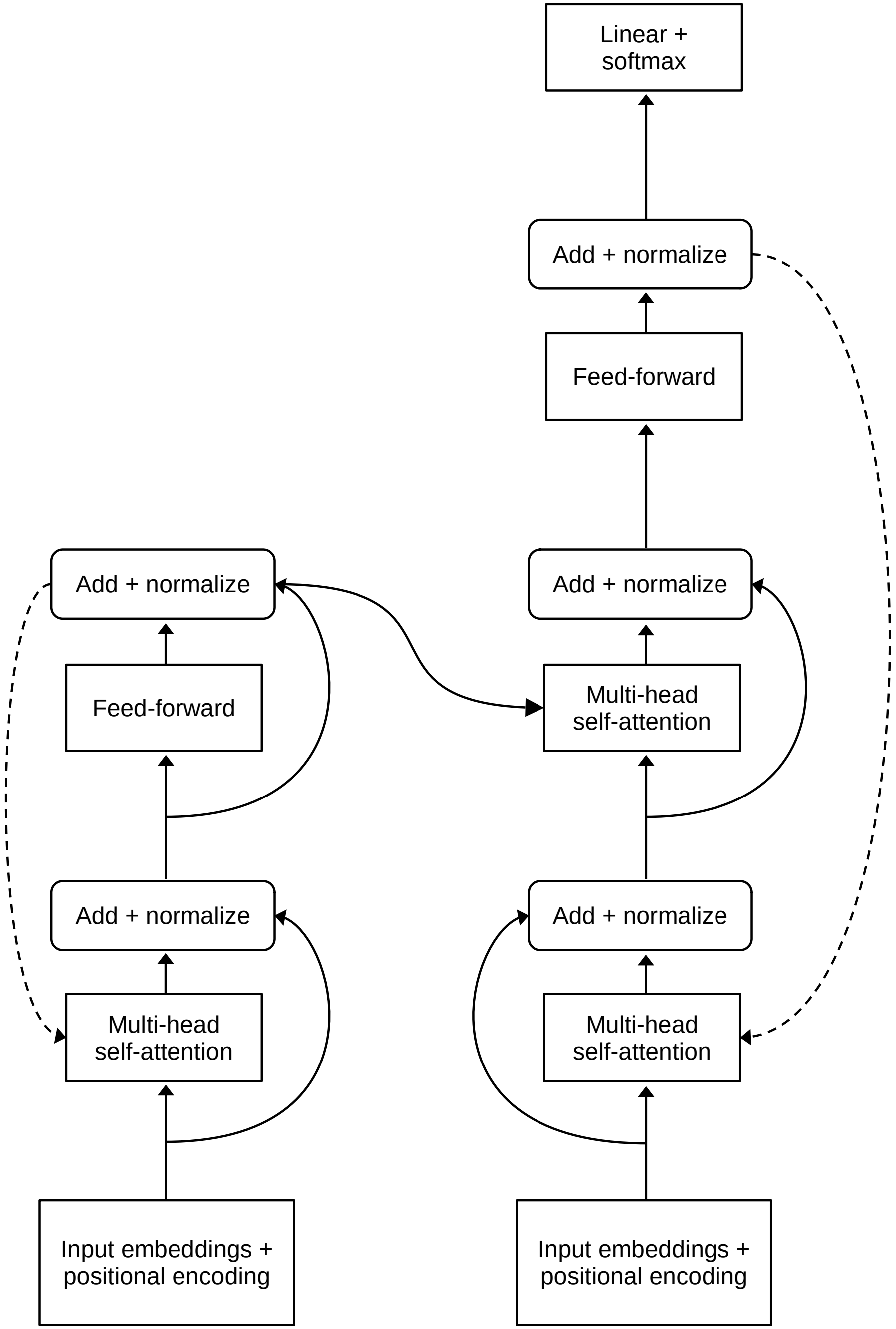
Łukasz Kaiser* Google Brain lukaszkaiser@google.com

Illia Polosukhin* ‡ illia.polosukhin@gmail.com

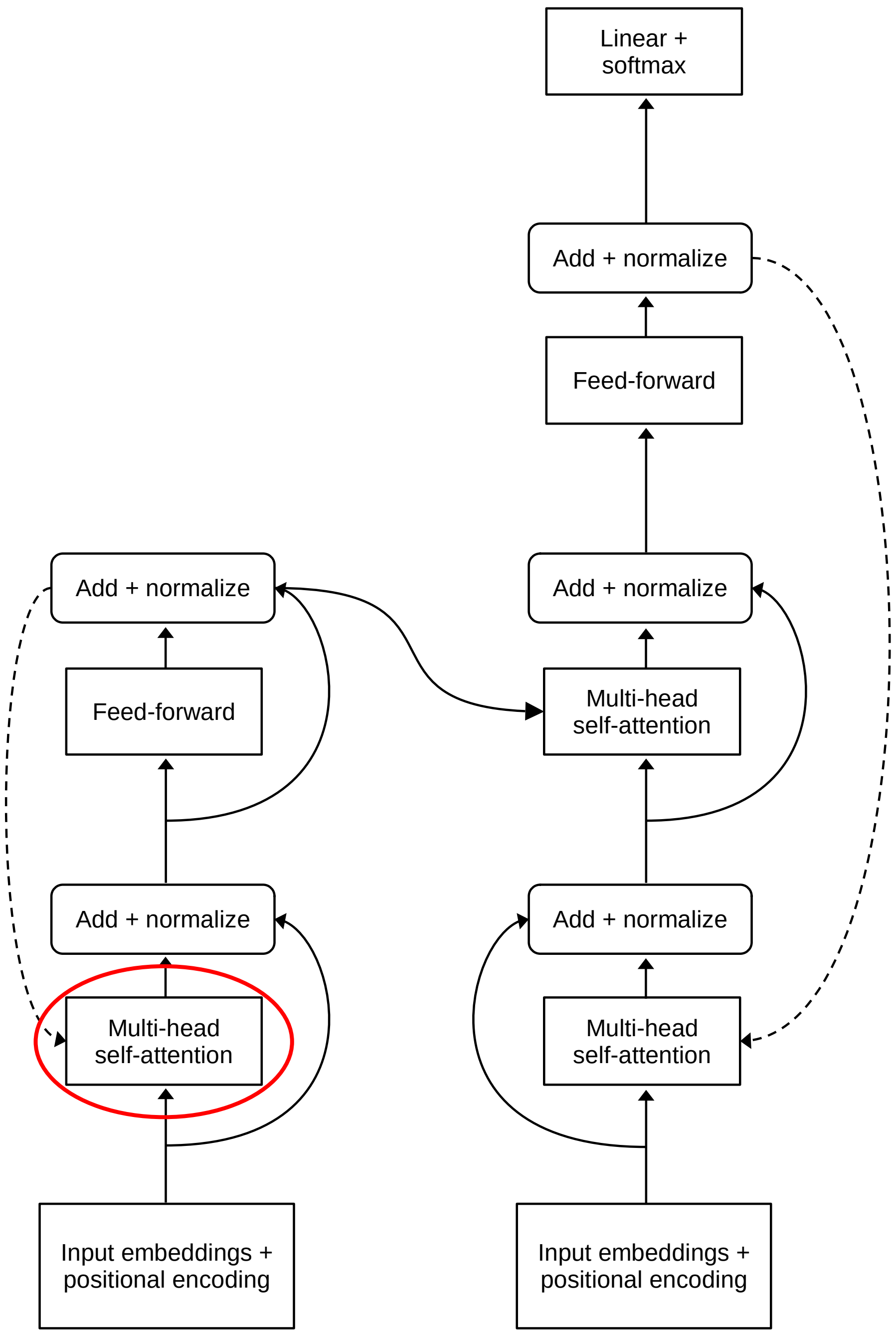
Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.0 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature.

