Combating Hallucination and Misinformation: Factual Information Generation with Tokenized Generative Transformer

Sourav Das, Sanjay Chatterji, and **Imon Mukherjee**
Indian Institute of Information Technology Kalyani

Joint 3rd International Conference on Natural Language Processing for Digital Humanities & 8th International Workshop on Computational Linguistics for Uralic Languages **(NLP4DH & IWCLUL 2023)**

November 28, 2023

Layout

Introduction

- Language generation with Large language models (LLM) has become paradigm-shifting research in Natural language processing (NLP).
- LLMs have demonstrated outstanding abilities in a multitude of tasks.

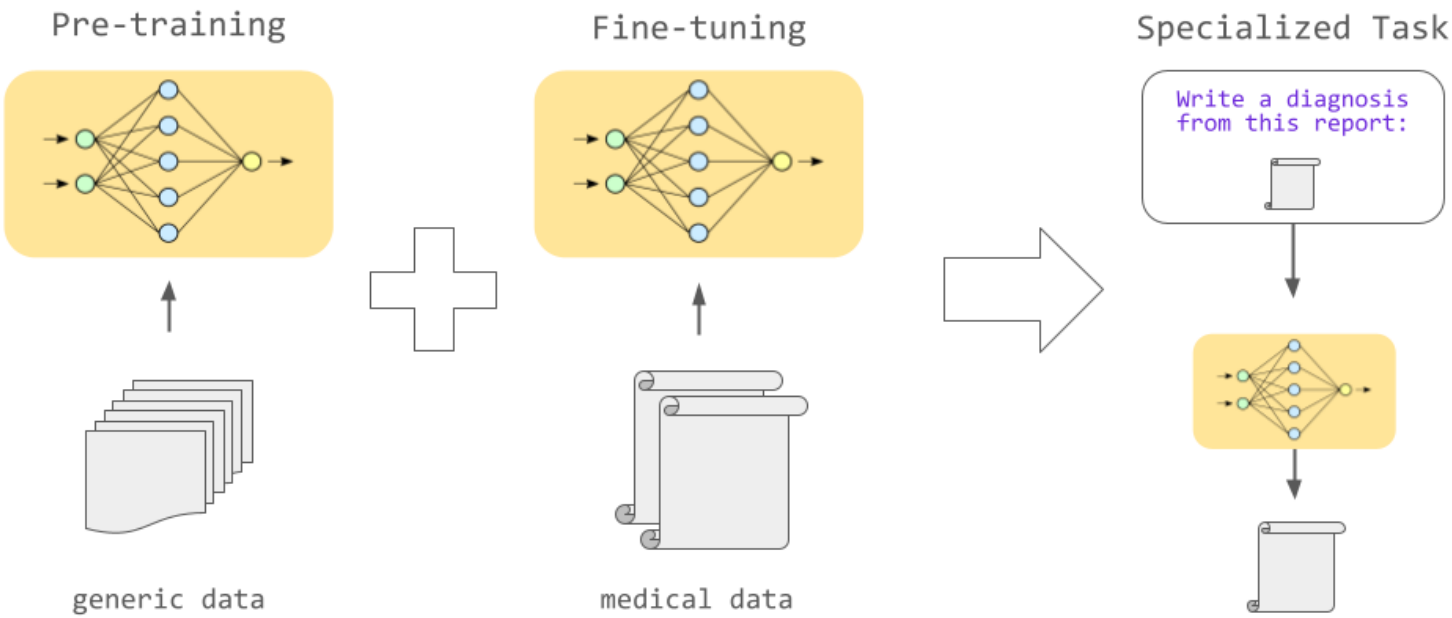


Fig. 1: Simplified working representation of LLMs. From **AssemblyAI**.

