Smashing the context window

Description

default watermark

In natural language processing LLMs, the â??context windowâ?• is the range of text around a target word or phrase that constrains the sequence of text processed in calculating positional and attentional encoding, and hence prediction of the text that follows.

A larger context window allows the AI model to capture more subtle relationships and dependencies across longer text sequences, leading to more accurate and coherent predictions and outputs. In LLMs, context windows can range from a few hundred tokens to several thousand tokens, with some advanced models supporting windows of up to 8,000 tokens or more.

A limit of 8,000 tokens (about 7,000 words) still imposes constraints on the quantity of text that you (the user) can introduce to the LLM. How can an LLM handle queries about a whole book, or even several books?

As I showed in my previous posts, you can upload texts of several hundred words to the NotebookLM language model. The platform can respond to prompts related to the text. You can ask questions about it, call on summaries and effectively â??discussâ?• the content of the text with the LLM. You can even upload several texts and ask the model to compare and contrast them, identify common themes and synthesise their content. NotebookLM seems thereby able to overcome the limits of context window size.

There are several means of accomplishing this ability to handle large quantities of text in excess of the context window size. Here, Iâ??ll focus on just one involving the Transformer model augmented with AI search.

RAG

Retrieval-Augmented Generation (RAG) is an approach used in some large language models to handle constraints posed by limited context window sizes. RAG combines retrieval mechanisms with text generation to improve the modelâ??s response quality, especially when generating responses that require information beyond what is directly within the context window.

The Retrieval-Augmented Generation (RAG) method uses AI-based search to identify relevant chunks of text in source documents, enhancing an LLMâ??s responses by embedding additional context beyond its immediate window. Rather than keyword search, RAG uses a neural network trained on the LLMâ??s original data, enabling semantic retrieval based on thematic similarity.

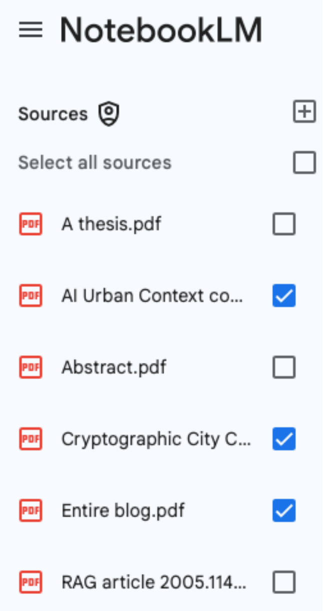
In this process, the training data is divided into overlapping chunks, each approximately 100 tokens long, to maintain coherence. During training, these chunks undergo semantic, positional, and attentional encodingâ??following Transformer model principlesâ??to produce high-dimensional vectors that encapsulate each chunkâ??s thematic content.

The neural network is trained to recognise these patterns, embedding them into vectors that allow efficient, similarity-based retrieval. At query time, the userâ??s prompt is also vectorised, producing a thematic representation that undergoes similarity search against the indexed vectors of the document chunks. Chunks that most closely match the query are then added to the context window, providing the LLM with detailed, relevant information to support accurate and context-rich responses without increasing the modelâ??s memory or window size.

The chunking and indexing of a query document typically occurs around the time it is uploaded. When a user uploads a document to be queried, the RAG system first processes this document by dividing it into manageable chunks. Each chunk is then immediately encoded into a high-dimensional vector representation using the pre-trained neural network, capturing its thematic and semantic patterns. These vectors are subsequently indexed in a database, making them readily accessible for similarity-based retrieval when a query is issued. This preprocessing step ensures that the system can perform real-time similarity searches, as it allows the relevant chunks to be efficiently retrieved and integrated into the LLMâ??s context window at query time. It can take some minutes to upload and preprocess a large query document and responses to queries take a little longer than the usual conversational AI response times.

default watermark

As an illustration of the capabilities of this approach I checked three of the text sources I had already uploaded to NotebookLM.



Select all sources ☐
A thesis.pdf ☐
AI Urban Context co... ☑
Abstract.pdf ☐
Cryptographic City C... ☑
Entire blog.pdf ☑
RAG article 2005.114... ☐

The first is a draft of my latest book on *AI and Language in the Urban Context*. The second is the publicly available PDF of my book *Cryptographic City*. The third is a PDF of this entire blog of 700,000 words. I asked â??The three sources are by me. Can you suggest any future book projects?â?•