Transformer

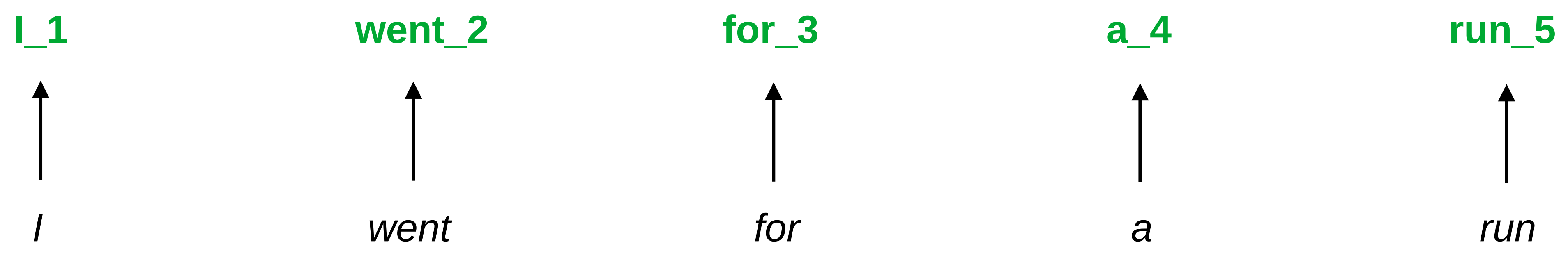


Transformer



Transformer

Each input word has an **embedding**, which is combined with **positional encoding**.