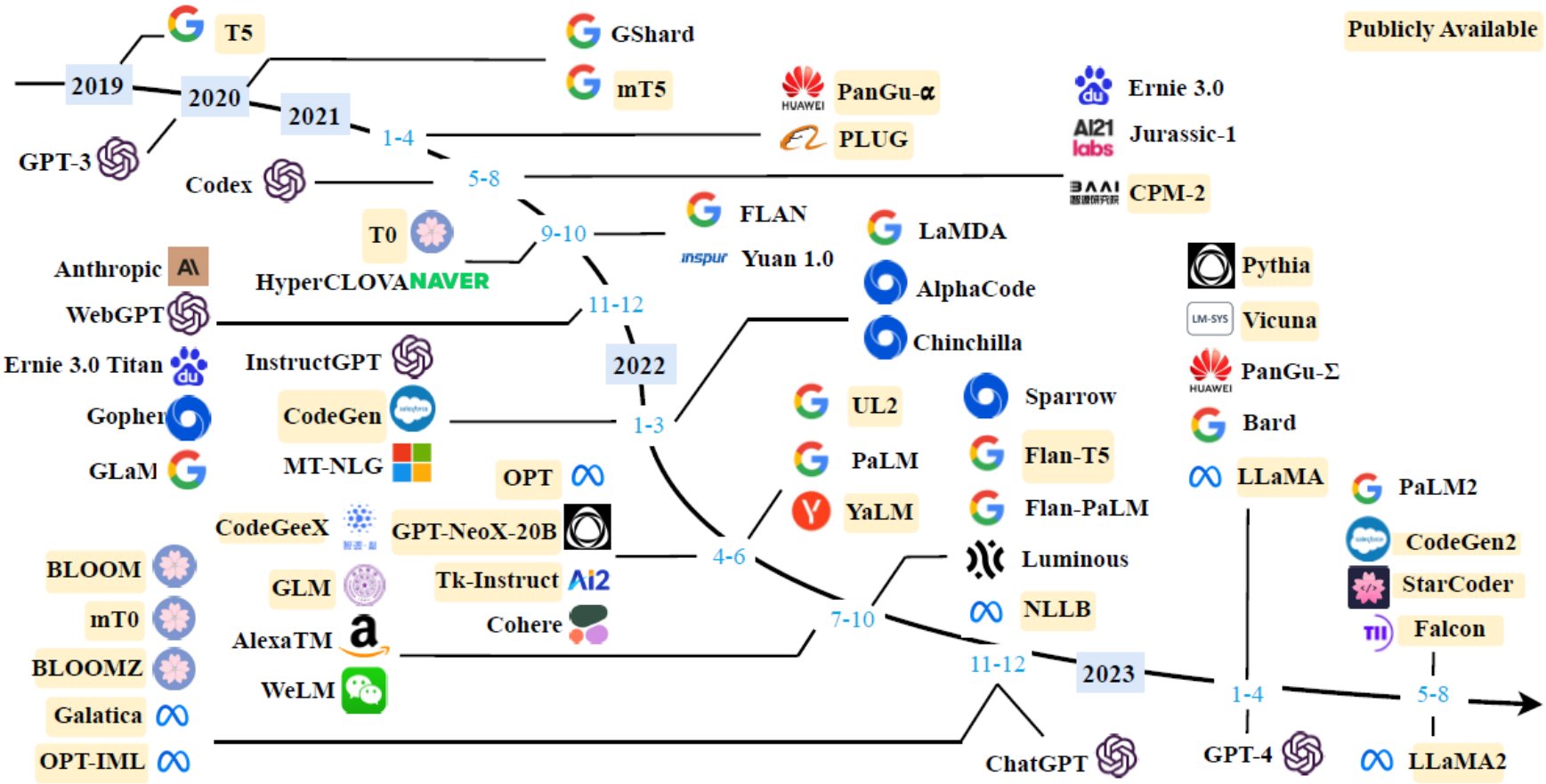
Introduction

- LLMs are characterized by their large size in terms of the number of parameters or weights they contain, which enables them to perform a wide range of natural language processing tasks with impressive accuracy.

- To discriminate the language models in different parameter scales, the research community has coined the term Large Language Models (LLM) for the models of significant size, with billions or even trillions of parameters [1].

Current State of LLMs

A timeline of existing prime Large Language Models [7] (up-to-date till November 2023). The timeline is shown according to the availability date of the paper for a model.



Why Factual Information Generation?

In numerous LLM-based language generation cases, a critical challenge arises when engaging conversational agents. Input prompts frequently result in misunderstood **topics** and **semantics**, leading to logical voids in the conversation.

