



Extending large language models

Description

I'm in the process of identifying parallels between urban semiotics and large language models (LLMs), arguing that core facets of language competence parallel aspects of urban life, experiences and processes. We can identify eight core functions that contribute to the success of the [Transformer model](#) of LLMs, as deployed in ChatGPT and other chatbot platforms. I derive these core functions from the seminal article from 2017 by Vaswani et al called "Attention Is All You Need" and various commentaries.

Core LLM functions

The core LLM functions are (i) the identification and manipulation of **patterns** in data, (ii) reliance on a vast collection of existing data (in text-based LLMs this is a **corpus** of existing texts), (iii) the division of the corpus into **tokens** (or words), (iv) the use of very large mathematical matrices that capture and process **semantic** information based on word proximities, (v) the establishment of **context** windows, (vi) adjustments to semantic matrices to account for the **positioning** of tokens in the context window, (vii) manipulations based on alternative distributions of **attention** (or emphasis) within texts, and (viii) the **tuning** of the model by human operatives.

Large language models and their applications are developing all the time. I draw on a helpful podcast *Last Week in AI* (<https://www.lastweekinai.com>) by Andrey Kurenkov and Jeremie Harris for updates.

There are LLM developments extrinsic to the 8 core functions that I've just identified, and that add functionality and customise the user experience for different contexts. Recent developments and enhancements extend beyond the eight core functions, encompassing a range of features and implementations. Such developments include

- expanded user interaction including voice chat
- integration into existing applications such as word processing, education tools and customer service apps
- web search integration with data taken directly from the web, as well as AI integration to enhance web search

- modules that deal with mathematics, logic, image recognition and generation, spatial inference and other specialised functions
- incorporation of spelling and grammar checkers and other text processing functions into LLM text generation

Other innovations focus on efficiency and coherence in extended conversations or tasks requiring large bodies of text. This may required expanding the context window to millions of tokens, enabling LLMs to consider a much larger span of text when generating responses. This capability is applicable to complicated tasks that require processing and referencing extensive documents or long-term contexts. One technique involves **context caching** for storing and reusing previously processed content, allowing the LLM to maintain continuity over longer interactions without reprocessing the entire context.

Google has championed context extension with Gemini 1.5. Gemini chat tells me:

â??**Limited Preview of 1 Million Tokens:** Google offers a limited preview program for developers and enterprise customers where the context window can be extended to a whopping 1 million tokens. This is still under development, but it showcases the potential of Geminiâ??s architecture. Itâ??s important to note that using the 1 million token window is currently limited and requires access to specific Google Cloud programs.â?•

On context windows see post: [AI learns abc](#).

References

- Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. â??Attention Is All You Need.â?• In *31st Conference on Neural Information Processing Systems*, 1-15. Long Beach, CA, USA, 2017.

Note

- In terms of urban parallels, cities are made of (i) *patterns*. We can recast the function of the LLM text (ii) *corpus* in terms of urban archives. The elements of the city are subject to partitioning in various ways: (iii) *tokenization*. Cities are formed by sets of relationships, readily construed as meaning structures or (iv) *semantic* relationships. Cities provide a rich set of (v) *contexts* in which meanings are negotiated. In spite of ubiquitous digital communications, it still matters where urban elements are (vi) *positioned* relative to one another. The elements of a city compete for (vii) *attention*, from prominent landmarks to graffiti tags on walls. Our entire experience of cities lends itself to metaphors of (viii) *tuning* in and tuning out.

Category

1. Artificial Intelligence

Tags

1. transformer

2. urbanism

Date Created

July 13, 2024

Author

rcoyne99

default watermark