ISSUE BRIEF

ARTIFICIAL INTELLIGENCE

Technologist Roundtable: Key Issues in AI and Data Protection
Post-Event Summary and Takeaways

November 27, 2024

Dr. Rob van Eijk, Stacey Gray, Marlene Smith

FUTURE OF PRIVACY FORUM | CENTER FOR ARTIFICIAL INTELLIGENCE

About FPF

The Future of Privacy Forum (FPF) is a non-profit organization that serves as a catalyst for privacy leadership and scholarship, advancing principled data practices in support of emerging technologies. For more about FPF, please visit us at fpf.org/about and learn more about the FPF Center for Artificial Intelligence at fpf.org/issue/ai-ml/.

Event Summary

On 27 November 2024, the Future of Privacy Forum (FPF) hosted a Technologist Roundtable with the goal of convening an open dialogue on complex technical questions that impact law and policy, and assisting global data protection and privacy policymakers in understanding the relevant technical basics of large language models (LLMs). We invited a wide range of academic technical experts to convene with each other and data protection regulators and policymakers from around the world.

As a result of the emergence of LLMs, data protection authorities and lawmakers are exploring a range of novel data protection issues, including how to ensure lawful processing of personal data in LLMs, and obligations to comply with obligations such as data deletion and correction requests. While LLMs can process personal data at different stages,[1] including in training and in the input and output of models, there is an emerging question of the extent to which personal data exists "within" a model itself.[2] Navigating these complex emerging issues increasingly requires understanding the technical building blocks of LLMs.

This post-event summary contains highlights and key takeaways from three parts of the Roundtable on 27 November.

[1] *See e.g.,* EDPB ChatGPT Taskforce Report, *available at* https://www.edpb.europa.eu/our-work-tools/our-documents/other/report-work-undertaken-chatgpt-taskforce_en

[2] *See, e.g.*, Future of Privacy Forum, Do LLMs Contain Personal Information? California AB 1008 Highlights Evolving, Complex Techno-Legal Debate (Oct 2024), https://fpf.org/blog/do-llms-contain-personal-information-california-ab-1008-highlights-evolving-complex-techno-legal-debate/

Table of Contents

1. Basics of Transformer Technology and Tokenization

- *What is involved in pre-training, training, and post-training of a Generative Pre-trained Transformer (GPT) approach to create Large Language Models (LLMs)?*
- *What are the most relevant aspects of tokenization, the transformer architecture, and foundation models?*

The Roundtable began with an introductory discussion of the technical basics of tokenization, the training pipeline stages, and the concept of memorization, which we supplement here with additional information. In recent years, the fields of artificial intelligence (AI) and machine learning (ML) have been galvanized by the emergence of **transformer architecture**, a type of neural network that has become the foundation for large language models (LLMs). The underlying technology has revolutionized natural language processing (NLP), computer vision, and multimodal applications, and most recently given rise to the broader category of **foundation models**, or large-scale general purpose AI models, pre-trained on very large datasets, which can be fine-tuned or adapted to a wide range of tasks. The emergence of general-purpose AI (GPAI) capabilities stems from increased compute and data, as well as key architectural elements such as tokenization and advanced multi-stage training.

a. Tokenization

When you interact with a Large language models (LLM) system, the text first goes through a preprocessor called a tokenizer, which breaks down the input into discrete tokens, representing words or portions of words. Tokens are converted into vectors, or dimensional numerical representations, which are processed through the LLM's transformer architecture. These numerical tokens are processed through neural networks to make predictions one token at a time, with each prediction based on all previous tokens, similar to how humans compose sentences by considering earlier words. For example, in case of generating an answer to the prompt "Plants need several (...)", the model predicts:

> "Plants" -> "need" -> "several" -> "essential" -> "components" -> "to" -> "grow" -> ":" -> "sunlight" -> "," -> "water" -> "," -> "and" -> "nutrients" -> "from" -> "the" -> "soil" -> "."