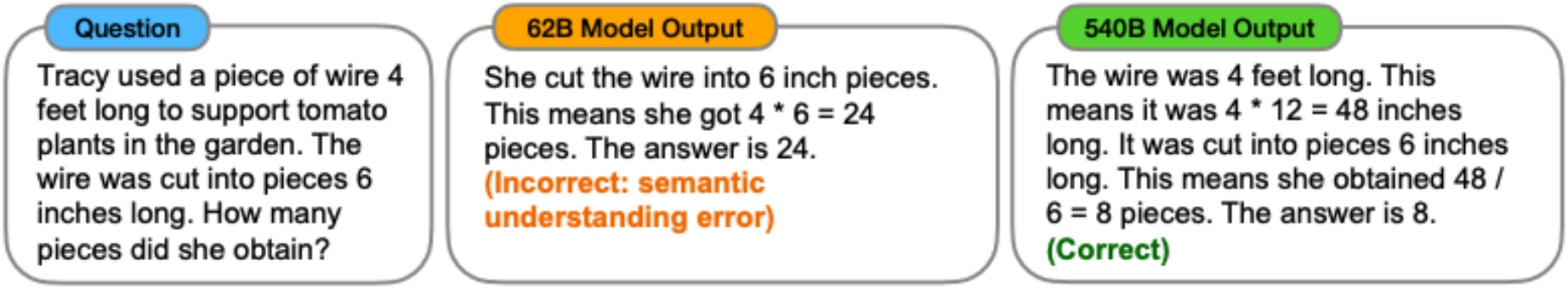


Fig. 2: A comparison between wrong and correct answer generated by LLM: From **Google Research**.

Problem Statement: Hallucination

- With the logical voids in generated languages from LLMs, scrutiny of their limitations and risks are also under the radar. [2]
- These approaches can lead to degeneration, characterized by bland, incoherent, or repetitive output. language generation often produces nonsensical or unfaithful text, known as **hallucinations**.

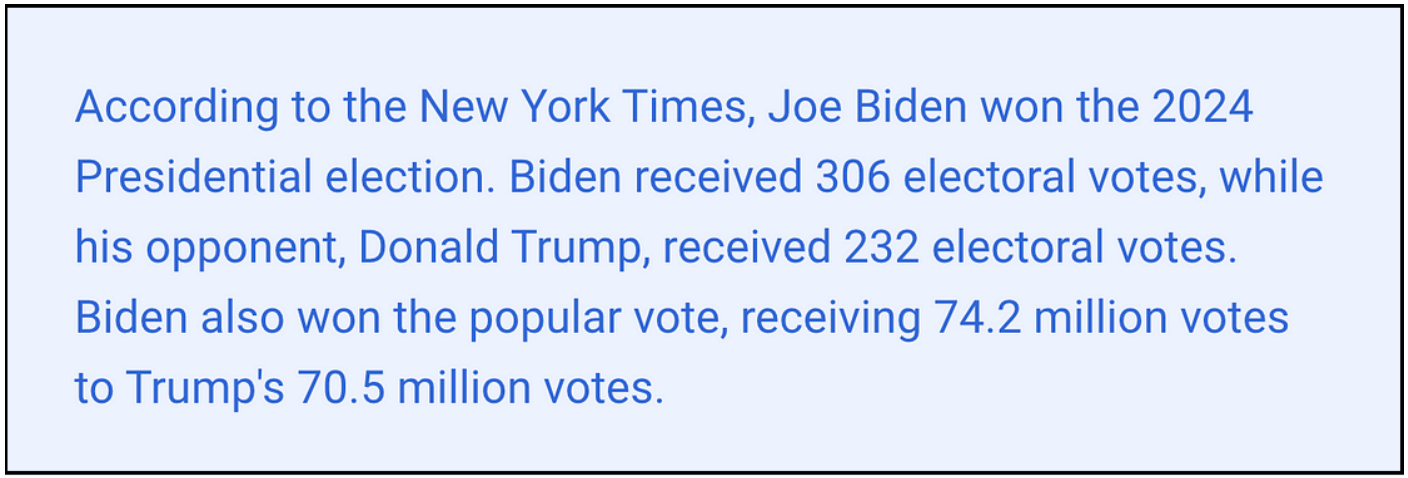


Fig. 3: Hallucinated text generated by **Google Bard**: Published on **Google Cloud Community**.

Problem Statement: Misinformation

- The growing accessibility of LLMs and their enhanced capacity to produce credibly-sounding text also raise concerns regarding their potential misuse for generating **misinformation** [3].
- The deliberate distribution of misinformation can lead to significant societal harm, including the manipulation of public opinion and the promotion of detrimental ideologies.