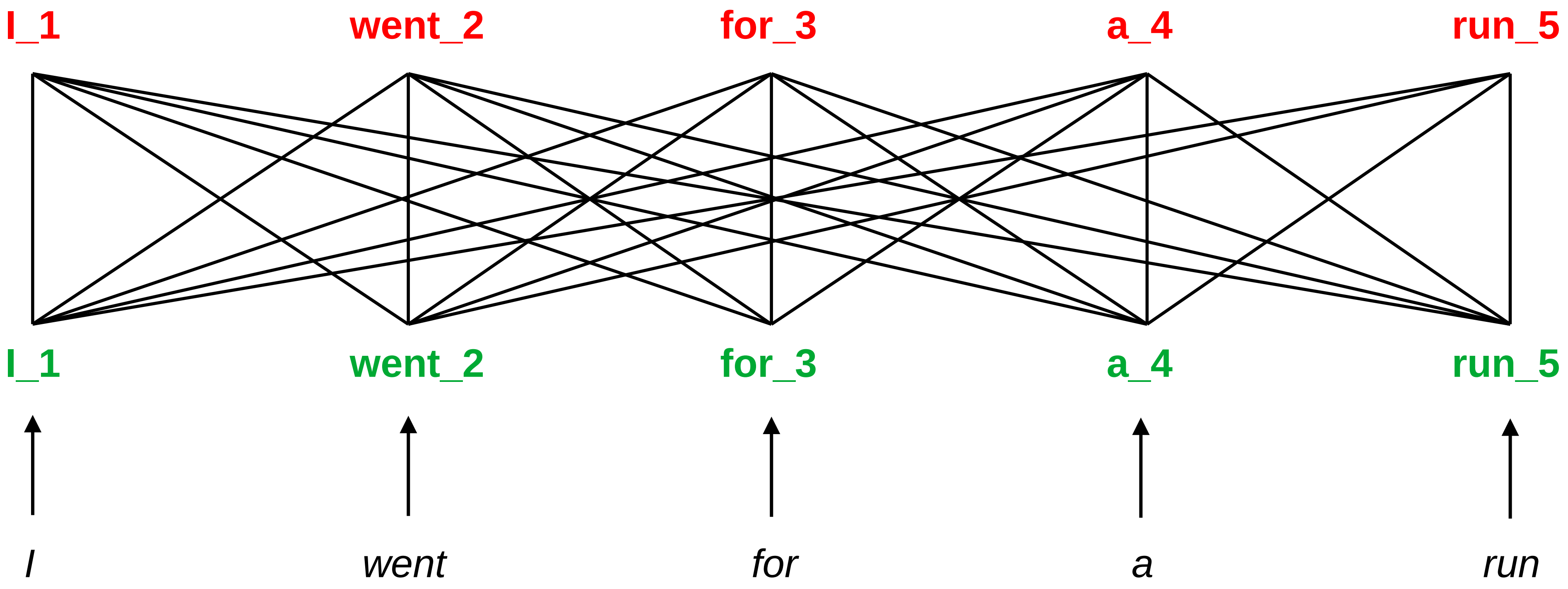


Transformer

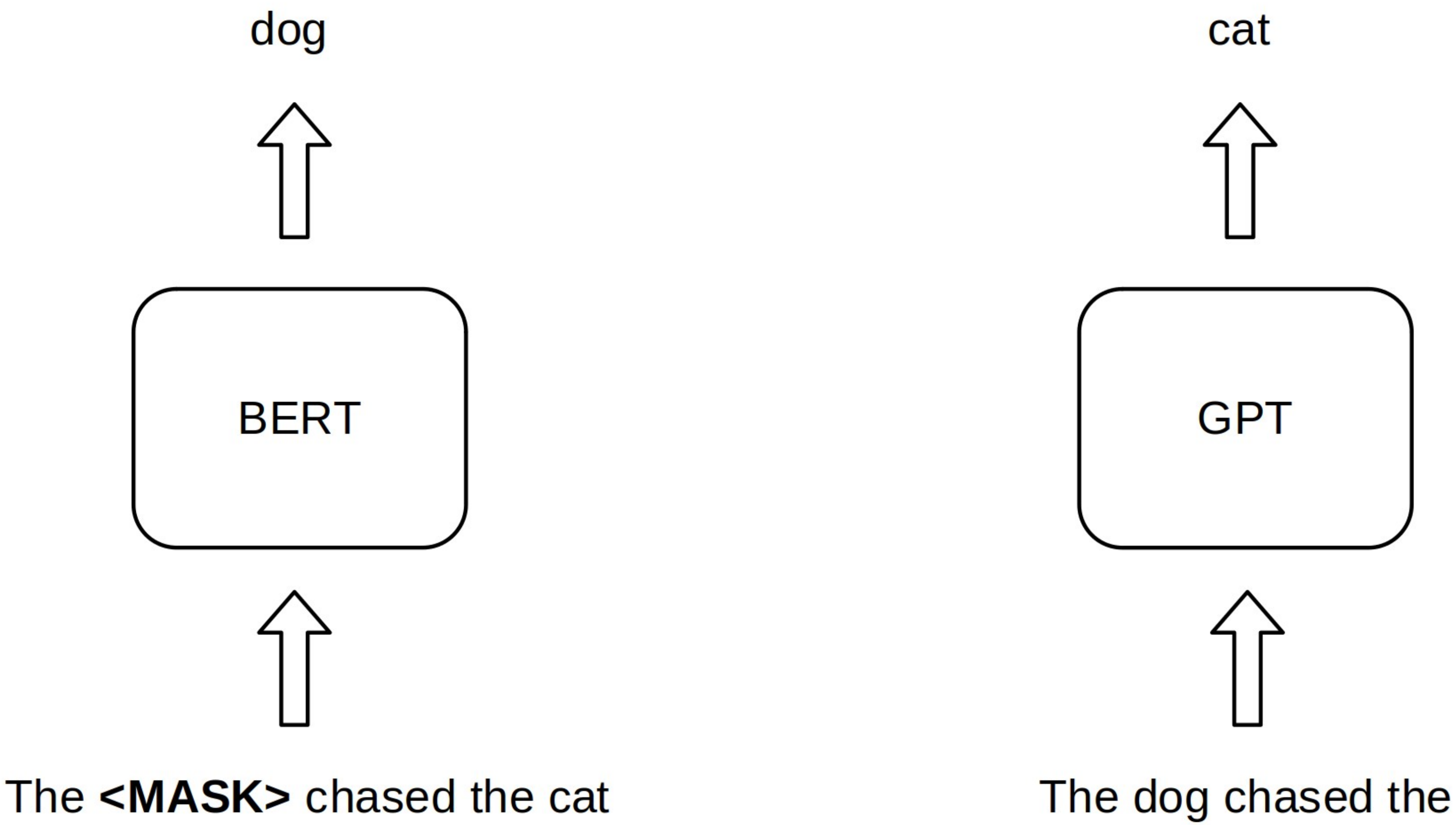
Each input word has an **embedding**, which is combined with **positional encoding**.

Input goes through **multi-head self-attention**, creating new **contextual encodings** for each token.

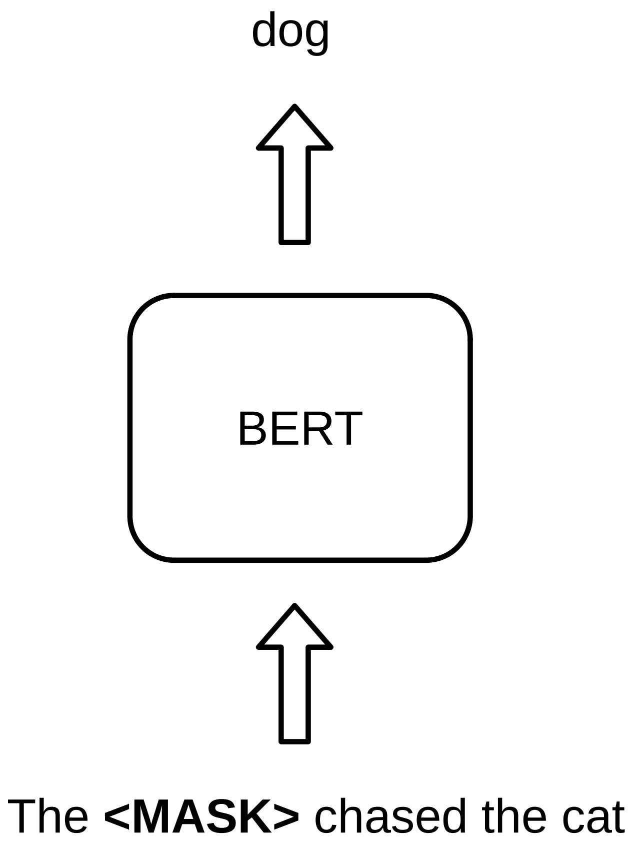
Contextual encoding for each token is calculated from previous embeddings of each token.



