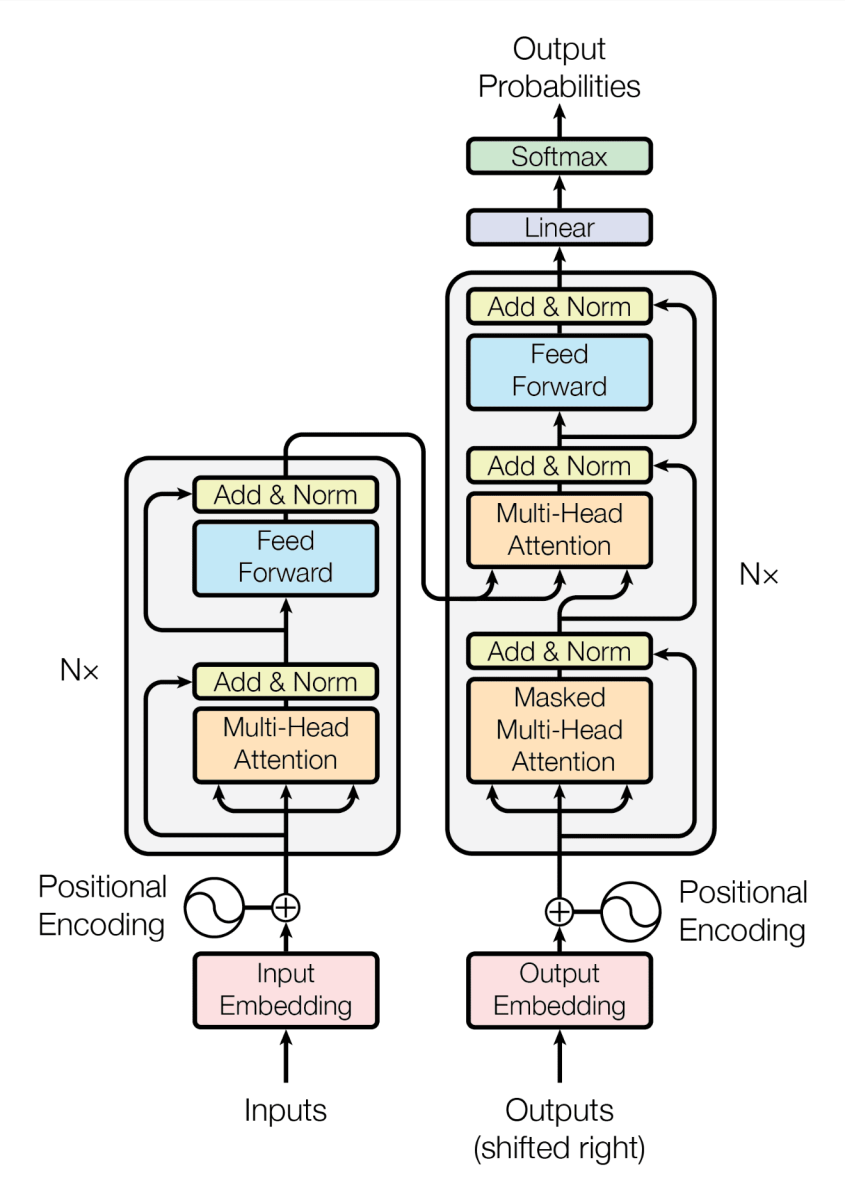
Syntax and grounding in adjective learning
General supplementary material

Adèle Hénot-Mortier (MIT)