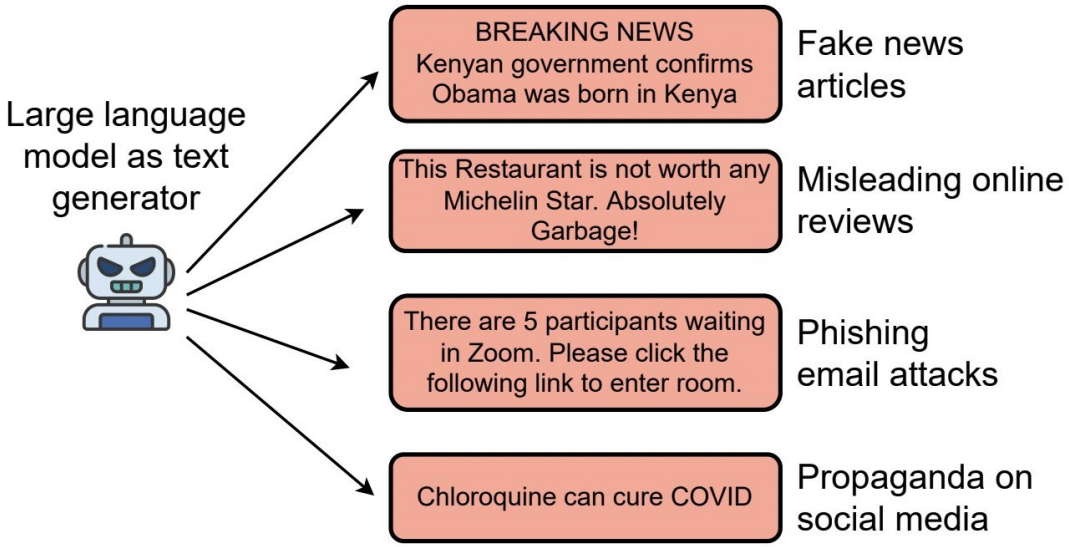


Fig. 4: Potential misinformation generation by LLMs: by Pu et al. (2022) [4].

Proposed Framework: Factual Information Generation with Tokenized Generative Transformer