The **context array**, represented as a sequence of numbers like [128006,882,128007,...], serves as the model's working memory by containing both the original input tokens and all generated tokens. These numbers function like a dictionary, where each number corresponds to a specific piece of text in the model's vocabulary, i.e., 128006 represents "Plants", 882 represents " need", and so on (see Appendix 1 for more detail). Rather than working with raw text, the model operates with these numerical tokens internally, using the entire context array to track what it has generated and inform each new prediction. Finally, the predicted numerical tokens are converted back into human-readable text for output.

b. *Stages of Training*

LLMs are part of the GPT (Generative Pre-trained Transformer) family, and represent a particularly successful application of this approach in the language domain. LLM development begins with (1) ***pre-training*** to learn statistical patterns and representations based on enormous datasets of uncategorized information. This is followed by (2) ***mid-training*** to enhance specific capabilities or align with desired behaviors, beginning to shape model behavior and responses. Mid-training often involves instruction tuning and can include, e.g., reinforcement learning from human feedback (RLHF) to improve task performance and alignment. The process concludes with (3) ***post-training*** where fine-tuning of models helps them to adapt to specific downstream tasks, such as sentiment analysis, named entity recognition, or text summarization. In this stage, fine-tuning adapts models to specific tasks using smaller, specialized datasets, and can introduce privacy risks due to (typically) greater data sensitivity.

The adaptability of these models is demonstrated through various applications. For instance, (1) developers can fine-tune models for *binary classification tasks* like spam detection by training on labeled examples, or (2) they can *instruction-tune* them to follow specific formats and commands through careful curation of instruction-response pairs. These advances in AI design and training methods have changed how machine learning works. Now, AI systems can learn many new tasks from a small number of examples.

c. *Memorization*

*M*emorization in the context of LLMs typically refers to verbatim memorization, or how a model can potentially reproduce exact or identical content from its training datasets. Memorization is usually considered distinct from a model's general ability to extrapolate and reproduce patterns in training data, although it emerges from the same fundamental architecture. It has particular implications for copyrighted material and personal information.[3] Experts discussed that memorization of information occurs at the parameter level, where repeated patterns in the vocabulary become encoded across the model's weights. The challenge is finding the right balance between extracting generalizable patterns from training data and avoiding verbatim reproduction of training sequences.

In a discussion of how to measure, detect, and mitigate memorization, experts observed that approaches are evolving, for example with fine-tuning to update the model's weights to improve text summarization. Another method of testing memorization, randomized controlled setups with ***canaries*** - synthetically generated, unique sentences inserted into training data - could potentially be used to detect memorization. Canaries have been used in randomized controlled setups, allowing researchers

[3] For example, experts observed the heightened urgency of this issue following the New York Times' lawsuit against OpenAI, which demonstrated that the LLM was able to produce nearly verbatim copies of articles when prompted with their first lines. https://www.nytimes.com/2023/12/27/business/media/new-york-times-open-ai-microsoft-lawsuit.html.

to causally demonstrate that longer sequences and higher perplexity (more “surprising”) content are more likely to be memorized.

Discussion Takeaways from Part 1 (Basics of Transformer Technology and Tokenization):

- LLMs face privacy risks at each step of training. Each stage has different risks and mitigation strategies. Fine-tuning - in contrast to pre-training and mid-training - presents particular privacy concerns due to potentially sensitive data.
- Memorization does not always have to be exact. Instead, we see near verbatim memorized content as well. Deduplication helps reduce memorized content, but deduplication is not always straightforward to implement because duplicates can be subjective or determined.[4] Other techniques like differential privacy or similar techniques may help prevent a model from overfitting (independent of training data).
- One expert highlighted the “Reversal Curse” - an apparent limitation in how LLMs learn about directional relationships that exist in their training data. For example: a model is likely to correctly answer 'Who is Tom Cruise's mother?' but may fail to answer ‘Who is Mary Lee Pfeiffer’s son?'.[5]

2. Training and Data Minimization

- *At what stages of the training pipeline is it possible to implement data minimization techniques or privacy-enhancing technologies? Should we pay extra attention to the fine-tuning stage? What technical tradeoffs exist or do not exist between data minimization and model performance across different stages?*
- *What is the role of synthetic data?*
- *What is the role of (human) feedback in post-training stages, for example in enabling reinforcement learning and Retrieval-Augmented Generation (RAG)?*

Experts agreed that data minimization techniques can and should be applied throughout multiple stages of the LLM training process. Specifically, experts discussed the uses of data cleaning and synthetic data in pretraining, the use of alignment in post training, and the use of differential privacy throughout the entire training process. Experts also agreed that while there are many possible solutions for data

[4] Lee K., Ippolito D., Nystrom A., Zhang C., Eck D., Callison-Burch C., Carlini N. (2021). *Deduplicating Training Data Makes Language Models Better*. https://doi.org/10.48550/arXiv.2107.06499.
[5] Berglund L., Tong M., Kaufmann M., Balesni M., Cooper Stickland A., Korbak T., Evans O. (2023). *The Reversal Curse: LLMs trained on "A is B" fail to learn "B is A"*. https://doi.org/10.48550/arXiv.2309.12288.

minimization, those currently in use predominantly include data cleaning and alignment, which are supplemented with output filters.