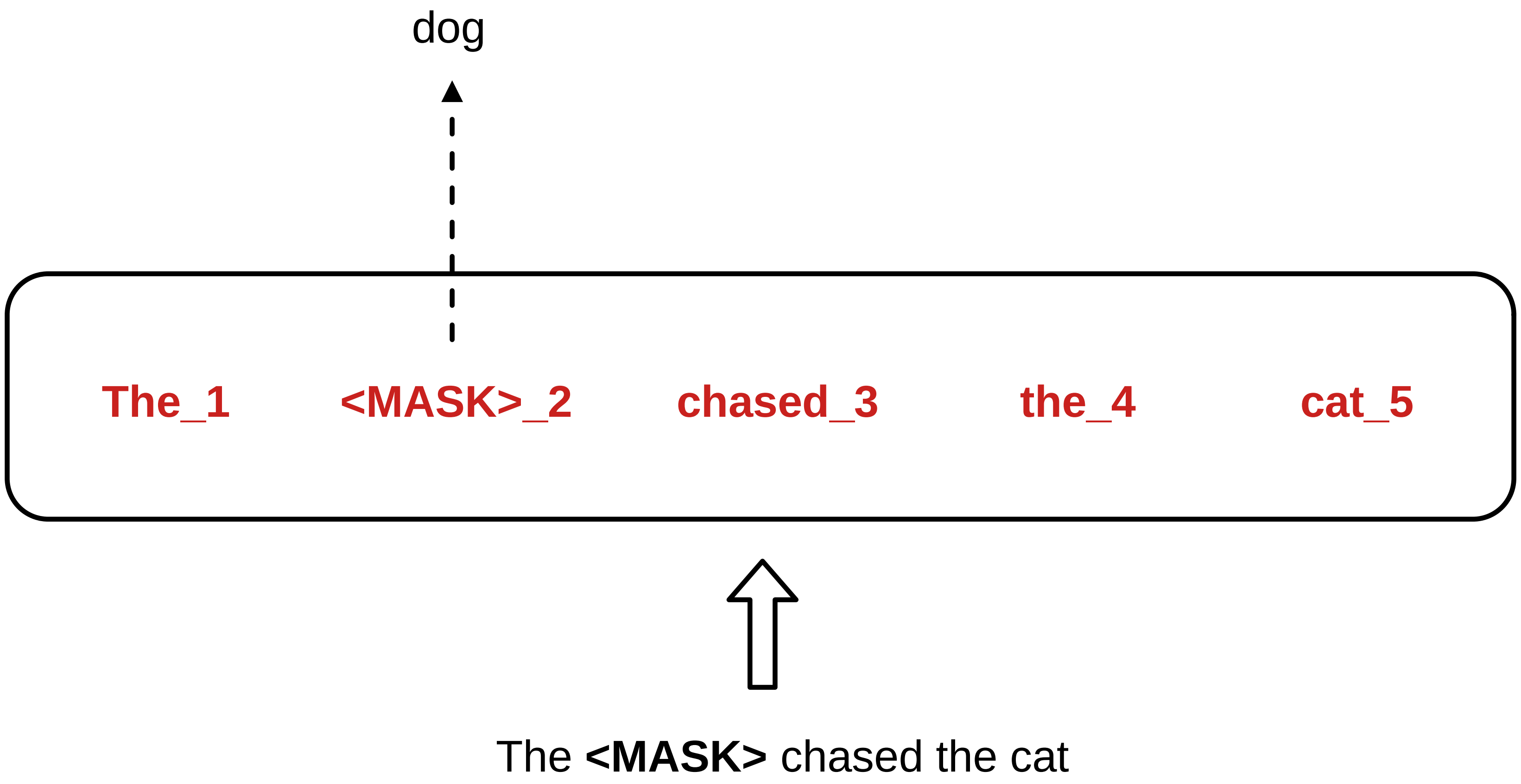
BERT and GPT



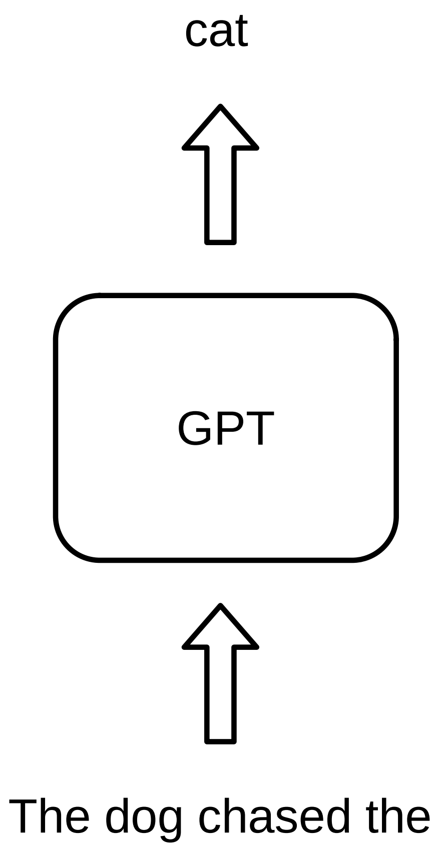
BERT: predicting masked tokens



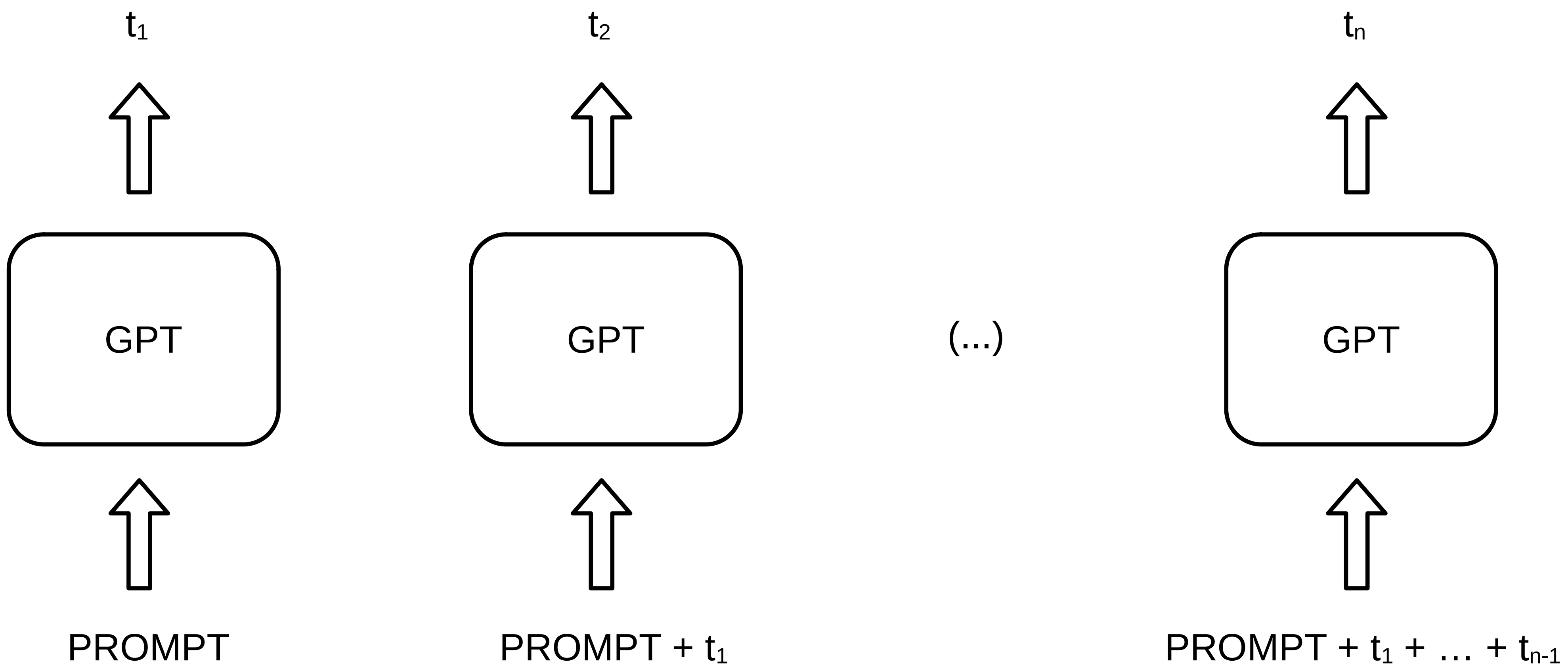
BERT: predicting masked tokens



GPT: predicting the next token



GPT: predicting the next token



Semantic representations in LLMs

Methods of studying LLMs

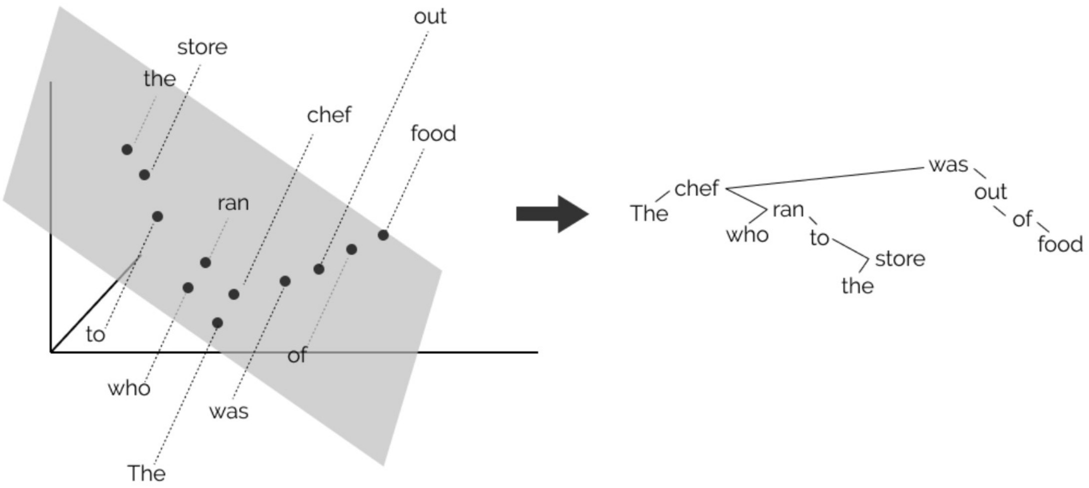
Behavioral

- Fine-tuning for specific tasks, measuring performance
- Prompting pre-trained models directly

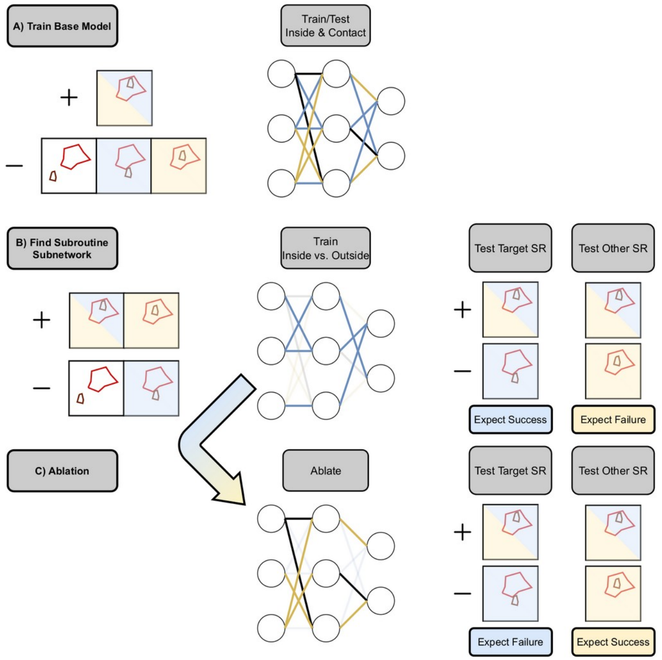
Methods of studying LLMs

Looking inside LLMs

- *Probing:* mapping activation patterns to linguistic/semantic labels
- *Mechanistic interpretation:* opening up the computational pipeline



https://nlp.stanford.edu/~johnhew/structural-probe.html



(Lepori et al. 2023)

Survey/commentary papers

Tyler A. Chang and Benjamin K. Bergen. 2024. Language Model Behavior: A Comprehensive Survey. *Computational Linguistics* 50 (1): 293–350.

Ellie Pavlick. 2023. Symbols and grounding in large language models. *Philosophical Transactions of the Royal Society* 381.

Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. 2024. Explainability for Large Language Models: A Survey. *ACM Transactions on Intelligent Systems and Technology*, 15(2): 1–38.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8: 842–866.

Lexical and pragmatic

Lexical semantics

- Predicting hypernyms in templates: “A robin is a _” (Hanna and Mareček 2021)
- Probability correlations between synonyms/co-hyponyms (Arefyev et al. 2020)
- Similes and analogies (Liu et al. 2022, He et al. 2022, Ushio et al. 2021)
- Challenges with e.g. names and numbers (Wallace et al. 2019, Balasubramanian et al. 2020)

Lexical and pragmatic

Lexical semantics

- Predicting hypernyms in templates: “A robin is a _” (Hanna and Mareček 2021)
