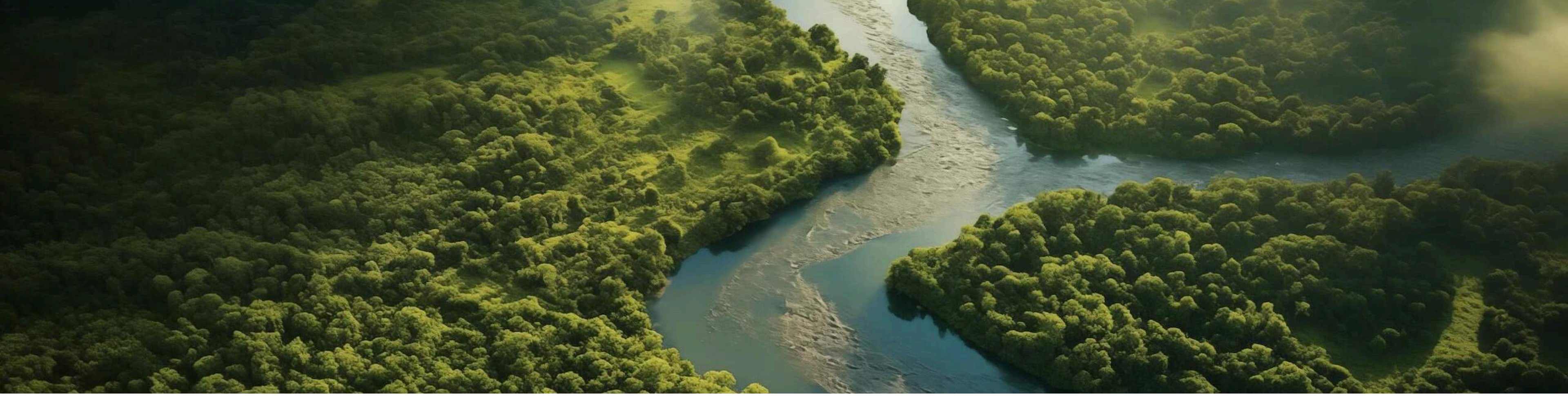


Assessing and Improving ClimateGPT's Faithfulness in Retrieval Augmented Generation

David Thulke

Oral Presentation at the Nordic Workshop on AI for Climate Change 2025

Joint Work with:

Jakob Kemmler, Christian Dugast and Hermann Ney
and many others from AppTek, Erasmus.AI and EqtyLab as part of the ClimateGPT project

Outline

Development of ClimateGPT

ClimateGPT: Towards AI Synthesizing Interdisciplinary Research on Climate Change

January, 2024

Assessing and Improving Faithfulness

Listen to the Context: Towards Faithful Large Language Models
for Retrieval Augmented Generation on Climate Questions

ClimateNLP 2025

Outline

Development of ClimateGPT

ClimateGPT: Towards AI Synthesizing Interdisciplinary Research on Climate Change

January, 2024

Assessing and Improving Faithfulness

Listen to the Context: Towards Faithful Large Language Models
for Retrieval Augmented Generation on Climate Questions

ClimateNLP 2025

Motivation

ClimateGPT: A climate-specific large language model

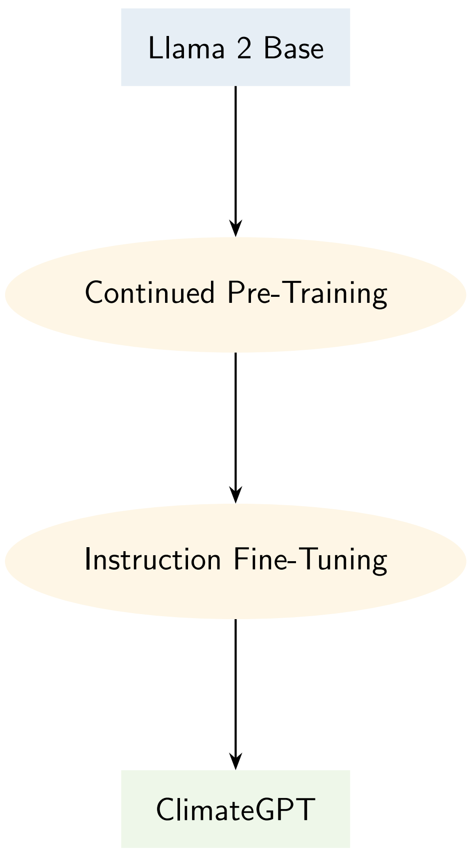
- Climate NLP Applications:
  - Make climate science and policy more accessible - e.g. ClimateQ&A[1], ChatReport[2], Climate Policy Radar[3]
  - Climate misinformation - e.g. Climinator (Leippold et al., 2024)
  - Extraction and analysis of climate data - e.g. Climate Impact Events (Li et al., 2024)
- Usually, large general-purpose models are used
- Other fields: domain-specific models can outperform general-purpose models
  - e.g. Science (Taylor et al., 2022), Medicine (Chen et al., 2023), Finance (Wu et al., 2023)