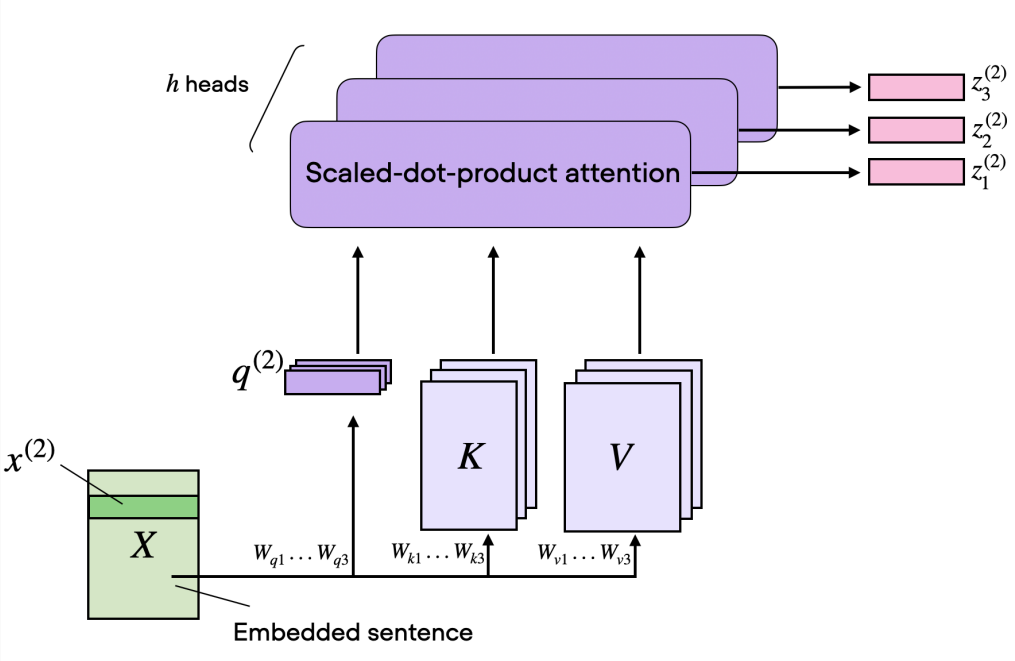
February 27, 2025

Queen Mary University of London

Full Transformer architecture



Multi-head self attention

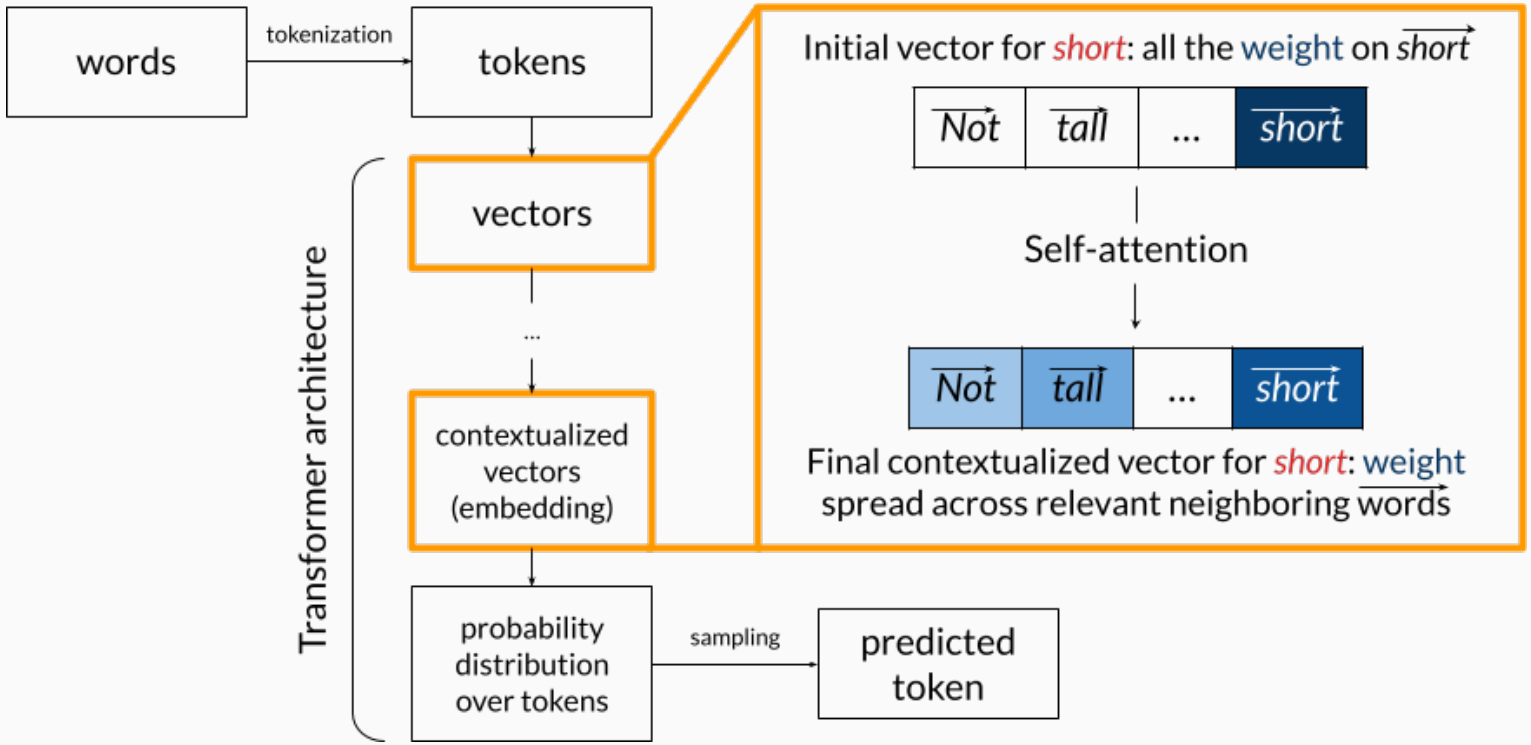


Model parameters

Model	Structure	Parameters	Layers	Hidden dimension
GPT2-Large	Decoder	774M	36	1280
BERT-Large	Encoder	340M	24	1024
RoBERTa-Large	Encoder	355M	24	1024
XLNet-Large	Decoder	340M	24	1024
Mistral 7B v0.1	Decoder	7B	32	4096

Self-attention creates context-sensitive word representations

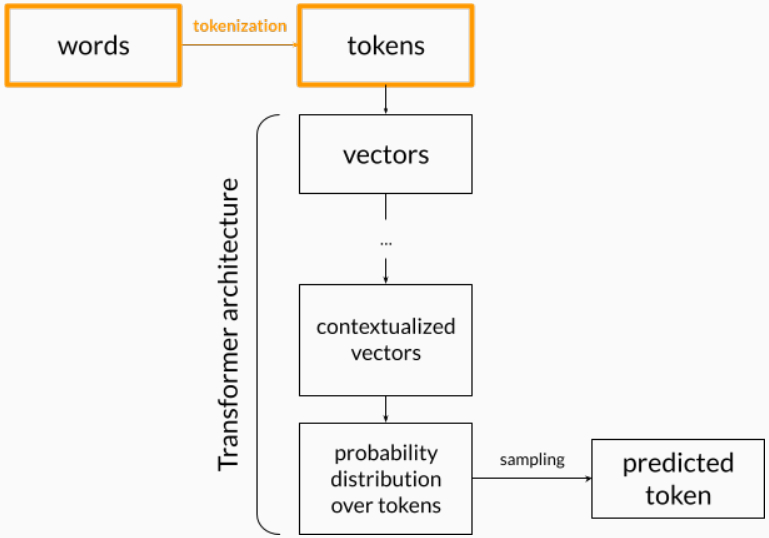
- **Self-attention** is at the core of Transformers, and should allow them to grasp the **contextualized meaning** of antonymic adjectives, and the **functional behavior** of negation.[1]



[1] Though see Peng et al. (2024).

Tokenization as a proxy for morphological decomposition

- **Tokenization** maps sentences into tokens, which represent words or pieces of words.
- Tokens are determined based on **character coocurrences**.
- Tokens may therefore reflect **morphology**,[2] and provide LLMs with useful distributional cues to derive ITA contrasts.



[2] Nair and Resnik, 2023