- Probability correlations between synonyms/co-hyponyms (Arefyev et al. 2020)
- Similes and analogies (Liu et al. 2022, He et al. 2022, Ushio et al. 2021)
- Challenges with e.g. names and numbers (Wallace et al. 2019, Balasubramanian et al. 2020)

Argument structure

- Thematic roles partly reconstructable via probing (Tenney et al. 2019)
- Sensitivity to verb classes (e.g. causatives) (Davis & van Schijndel 2020)
- Layer separation: lexical → syntactic → thematic (Tenney et al. 2019, Manning et al. 2020)
- Theoretical formalism influences results (Kuznetsov & Gurevych 2020, Kulmizev et al. 2020)

Lexical and pragmatic

Discourse & pragmatics

- Tracking entities across a discourse (e.g. characters in a story) (Schuster and Linzen 2022)
- Sensitivity to the literal-figurative distinction (Pedinotti et al. 2021; Griciūtė et al. 2022)

Lexical and pragmatic

Discourse & pragmatics

- Tracking entities across a discourse (e.g. characters in a story) (Schuster and Linzen 2022)
- Sensitivity to the literal-figurative distinction (Pedinotti et al. 2021; Griciūtė et al. 2022)

World-knowledge

- Success at certain pragmatic reasoning tasks (Petroni et al. 2019, Jiang et al. 2020)
- Some knowledge of physical properties (Apidianaki and Garí Soler 2021, Shi and Wolff 2021)
- Inferring knowledge/desires of characters (Summers-Stay et al. 2021, Sap et al. 2022)
- Explaining behavior of characters in common-sense terms (Lal et al. 2022)
- Improvement with model size (Sahu et al. 2022, Kalo & Fichtel 2022)
- Reliance on simple heuristics (Poerner et al. 2019, Lin et al. 2020, Cao et al. 2021)

Lexical and pragmatic: summary

LLMs have vast amounts of distributional information about words, hierarchically organized to different levels of abstraction: lexical relations, argument structure, discourse.