A number of experts focused on addressing privacy risks in the pre-training and training stages, including “cleaning” datasets by removing obvious occurrences of personal information, as well as redundant or repetitive instances of data (deduplication) to limit risks of memorization. While there was some agreement that synthetic or artificial datasets can reduce privacy risks, experts emphasized that synthetic data alone does not guarantee privacy. Careful consideration must be given to how it's generated and implemented.

Finally, experts provided some insights into Reinforcement Learning from Human Feedback (RLHF), a machine learning technique that fine-tunes models using human feedback (such as thumbs up/down responses) as a signal to align outputs with human preferences and values, often scored against criteria of being “helpful, honest, and harmless.” Experts agreed that this process is inherently complicated as ethics are contextual, personal, and culturally informed. For example, one expert voiced concerns about collecting human feedback from low-wage workers in developing countries. The discussion included the alternative of using AI feedback for alignment, though this presents a chicken-and-egg problem of using unaligned models to achieve alignment.

Discussion Takeaways from Part 2 (Training and Data Minimization):

- One expert noted that **synthetic data** should not be used for membership inference testing (i.e., assessing whether specific data was part of a model’s training dataset through comparing confidence scores and analyzing response patterns), because a model may treat synthetic data as extra legitimate and high-quality, even more so than the original training data. This would make synthetic data unsuitable for testing whether something was part of the model's original training data.
- Among data privacy safeguards, the experts noted that **differential privacy** can protect sensitive information, but has limitations. ***(More below)***
- An expert observed current industry practices with respect to retention policies triggered by users giving thumbs up/down feedback on conversations, and the use of this feedback to give consent for use of the conversation in model training.

3. Memorization, Filters, and "Un-learning"

- *How do current techniques for detecting or measuring memorization or information leakage in language models work?*
- *What are different kinds of "un-learning" being explored in literature, and what promises vs. technical challenges or limitations do they have? (Is this expected to change over time?)*
- *What is the current state of model "reasoning" (neurosymbolic AI) and what connections does it have to model accuracy or to enabling output filtering?*
- *What role does (or can) differential privacy have in model systems?*

In the final portion of the Roundtable, experts discussed the concept of "unlearning," or the mechanisms involved in modifying an LLM to "forget" information that it has been trained on. Because of the way LLMs learn from datasets, removing information from a model is more difficult than retroactively deleting data from a data set.

Experts discussed a few methods for unlearning but agreed that most approaches remain theoretical. The most robust method of unlearning would involve complete retraining, which involves retraining a model on the initial dataset, from which the data to be "forgotten" has been fully removed. Notably, this assumes that the full scope of information that should be forgotten can be identified, which can prove challenging for large unstructured datasets with potentially large amounts of diverse kinds of information related to an individual. In addition, there was some discussion of large computational costs involved in re-training entire models.

An approach known as *sharding* aims to improve efficiency by splitting training data into multiple, disjoint "shards," training a component model on each piece of the training set, and eventually synthesizing the component models. Additionally, each small model is trained incrementally, and multiple versions of each model are stored as they are introduced to more data. Because of the scaffolded nature of the training process, a piece of training data can be removed from the larger model by retraining a single component model from an intermediate stage. While this approach may reduce the costs associated with retraining, it introduces additional complications, such as storing intermediate models and tracking what data comprises each shard. Another expert alluded to a spectrum of "**approximate unlearning**" techniques, where the goal is to modify the model's parameters to mimic a completely retrained version of the model.

At several points during the Roundtable, the group discussed the potential of **differential privacy** to mitigate risks related to LLMs. Differential privacy can be applied at many different stages of an AI system to mitigate privacy risks, and so its benefits in the context of LLMs may also be unevenly distributed. Because differential privacy requires some precision in defining the scope of information it is intended to protect, it may be more useful in fine-tuning than, for instance, in pre-training for a large unstructured dataset. At least some experts also pointed out that it could impact performance on rare but valuable knowledge, and that differential privacy applied to training data can overall lead to poorer model performance. Specifically, there was some discussion that differential privacy may not be an immediately applicable solution for memorization in LLMs, given LLMs' training on unstructured data and

reliance on forms of memorization to make predictions about infrequent uses of language. A frequent theme throughout the conversation was the challenge of balancing privacy guarantees with maintaining model utility.

Finally, neurosymbolic AI was mentioned in the conversation as a potentially promising area of AI research. It focuses on training models to have better "understanding" about relationships in data through reasoning.