> â??Here are some potential book projects, based on the sources you provided and our conversation:
>
> **AI and Urban Design: A Critical and Creative Exploration:** This book could expand on your existing research on the intersection of AI and urban design. It could provide a comprehensive overview of the different ways that AI can be used to analyse urban data, generate design solutions, simulate urban scenarios, and facilitate public engagement. The book could also address the ethical and cultural implications of AI in urban design, drawing on insights from the humanities and social sciences.
>
> - This project could draw heavily on source 1 (â??AI Urban Context combined.pdfâ?•), which explicitly discusses AIâ??s application in urban environments. It could elaborate on concepts like scripting, tuning, and tokenization in LLMs, discussed in the source, and their relevance to urban design.
> - Additionally, you could incorporate content from our conversation, where we discussed using AI to analyse urban data, simulate scenarios, and engage the public in urban design.
> - The book could also explore the broader cultural and societal impacts of AI in cities, drawing on themes and insights from source 3 (â??no pics 3 Nov 2024 Reflections on Technology, Media & Culture.pdfâ?•), which addresses the impact of digital technologies on various aspects of culture and society.
>
> **The Cryptographic City: Unmasking the Hidden Dimensions of Urban Life:** This book could build on your existing work on cryptography and the city (source 2: â??Cryptographic City COYNE ebook_9780262374811.pdfâ?•). It could explore the different ways that cryptography is used to shape urban spaces, from surveillance systems and data collection to the use of encryption in communication and financial transactions. The book could also examine the social, political, and ethical implications of cryptography in urban life.
>
> This project could expand on the themes of source 2, such as:
>
> - The use of cryptography in smart cities and its implications for privacy and security.
> - The role of cryptography in shaping urban power dynamics and control.
> - The potential for cryptography to be used for both oppressive and liberating purposes in cities.
>
> The book could also connect the concept of cryptography with broader discussions on technology, media, and culture, as explored in source 3. For example, you could discuss how cryptography interacts with social media, digital art, and online activism in urban contexts.

default watermark

Designing for a Post-Digital City: This book could explore the emerging field of post-digital design and its implications for urban environments. It could examine how designers are using new technologies, such as AI, virtual reality, and 3D printing, to create more responsive, sustainable, and equitable cities. The book could also discuss the challenges and opportunities of designing for a post-digital city, where the boundaries between the physical and digital are increasingly blurred.

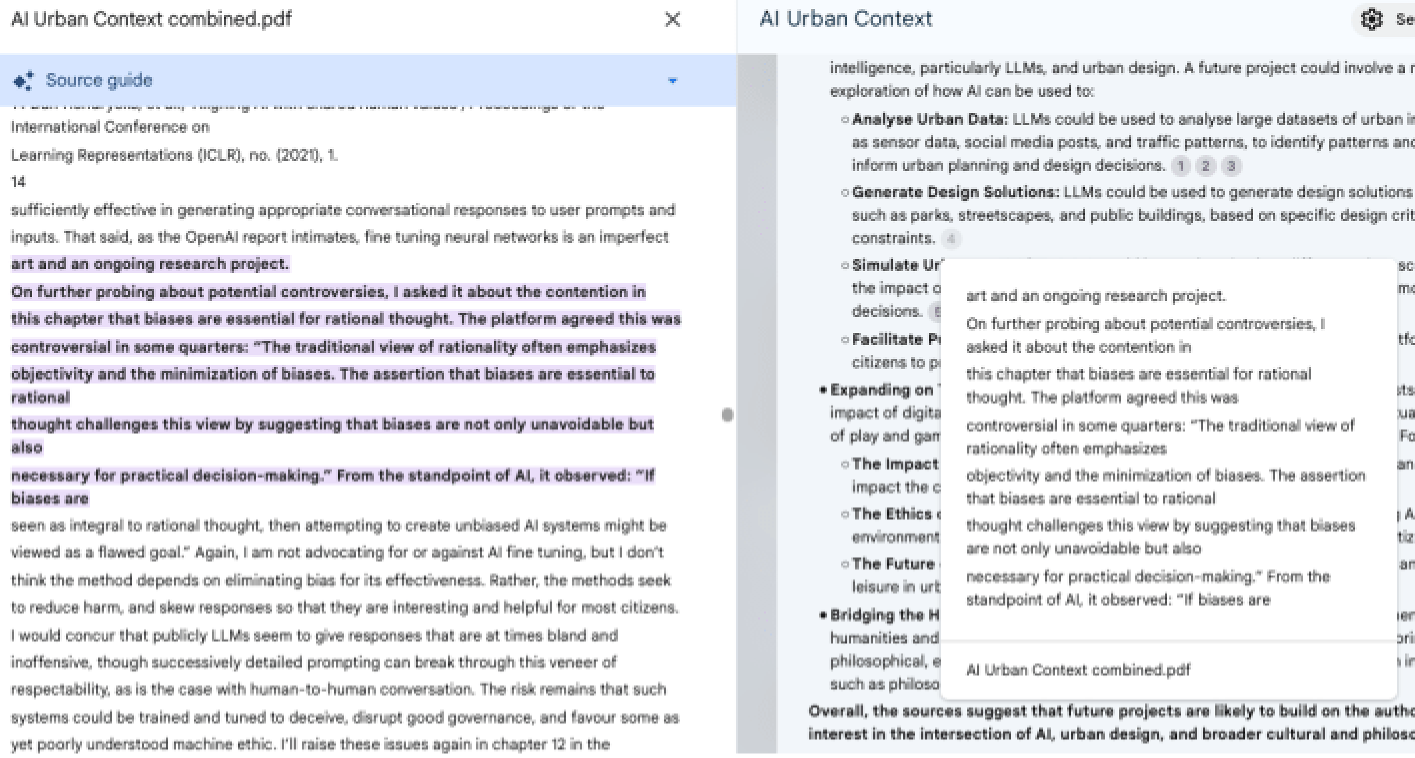
This project could synthesise insights from all three sources.