Much of “world-knowledge” in LLMs is memorized and relies on superficial heuristics.

Formal semantics

Compositionality

- Challenges with systematic reasoning (Hupkes et al. 2020, Kassner et al. 2020)
- LLMs can be prompted to produce semantic parses (Qiu et al. 2022, Hosseini et al. 2022)
- Partial dissociability/modularity of representations (Lovering & Pavlick 2022)
- Candidates for LLM-internal symbolic processes (Geva et al. 2021, Olsson et al. 2022)
- Competence vs. performance? (Pavlick 2023)

Formal semantics

Compositionality

- Challenges with systematic reasoning (Hupkes et al. 2020, Kassner et al. 2020)
- LLMs can be prompted to produce semantic parses (Qiu et al. 2022, Hosseini et al. 2022)
- Partial dissociability/modularity of representations (Lovering & Pavlick 2022)
- Candidates for LLM-internal symbolic processes (Geva et al. 2021, Olsson et al. 2022)
- Competence vs. performance? (Pavlick 2023)

Logic

- “Chain-of-thought-prompting”: asking LLMs to reason step-by-step (Kojima et al. 2022)
- Multi-step reasoning is hard (Forbes et al. 2019, Kassner et al. 2020, Saparov and He 2023)
- Notable troubles with negation

Formal semantics

Negation

- Ignoring negation: “A robin is [not] _” (Ettinger 2020, Kassner and Schütze 2020)
- Reasoning is more difficult with negated prompts (Jang et al. 2022)
- Performance of fine-tuned LLMs deteriorates significantly with negation-focused datasets (Hossain et al. 2020, Geiger et al. 2020, Tejada et al. 2021, Truong et al. 2022)
- Performance on negated prompts *descreases* as models *increase* in size (Jang et al. 2022)

Formal semantics: summary

Compositionality in LLMs is a topic of contention; algorithmic/mechanistic interpretation needed.

LLMs have some level of logical capacity, but struggle with complex inferences.

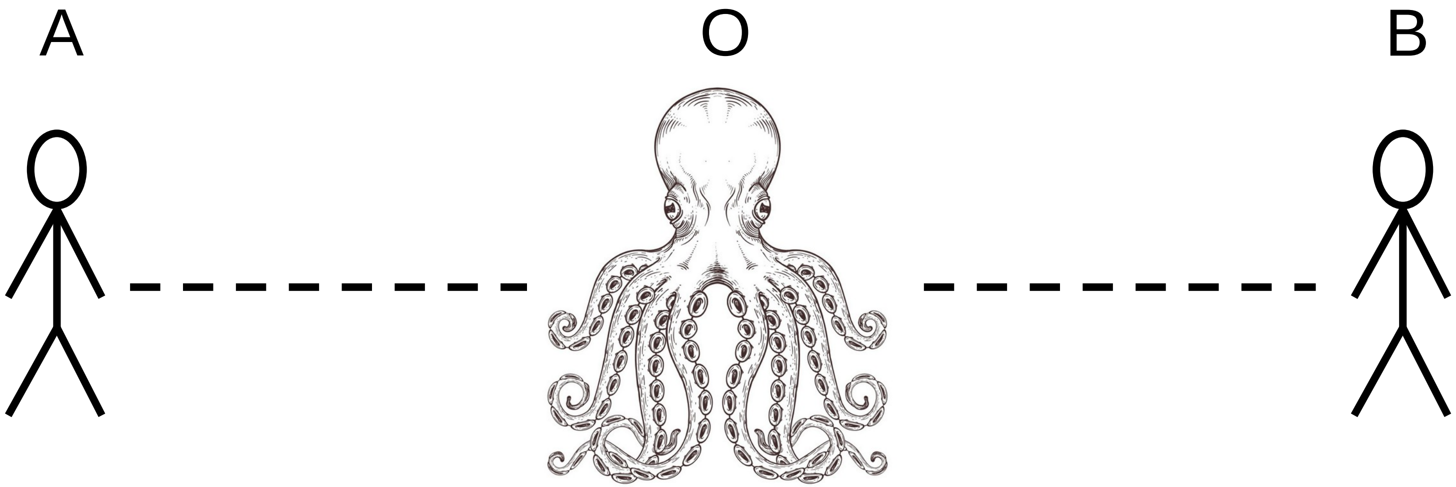
Negation is a major problem, and increasing model size does not help (it even hinders).

Grounding

(Bender & Koller 2020)

Grounding

“O knows nothing about English initially, but is very good at detecting statistical patterns. Over time, O learns to predict with great accuracy how B will respond to each of A’s utterances” (p. 5188)



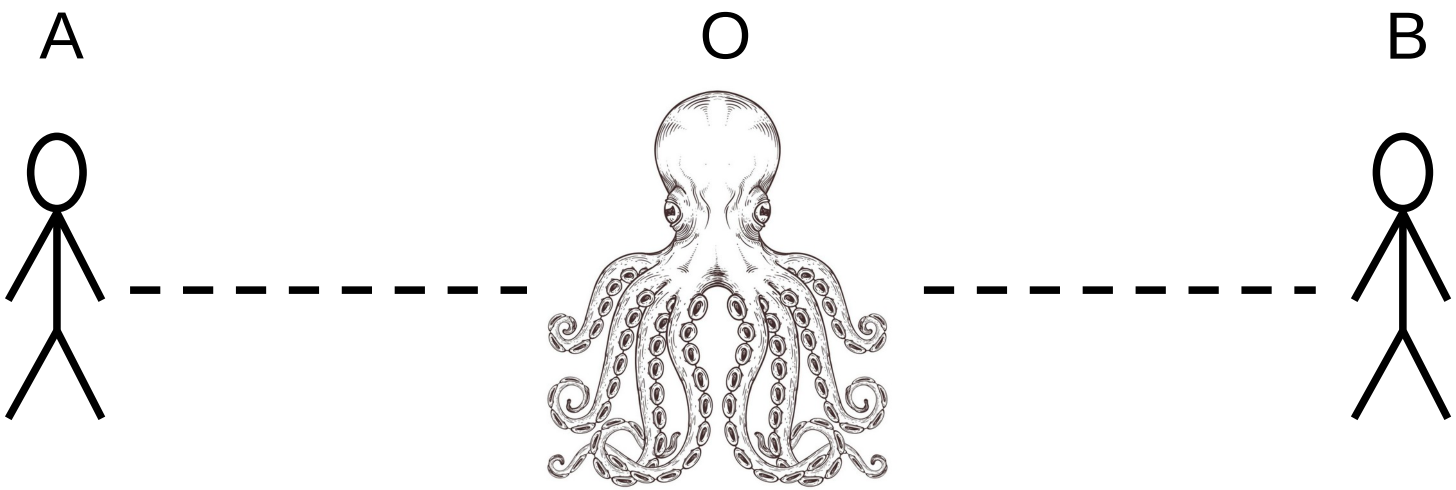
(Bender & Koller 2020)