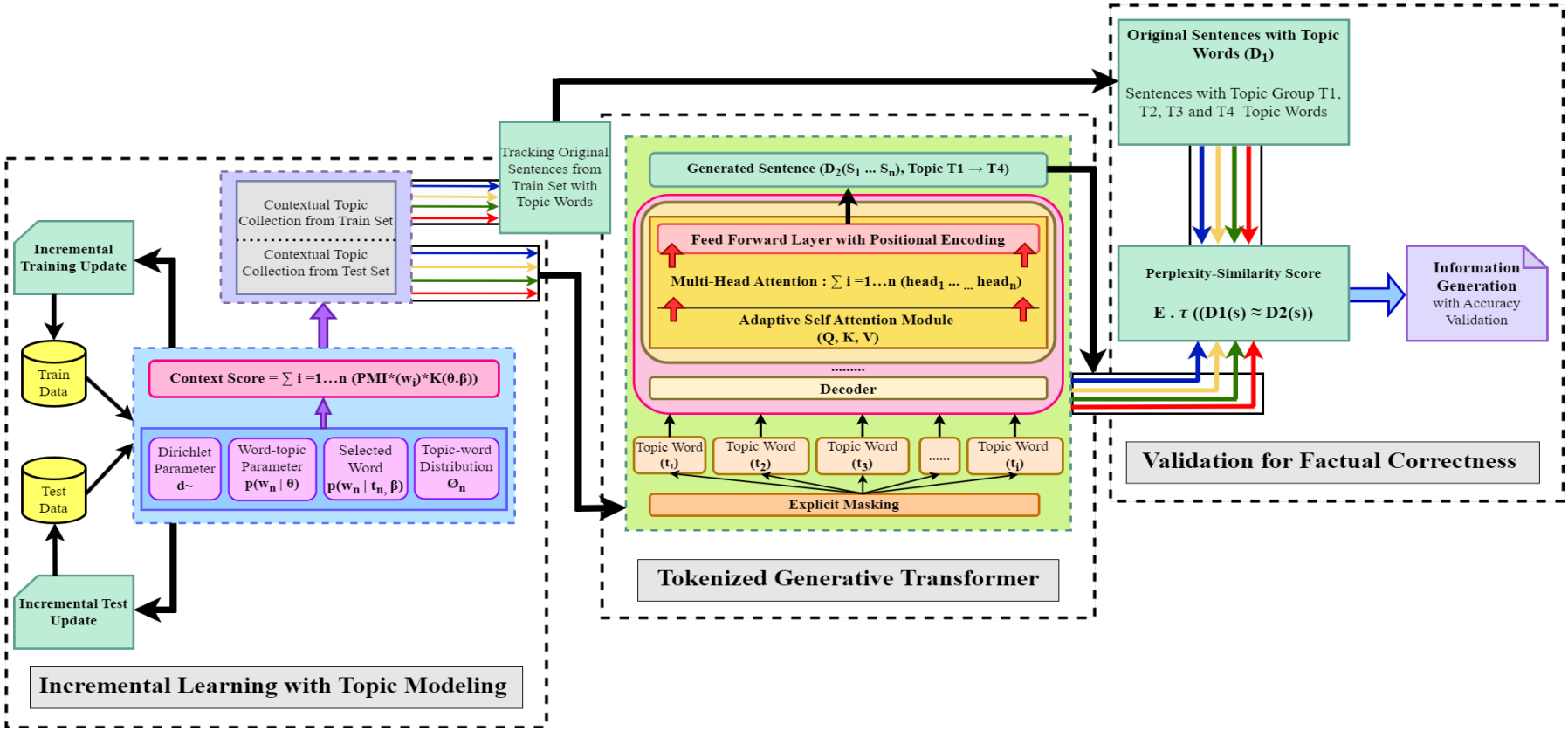


Fig. 5: Combating hallucination and misinformation generation: Overview of the system framework for contextual topic modeling for tokenized generative transformer-based information generation.

Contributions

Fundamental phases of the experiment:

- **Incremental learning** from a collection of large research literature from arXiv,
- **Contextual LDA (Co-LDA)** for topic modeling,
- Information (in the form of sentences) generation using **Tokenized Generative Transformer (TGT)**,
- **Information validation** parameters,
- **Perplexity Similarity Score** generation,
- **Benchmark evaluation** (Gold standard corpus + SoTA models).

Incremental Learning

Training Corpus:

5000 research papers were collected to construct the train set from the **arXiv** repository, focusing on COVID-19 research published between March 2020 to July 2023.

- When new literature is fetched during the API call, the model can be adjusted to learn from the extracted topics without having to completely retrain it from scratch.

- We assume to represent DataFrame($\mathcal{L}$) as a collection of data points:

$$\text{DataFrame}(\mathcal{L}) = \{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\} \tag{1}$$

Incremental Learning

- Key challenge is to adapt the model's parameters to the new data without forgetting the knowledge gained from the old data.
- We aim to minimize a loss function J that measures the difference between the model's predictions and the baseline parameters.
- Loss determined on the old data:

$$J_{\text{old}}(\Phi) = \sum_{i=1}^{N_{\text{old}}} \ell(f(x_i; \Phi), y_i) \tag{2}$$

- Loss determined on the new data:

$$J_{\text{new}}(\Phi) = \sum_{i=1}^{N_{\text{new}}} \ell(f(x_i; \Phi), y_i) \tag{3}$$

Incremental Learning

- The overall objective in *incremental learning* is to find a set of updated parameters Φ^* that minimize the combined loss:

$$\Phi^* = \arg\min_{\Phi} \left(\alpha J_{\text{old}}(\Phi) + (1-\alpha) J_{\text{new}}(\Phi)\right) \tag{4}$$

- α is a hyperparameter that controls the balance between preserving knowledge from the old contextual topics (J_{old}) and adapting to the new contextual topics (J_{new}).

Contextual Topic Modeling

- For the improved topic modeling with context, we propose the Latent Dirichlet allocation embedded with **Context Scores** for emphasizing contextuality in extracting meaningful topics from the developed corpus.
- We call this scheme the **Contextual LDA (Co-LDA)**.
- Four Topic Domains or Groups are observed and derived from this method, corresponding to
  T1: Medical Topic,
  T2: Social Topic,
  T3: Research Topic, and
  T4: Generic Topic.
- The labeling aids in computing context scores for different domains.

Contextual Topic Modeling

- In traditional LDA:

$$\alpha \sim \text{Dirichlet}(\alpha_1, \alpha_2, \ldots, \alpha_K)$$

and

$$\beta \sim \text{Dirichlet}(\beta_1, \beta_2, \ldots, \beta_V)$$

- Additionally, We utilize the **Pointwise Mutual Information (PMI)** metric to perform this. Higher PMI scores indicate more contextual and meaningful topics.
- The context score for a set of topics is computed from the PMI score as follows:

$$\textit{Context Score} = \sum_{i=1\ldots n}^{N} (PMI(w_i) * K(\theta.\beta)) \tag{5}$$