Discussion Takeaways from Part 3 (Memorization, Filters, and "Un-Learning") and Final Wrap-Up Conversation:

- Different methods for "**unlearning**" exist and represent an active field of emerging research in technical literature. Many remain high-cost and theoretical to implement in practice.
- Some experts advocated for a user-centric approach, including encouraging the development of tools and APIs for querying models about personal information. This could enable automated privacy reports and adjustment mechanisms.
- **Differential privacy** may or may not offer solutions for data minimization or memorization in LLMs, given the potential tradeoffs with model utility for rare but valuable knowledge and uses of language.
- Current privacy protection approaches face significant tradeoffs - while differential privacy offers formal guarantees specifically during fine-tuning, it degrades model performance on rare but valuable knowledge during pre-training and mid-training; synthetic data can lead to unexpected model behaviors and overconfidence; and though unlearning could enable targeted data removal, it remains expensive for large models.
- User privacy controls and consent mechanisms need careful consideration, as demonstrated by concerns over using feedback mechanisms as implicit consent for long-term data retention.
- Industry practices may face growing privacy challenges as the scarcity of high-quality open datasets pushes companies toward increased use of user data for training.

Did we miss anything? Reach out to us at ai@fpf.org.

Appendix 1. LLM Vocabulary Mapping: From Natural Text to Tokens

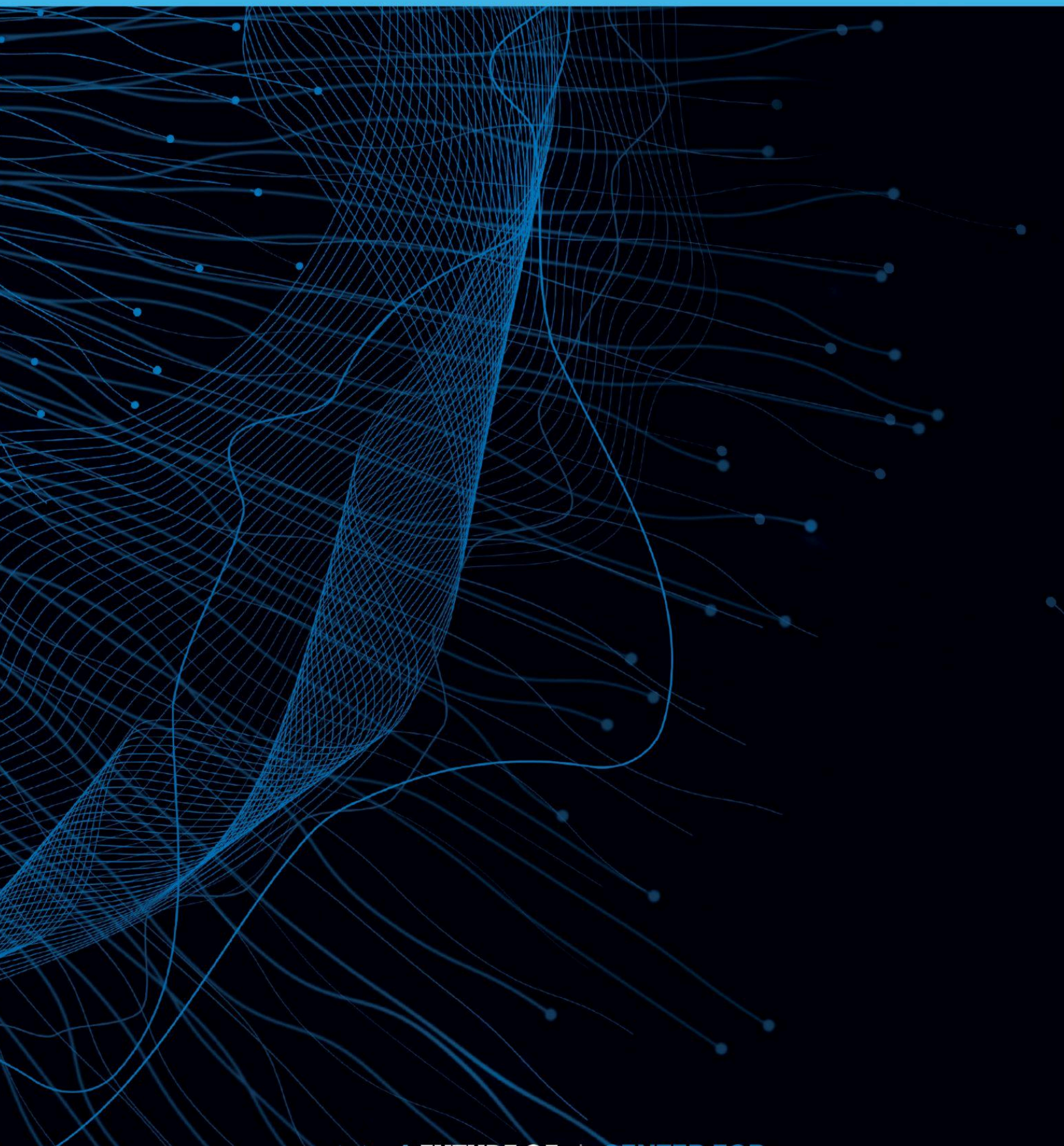
We can actually inspect the dictionary of an LLM. To demonstrate this, the authors decoded the content of Google's Gemma 7b model (version 1.1). Our tool 'List Tokens' analyzes and displays tokens from the LLM model file's vocabulary section.[6] In the table below, we list the first 11 tokens in the LLM's vocabulary, followed by the 18 tokens of the answer to our prompt.