Grounding

“O knows nothing about English initially, but is very good at detecting statistical patterns. Over time, O learns to predict with great accuracy how B will respond to each of A’s utterances” (p. 5188)

“Having only form available as training data, O did not learn meaning.” (p. 5189)

“O’s language use will eventually diverge from the language use of an agent who can ground their language in coherent communicative intents.” (p. 5188)



(Bender & Koller 2020)

Grounding

Pavlick (2023); grounding not necessary for (all) semantics

- Internalist conceptual role semantics: possible even if ungrounded
- Externalist causal/informational semantics: possible even for ungrounded representations
- Mapping representations: ungrounded → grounded (Scialom 2020, Abdou et al. 2021)

Grounding

Pavlick (2023); grounding not necessary for (all) semantics

- Internalist conceptual role semantics: possible even if ungrounded
- Externalist causal/informational semantics: possible even for ungrounded representations
- Mapping representations: ungrounded → grounded (Scialom 2020, Abdou et al. 2021)

Much depends on meta-semantics...

- Causal/informational (Cappelen and Dever 2021, Mandelkern and Linzen 2023)
- Descriptivist/inferential (Pavlick 2023)
- Interpretationist (Lederman and Mahowald 2024)

Grounding

Pavlick (2023); grounding not necessary for (all) semantics

- Internalist conceptual role semantics: possible even if ungrounded
- Externalist causal/informational semantics: possible even for ungrounded representations
- Mapping representations: ungrounded → grounded (Scialom 2020, Abdou et al. 2021)

Much depends on meta-semantics...

- Causal/informational (Cappelen and Dever 2021, Mandelkern and Linzen 2023)
- Descriptivist/inferential (Pavlick 2023)
- Interpretationist (Lederman and Mahowald 2024)

Thank you!

References

Mostafa Abdou, Artur Kulmizev, Daniel Hershcovich, Stella Frank, Ellie Pavlick, Anders Søgaard. 2021. Can Language Models Encode Perceptual Structure Without Grounding? A Case Study in Color. In Proceedings of the 25th Conference on Computational Natural Language Learning, pages 109–132.

Marianna Apidianaki and Aina Garí Soler. 2021. ALL dolphins are intelligent and SOME are friendly: Probing BERT for nouns' semantic properties and their prototypicality. In Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP, pages 79–94.

Nikolay Arefyev, Boris Sheludko, Alexander Podolskiy, and Alexander Panchenko. 2020. Always keep your target in mind: Studying semantics and improving performance of neural lexical substitution. In Proceedings of the 28th International Conference on Computational Linguistics, pages 1242–1255.

Sriram Balasubramanian, Naman Jain, Gaurav Jindal, Abhijeet Awasthi, and Sunita Sarawagi. 2020. Proceedings of the 5th Workshop on Representation Learning for NLP (RepL4NLP-2020), pages 205–214.

Emily M. Bender, Alexander Koller. 2020. Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 5185–5198.

Tyler A. Chang & Benjamin K. Bergen. 2024. Language Model Behavior: A Comprehensive Survey. Computational Linguistics 50 (1): 293–350.

Boxi Cao, Hongyu Lin, Xianpei Han, Le Sun, Lingyong Yan, Meng Liao, Tong Xue, and Jin Xu. 2021. Knowledgeable or educated guess? Revisiting language models as knowledge bases. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 1860–1874.

Herman Cappelen and Josh Dever. 2021. Making AI intelligible: Philosophical foundations. Oxford: Oxford University Press.

Forrest Davis and Marten van Schijndel. 2020. Discourse structure interacts with reference but not syntax in neural language models. In Proceedings of the 24th Conference on Computational Natural Language Learning, pages 396–407.

References

Ettinger, Allyson. 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.

Maxwell Forbes, Ari Holtzman, and Yejin Choi. 2019. Do Neural Language Representations Learn Physical Commonsense? In Proceedings of the 41st Annual Conference of the Cognitive Science Society (CogSci 2019), page 7.

Atticus Geiger, Kyle Richardson, and Christopher Potts. 2020. Neural natural language inference models partially embed theories of lexical entailment and negation. In Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP, pages 163–173.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer Feed-Forward Layers Are Key-Value Memories. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 5484–5495.

Bernadeta Griciūtė, Marc Tanti, and Lucia Donatelli. 2022. On the cusp of comprehensibility: Can language models distinguish between metaphors and nonsense? In Proceedings of the 3rd Workshop on Figurative Language Processing (FLP), pages 173–177.

Michael Hanna and David Mareček. 2021. Analyzing BERT's knowledge of hypernymy via prompting. In Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP, pages 275–282.

Qianyu He, Sijie Cheng, Zhixu Li, Rui Xie, and Yanghua Xiao. 2022a. Can pre-trained language models interpret similes as smart as human? In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 7875–7887.

Md Mosharaf Hossain, Venelin Kovatchev, Pranoy Dutta, Tiffany Kao, Elizabeth Wei, and Eduardo Blanco. 2020. An analysis of natural language inference benchmarks through the lens of negation. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 9106–9118.

References

Arian Hosseini, Ankit Vani, Dzmitry Bahdanau, Alessandro Sordoni, and Aaron Courville. 2022. On the compositional generalization gap of in-context learning. In Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP, pages 272–280.

Dieuwke Hupkes, Verna Dankers, Mathijs Mul, and Elia Bruni. 2020. Compositionality decomposed: How do neural networks generalise? Journal of Artificial Intelligence Research, 67:757–795.