Contextual Topic Modeling

Topic Groups	Topic Words
Topic 1: Medical	pandemic, epidemic, vaccine, GSA, virus, health, disease, infected, booster, death
Topic 2: Social	social, distance, isolation, lockdown, migration, remote, online, curfew, mask, sanitizer
Topic 3: Research	dataset, measures, count, model, analysis, prediction, simulation, optimization, spread, results
Topic 4: Generic	approach, crisis, initiatives, education, precaution, spread, resilience, transport, efficiency, paper

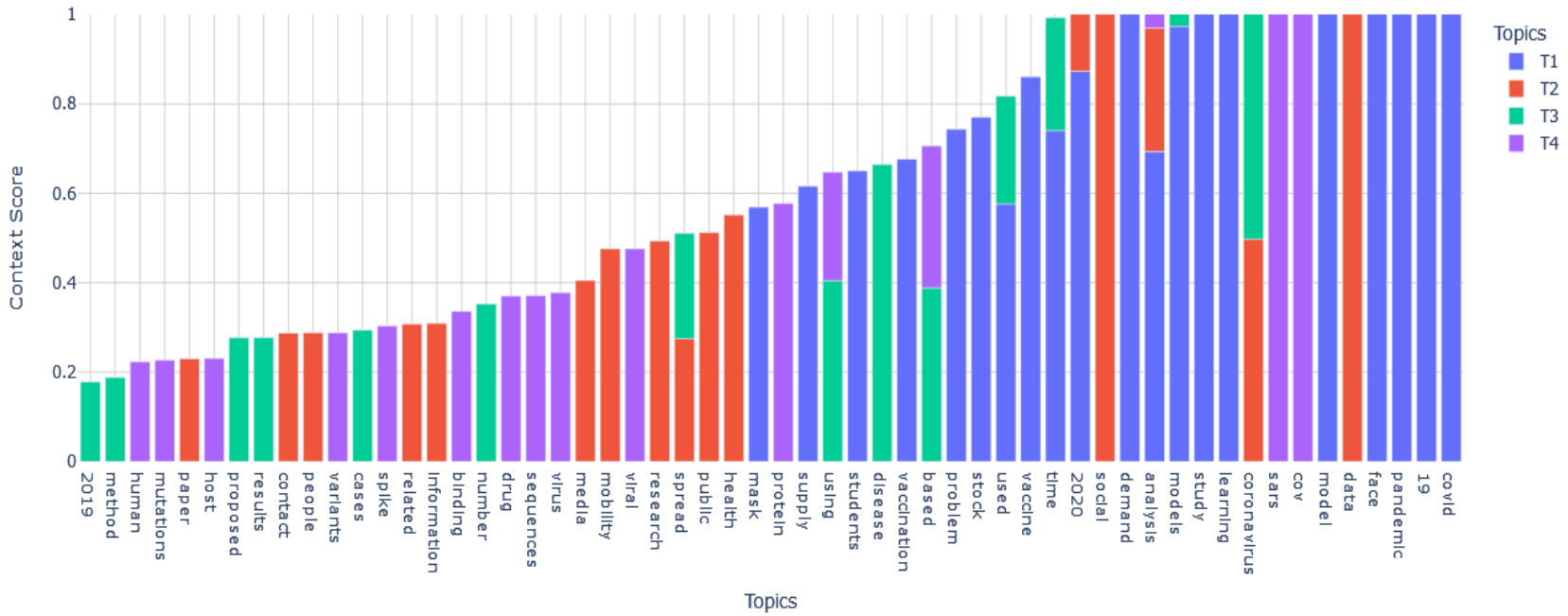


Fig. 6: Topic words arranged by Context Scores, with overlapping colors indicating shared words across groups, ensuring contextual sensitivity in proposed topic modeling.

Text Generation

- We utilize the GPT-3 language model as the platform for developing the Tokenized Generative Transformer model.
- After masking the topic words, the model imposes the self-attention mechanism for normalizing the Context Scores for further computation.
- The masking for self-attention is performed using three trained linear projections: Query (Q), Key (K), and Value (V). The attention scores are generically calculated as follows:

$$\textit{Self-Attention}(Q,K,V) = \textit{softmax}\frac{(QK^T)}{(\sqrt{d_k})} \cdot V \tag{6}$$

Text Generation

- The multi-head attention mechanism computes multiple attention scores in parallel, allowing the model to attend to different parts of the input sequence simultaneously.