#	Offset	Len	Raw Hex	Token Text
0	0x00000dad	4	e2 96 81 21	_!
1	0x00000dd6	4	e2 96 81 24	_$
2	0x00000e02	4	e2 96 81 27	_'
3	0x00000e31	4	e2 96 81 2a	_*
4	0x00000e63	4	e2 96 81 2d	_-
5	0x00000e98	4	e2 96 81 30	_0
6	0x00000ed0	4	e2 96 81 33	_3
7	0x00000f0b	4	e2 96 81 36	_6
8	0x00000f49	4	e2 96 81 39	_9
9	0x00000f8a	4	e2 96 81 3c	_<
10	0x00000fce	4	e2 96 81 3f	_?
(...)				
15492	0x0007843d	10	e2 96 81 50 6c 61 6e 74 73 0a	_Plants
340	0x0000525a	8	e2 96 81 6e 65 65 64 06	_need
1241	0x0000c6ed	11	e2 96 81 73 65 76 65 72 61 6c 04	_several
3586	0x0001d77a	13	e2 96 81 65 73 73 65 6e 74 69 61 6c 05	_essential
3638	0x0001dd51	13	e2 96 81 63 6f 6d 70 6f 6e 65 6e 74 73	_components
42	0x0000285c	6	e2 96 81 74 6f 06	_to
783	0x00008da3	8	e2 96 81 67 72 6f 77 04	_grow
148	0x00003558	5	e2 96 81 3a 03	_:
15308	0x00076e89	12	e2 96 81 73 75 6e 6c 69 67 68 74 08	_sunlight
415	0x00005cc5	5	e2 96 81 2c 02	_,
520	0x00006c1d	9	e2 96 81 77 61 74 65 72 07	_water
415	0x00005cc5	5	e2 96 81 2c 02	_,
43	0x00002869	7	e2 96 81 61 6e 64 03	_and
17139	0x00085000	13	e2 96 81 6e 75 74 72 69 65 6e 74 73 07	_nutrients
127	0x00003142	8	e2 96 81 66 72 6f 6d 03	_from
39	0x00002829	7	e2 96 81 74 68 65 03	_the
3418	0x0001c300	8	e2 96 81 73 6f 69 6c 07	_soil
177	0x00003988	5	e2 96 81 2e 03	_.

```
Analyzing file: /usr/share/ollama/.ollama/models/blobs/sha256-ef311de6af9db0
43d51ca4b1e766c28e0a1ac41d60420fed5e001dc470c64b77
File size: 5,011,844,064 bytes
```

The first column (#) displays the position of a token, i.e., the exclamation mark is the first token we found, followed by the ampersand in the next row. Words that are used frequently in the vocabulary section of the Gemma LLM have a lower ranking in the vocabulary in comparison to words that we use less frequently. For example, we found 'to' at position 42 and 'the' at position 39 in contrast to, e.g., the word 'nutrients' with a position of 17,139. The second column (Offset) indicates the relative position of the token text in the file, followed by the number of hexadecimal characters (Len), the hexadecimal representation (Raw Hex), and in the last column (Token Text) the decoded text of the token preceded by a marker character '_'.

[6] See, https://github.com/rvaneijk/list-tokens.



FUTURE OF PRIVACY FORUM | CENTER FOR ARTIFICIAL INTELLIGENCE

Washington, DC | Brussels | Singapore | Tel Aviv

info@FPF.org | FPF.org