Joel Jang, Seonghyeon Ye, and Minjoon Seo. 2022. Can large language models truly understand prompts? A case study with negated prompts. In Proceedings of the 1st Transfer Learning for Natural Language Processing Workshop, pages 52–62.

Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. 2020b. How can we know what language models know? Transactions of the Association for Computational Linguistics, 8:423–438.

Kassner, Nora and Hinrich Schütze. 2020. Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 7811–7818.

Nora Kassner, Benno Krojer, and Hinrich Schütze. 2020. Are Pretrained Language Models Symbolic Reasoners over Knowledge? Proceedings of the 24th Conference on Computational Natural Language Learning, pages 552–564.

Jan Christoph Kalo and Leandra Fichtel. 2022. KAMEL: Knowledge analysis with multitoken entities in language models. In 4th Conference on Automated Knowledge Base Construction.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In Advances in Neural Information Processing Systems, volume 35, pages 22199–22213.

Artur Kulmizev, Vinit Ravishankar, Mostafa Abdou, and Joakim Nivre. 2020. Do Neural Language Models Show Preferences for Syntactic Formalisms? Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 4077–4091.

References

Ilia Kuznetsov and Iryna Gurevych. 2020. A matter of framing: The impact of linguistic formalism on probing results. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 171–182.

Yash Kumar Lal, Niket Tandon, Tanvi Aggarwal, Horace Liu, Nathanael Chambers, Raymond Mooney, and Niranjan Balasubramanian. 2022. Using commonsense knowledge to answer why questions. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 1204–1219.

Harvey Lederman and Kyle Mahowald. Are Language Models More Like Libraries or Like Librarians? Bibliotechnism, the Novel Reference Problem, and the Attitudes of LLMs. Arxiv preprint arXiv:2401.04854.

Bill Yuchen Lin, Seyeon Lee, Rahul Khanna, and Xiang Ren. 2020. Birds have four legs?! NumerSense: Probing numerical commonsense knowledge of pre-trained language models. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 6862–6868.

Emmy Liu, Chenxuan Cui, Kenneth Zheng, and Graham Neubig. 2022a. Testing the ability of language models to interpret figurative language. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 4437–4452.

Charles Lovering and Ellie Pavlick. 2022 Unit testing for concepts in neural networks. Transactions of the Association for Computational Linguistics 10: 1193–1208.

Matthew Mandelkern and Tal Linzen. Do Language Models' Words Refer? Arxiv preprint arXiv:2308.05576.

Christopher D. Manning, Kevin Clark, John Hewitt, Urvashi Khandelwal, and Omer Levy. 2020. Emergent linguistic structure in artificial neural networks trained by self supervision. Proceedings of the National Academy of Sciences, page 201907367.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. 2022. In-context Learning and Induction Heads. Arxiv preprint arXiv:2209.11895.

References

Ellie Pavlick. 2023. Symbols and grounding in large language models. Philosophical Transactions of the Royal Society 381.

Paolo Pedinotti, Eliana Di Palma, Ludovica Cerini, and Alessandro Lenci. 2021a. A howling success or a working sea? Testing what BERT knows about metaphors. In Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP, pages 192–204.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 2463–2473.

Nina Poerner, Ulli Waltinger, and Hinrich Schütze. 2019. BERT is not a knowledge base (yet): Factual knowledge vs. name-based reasoning in unsupervised QA. ArXiv, arXiv:1911.03681.

Linlu Qiu, Peter Shaw, Panupong Pasupat, Tianze Shi, Jonathan Herzig, Emily Pitler, Fei Sha, and Kristina Toutanova. 2022. Evaluating the impact of model scale for compositional generalization in semantic parsing. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 9157–9179.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. Transactions of the Association for Computational Linguistics, 8: 842–866.

Pritish Sahu, Michael Cogswell, Yunye Gong, and Ajay Divakaran. 2022. Unpacking large language models with conceptual consistency. ArXiv preprint arXiv:2209.15093.

Maarten Sap, Ronan Le Bras, Daniel Fried, and Yejin Choi. 2022. Neural theory-of-mind? On the limits of social intelligence in large LMs. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 3762–3780.

Abulhair Saparov and He He. 2023. Language models are greedy reasoners: A systematic formal analysis of chain-of-thought. In International Conference on Learning Representations.

References

Sebastian Schuster and Tal Linzen. 2022. When a sentence does not introduce a discourse entity, transformer-based models still sometimes refer to it. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 969–982.

Thomas Scialom, Patrick Bordes, Paul-Alexis Dray, Jacopo Staiano, Patrick Gallinari. 2020. What BERT Sees: Cross-Modal Transfer for Visual Question Generation. In Proceedings of the 13th International Conference on Natural Language Generation, pages 327–337.

Haohan Shi and Phillip Wolff. 2021. What Transformers might know about the physical world: T5 and the origins of knowledge. In Proceedings of the Annual Meeting of the Cognitive Science Society, volume 43, pages 2218–2224.

Douglas Summers-Stay, Claire Bonial, and Clare Voss. 2021. What can a generative language model answer about a passage? In Proceedings of the 3rd Workshop on Machine Reading for Question Answering, pages 73–81.

Giorgia Nidia Carranza Tejada, Johannes Scholtes, and Gerasimos Spanakis. 2021. A study of BERT's processing of negations to determine sentiment. Benelux Conference on Artificial Intelligence and the Belgian Dutch Conference on Machine Learning, pages 47–59.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT Rediscovers the Classical NLP Pipeline. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 4593–4601.

Thinh Hung Truong, Yulia Otmakhova, Timothy Baldwin, Trevor Cohn, Jey Han Lau, and Karin Verspoor. 2022. Not another negation benchmark: The NaN-NLI test suite for sub clausal negation. In Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 883–894.

Asahi Ushio, Luis Espinosa Anke, Steven Schockaert, and Jose Camacho-Collados. 2021. BERT is to NLP what AlexNet is to CV: Can pre-trained language models identify analogies? In Proceedings of the 59th Annual Meeting of the Association forComputational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 3609–3624.

References

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser & Illia Polosukhins. 2017. Attention is all you need. Proceedings of the 31st International Conference on Neural Information Processing, pages 6000–6010.

Eric Wallace, Yizhong Wang, Sujian Li, Sameer Singh, and Matt Gardner. 2019. Do NLP models know numbers? Probing numeracy in embeddings. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 5307–5315.

Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin & Mengnan Du. 2024. Explainability for Large Language Models: A Survey. ACM Transactions on Intelligent Systems and Technology, 15(2): 1–38.