- Source 1 provides a foundation for discussing AIâ??s role in post-digital urban design.
- Source 2 offers insights into how cryptography and data security will shape post-digital urban spaces.
- Source 3, with its focus on technologyâ??s impact on media, culture, and society, could inform discussions on the social and cultural dimensions of post-digital cities.

These are just a few initial suggestions. The best book project will depend on your specific interests and goals. It is important to consider the intended audience, the scope of the project, and the potential contribution to the field.â?•

default watermark

The method also enables the LLM to hyperlink to specific paragraphs in the source text. Hereâ??s a screenshot as copy-paste does not preserve the links.



International Conference on

Learning Representations (ICLR), no. (2021), 1.

14

sufficiently effective in generating appropriate conversational responses to user prompts and inputs. That said, as the OpenAI report intimates, fine tuning neural networks is an imperfect

art and an ongoing research project.

On further probing about potential controversies, I asked it about the contention in this chapter that biases are essential for rational thought. The platform agreed this was controversial in some quarters: "The traditional view of rationality often emphasizes objectivity and the minimization of biases. The assertion that biases are essential to rational thought challenges this view by suggesting that biases are not only unavoidable but also necessary for practical decision-making." From the standpoint of AI, it observed: "If biases are

seen as integral to rational thought, then attempting to create unbiased AI systems might be viewed as a flawed goal." Again, I am not advocating for or against AI fine tuning, but I don't think the method depends on eliminating bias for its effectiveness. Rather, the methods seek to reduce harm, and skew responses so that they are interesting and helpful for most citizens. I would concur that publicly LLMs seem to give responses that are at times bland and inoffensive, though successively detailed prompting can break through this veneer of respectability, as is the case with human-to-human conversation. The risk remains that such systems could be trained and tuned to deceive, disrupt good governance, and favour some as yet poorly understood machine ethic. I'll raise these issues again in chapter 12 in the

intelligence, particularly LLMs, and urban design. A future project could involve a
exploration of how AI can be used to:

- **Analyse Urban Data:** LLMs could be used to analyse large datasets of urban
  as sensor data, social media posts, and traffic patterns, to identify patterns and
  inform urban planning and design decisions. ① ② ③
- **Generate Design Solutions:** LLMs could be used to generate design solutions
  such as parks, streetscapes, and public buildings, based on specific design crit
  constraints. ④
- **Simulate Ur**
  the impact o
  decisions.
- **Facilitate P**
  citizens to p

- **Expanding on**
  impact of digita
  of play and gam
  - **The Impact**
    impact the c
  - **The Ethics**
    environment
  - **The Future**
    leisure in urb
- **Bridging the H**
  humanities and
  philosophical, e
  such as philoso

art and an ongoing research project.

On further probing about potential controversies, I asked it about the contention in this chapter that biases are essential for rational thought. The platform agreed this was controversial in some quarters: "The traditional view of rationality often emphasizes objectivity and the minimization of biases. The assertion that biases are essential to rational thought challenges this view by suggesting that biases are not only unavoidable but also necessary for practical decision-making." From the standpoint of AI, it observed: "If biases are

Overall, the sources suggest that future projects are likely to build on the autho
interest in the intersection of AI, urban design, and broader cultural and philoso

Bibliography

- Lewis, P., et al. (2021), â??Retrieval-Augmented Generation for Knowledge-Intensive Nlp Tasksâ??. *arXiv*, 2005.11401. Available online: https://arxiv.org/abs/2005.11401 (accessed 9

November 2024).

Note

- Featured image is by WordPress generative image tool.

Category

1. Artificial Intelligence

Tags

1. Book proposal
2. context window
3. LLM
4. RAG

default watermark

Date Created
November 9, 2024
Author
rcoyne99