- The outputs of the attention heads are concatenated and linearly transformed to produce the final output:

$$\textit{Multi-Head}(Q,K,V) = \sum_{i=1,\ldots,n}^{N} \text{Head} : (h_1, \ldots, h_n) \cdot W^O \tag{7}$$

Text Generation

- To evaluate the quality of generated sentences, we first calculate the accuracy of the 100 sentences generated for each topic:

$$Topic\text{-}Acc(T_i) = \frac{(Original_{Sentences}\ T_i)}{(Generated_{Sentences}\ T_i\ (100))} \tag{8}$$

- Consecutively, we check, within the generated 100 sentences for a topic, what percentage contain topic words from the corresponding extended list. Here T_i represents topic $T1$ to $T4$:

$$\bar{CM}_{Accuracy} = \frac{\sum_{i=1}^{n} Topic\_Acc(T_i)}{4} \tag{9}$$

- The cumulative mean accuracy is obtained for all the generated sentences for topic $T1$ to $T4$ in terms of simultaneous sentence generation per topic.

Topic-wise Generated Information

Generated sentence for each topic word:

Index	Topic Word	Generated Sentence Corresponding to Each Topic Word
		Topic T1: Medical Topic; Topicwise Accuracy: 0.88
1.	**pandemic**	**'Pandemic'** is a phenomenon when a large number of people are infected with the virus.
2.	**epidemic**	The **'epidemic'** is the number of large contaminations that are detected in a given period.
3.	**vaccine**	Pfizer was the first coronavirus **'vaccine'**.
		Topic T2: Social Topic; Topicwise Accuracy: 0.85
1.	**social**	**'Social'** is a term that describes the social interaction of individuals.
2.	**distance**	The **'distance'** between two points on a two-dimensional coordinate is Euclidean.
3.	**isolation**	WHO reports **'isolation'** as one of the core reasons for depression.
		Topic T3: Research Topic; Topicwise Accuracy: 0.87
1.	**dataset**	Many Covid-19 related **'datasets'**, tools, and software are created and shared.
2.	**measures**	**'Measures'** is the number of steps that can be taken to achieve the desired outcome.
3.	**count**	Daily **'count'** of the Covid-19 cases was at an all-time high in the first quarter of 2020.
		Topic T4: Generic Topic; Topicwise Accuracy: 0.78
1.	**approach**	The **'approach'** is a simple, straightforward, and cost-effective way to reduce the cost.
2.	**crisis**	A setback is not a **'crisis'**, but a scope for analytical examination.
3.	**initiatives**	All the necessary **'initiatives'** have been taken to slow down the rate of transmission.

Table 2: Topic wise information generation.

Information Validation Parameters

- For verifying the authenticity of the generated information, comparison with original information containing the same topic(s) is crucial.
- To do that, the semantic evaluation is necessary. It can be derived from semantic embedding between the comparable information (sentences).
- To fully utilize the semantic embedding, we propose the **Perplexity-Similarity Score** to achieve the similarity between the comparing documents to understand the complex similarity or polarity structure of the sentences within.

Information Validation Parameters

- Next, the multi-document summarization is performed for encoded tokenized representation of the comparable documents.
- This is done because of optimization and efficiency, to tackle the bottleneck of the system while comparing numerous sentences at each step:

$$E = Comp(\theta^{(D_1 \rightarrow D_2 \ldots)}) \times (s(T \rightarrow t_1, \ldots, t_n)) \tag{10}$$

- The encoded input for any sentence s can be represented as a matrix X of size $n \times t$, where n is the number of tokens and t is the dimension of the token embeddings.
- Formally, the Perplexity-Similarity Score can be proposed:

$$Per\text{-}Sim\ Score = E \cdot \tau(D_1(s) \approx D_2(s)) \tag{11}$$

Benchmarking

We evaluate the performances of a few *comparable* SoTA models against our proposed framework on the test corpus **COVID-19 Open Research Dataset (CORD-19)** [5]:

- BERT-Base,
- BERTGeneration,
- T5-Base,
- LLama-13B,
- MPT-7B,
- GPT Neo.

Benchmarking

Models	ROUGE		METEOR		BLEU		Per-Sim Score	
	Train	Test	Train	Test	Train	Test	Train	Test
T5-11B [6]	0.75	0.72	0.80	0.78	0.85	0.83	0.85	0.85
LLaMA-13B [7]	0.78	0.76	0.82	0.80	0.85	0.82	0.87	0.86
MPT-7B [8]	0.76	0.74	0.81	0.78	0.83	0.81	0.84	0.83
GPT Neo-20B [9]	0.79	0.77	0.83	0.81	0.87	0.85	0.89	0.89
CoLDA+TGT (Ours)	**0.82**	**0.80**	**0.86**	**0.84**	**0.87**	**0.86**	**0.93**	**0.91**

Table 2: Benchmarking of our proposed framework with transformer-based state-of-the-art language models on the open corpus CORD-19.

Future Prospects

- To integrate domain-specific ontologies and relational knowledge base for enhancing contextual relevance in LLMs.
- To analyze the domains and underlying topics in LLM-based text generation to improve factual consistency and mitigate hallucination issues.
- To conduct further research to make the developed model more efficient at multitask learning and multi-view capabilities.

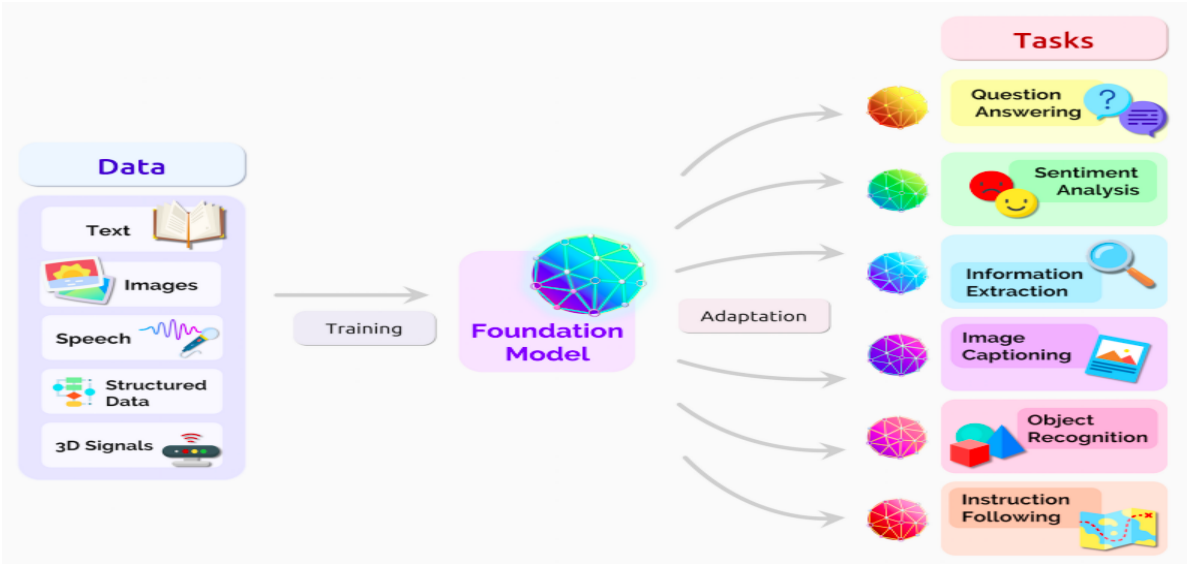


Fig 7. A visual overview of how LLM could be used as a foundation for multitask learning.

From NVIDIA Blog.

References

1. Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., ... & Wen, J. R. (2023). A survey of large language models. arXiv preprint arXiv:2303.18223.
2. Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., ... & Fung, P. (2023). Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12), 1-38.
3. Pan, Y., Pan, L., Chen, W., Nakov, P., Kan, M. Y., & Wang, W. Y. (2023). On the Risk of Misinformation Pollution with Large Language Models. arXiv preprint arXiv:2305.13661.
4. Pu, J., Sarwar, Z., Abdullah, S. M., Rehman, A., Kim, Y., Bhattacharya, P., ... & Viswanath, B. (2022, October). Deepfake Text Detection: Limitations and Opportunities. In 2023 IEEE Symposium on Security and Privacy (SP) (pp. 19-36). IEEE Computer Society.

References

5. Wang, L. L., Lo, K., Chandrasekhar, Y., Reas, R., Yang, J., Burdick, D., ... & Kohlmeier, S. (2020, June). CORD-19: The COVID-19 Open Research Dataset. In ACL 2020 Workshop on Natural Language Processing for COVID-19 (NLP-COVID).
6. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ...& Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. The Journal of Machine Learning Research, 21(1), 5485-5551.
7. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M. A., Lacroix, T., ... & Azhar, F. Llama: Open and efficient foundation language models. arXiv 2023. arXiv preprint arXiv:2302.13971.
8. MosaicML NLP Team. 2023. Introducing mpt-7b: A new standard for open-source, commercially usable llms. 2023.
9. Black, S., Gao, L., Wang, P., Leahy, C., & Biderman, S. (2021). Gpt-neo: Large scale autoregressive language modeling with mesh-tensorflow.

Any Questions?

Thank You.