- ClimateGPT Project: Development of a climate-specific large language model
  - Main project time: September - November 2023
  - Prototype application: multilingual Q&A system demoed at COP28 in December
  - Partners: AppTek, Erasmus.AI, EqtyLab

[1] climateqa.com
[2] reports.chatclimate.ai
[3] climatepolicyradar.org

Development of ClimateGPT

Method



Continued Pre-Training

- Large corpus of climate-related documents from Erasmus.AI (3.2B words)
- Consisting of scientific papers, reports and news articles
- Goal: Improve the domain knowledge of the system

Instruction Fine-Tuning

- Pairs of input and expected model output
- Large set of climate-related fine-tuning data generated by working with domain experts and annotators
- Additional publicly available general domain data
- Goal: Instruction-following capabilities on climate-related tasks

Development of ClimateGPT

Instruction Fine-Tuning

Data Collection with 10 experts in climate-related fields and 99 non-experts from AppTek's data team.

Types of instructions:

- Closed-Ended: Instructions on a given documents, e.g. summarisation, extraction, etc.
- Open-Ended: Instructions for which the model is expected to “know” the answer, e.g. questions, brainstorming, etc.
  - Grounded: Annotators provided sources for outputs. To simulate RAG conditions, these sources were added in training as additional context.

Development of ClimateGPT

Evaluation Tasks

Few-shot evaluation on climate-specific NLP tasks

Datasets	Subset	Task Type
ClimaBench	ClimaText	Topic Classification
	ClimateStance	Sentiment Analysis
	ClimateEng	Topic Classification
	ClimateFever	Fact Checking
	CDP-QA	QA-Pair Prediction
Pira 2.0	MCQ	Multiple-Choice QA
CARDS	Binary	Misinformation Classification

+ Commonly used general-domain benchmarks (e.g. MMLU)

Development of ClimateGPT

Results on climate-specific benchmarks

Model	ClimaBench	Pira 2.0 MCQ	CARDS	Weight. Avg.
Llama-2-Chat-7B	67.8	72.0	64.3	68.5
ClimateGPT-7B	75.3	86.6	65.9	77.1
w/o CPT	72.7	85.7	66.5	75.3
Llama-2-Chat-70B	72.7	88.8	72.5	77.0
ClimateGPT-70B	72.4	89.9	73.4	77.2

Accuracies on the climate-specific benchmarks.

No degradations on general-domain benchmarks

Limited human evaluation results confirm these trends

Models are available on `huggingface.co/eci-io`

Outline

Development of ClimateGPT

ClimateGPT: Towards AI Synthesizing Interdisciplinary Research on Climate Change

January, 2024

Assessing and Improving Faithfulness

Listen to the Context: Towards Faithful Large Language Models
for Retrieval Augmented Generation on Climate Questions

ClimateNLP 2025

Assessing and Improving Faithfulness

Faithfulness

- Current LLMs, including ClimateGPT, are prone to hallucinating facts
- Retrieval Augmented Generation (RAG) systems aim to ground answers in retrieved context to reduce hallucinations
- **But:** Does the model truly adhere to the retrieved context?
- Faithfulness evaluates whether each generated claim is *supported* by the provided context
- **Why care?**
  - All information that goes beyond the provided context might be hallucinated
  - Allows to trace back claims to the original source
- **Faithfulness vs. Factuality:**
  - Faithfulness: Does output reflect the *retrieved context*?
  - Factuality: Is output correct w.r.t. broader *world knowledge*?
- **How to measure faithfulness?**
  - Automatic method based on RAGAs (Es et al., 2024)

Assessing Faithfulness

Prompt
What is the purpose of the Global Stocktake?

Input:
prompt x

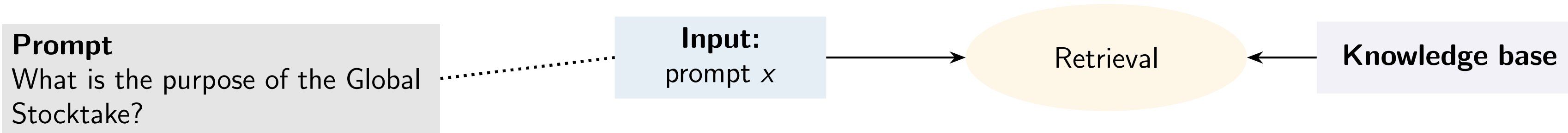
Assessing Faithfulness

Prompt
What is the purpose of the Global Stocktake?

Input:
prompt x

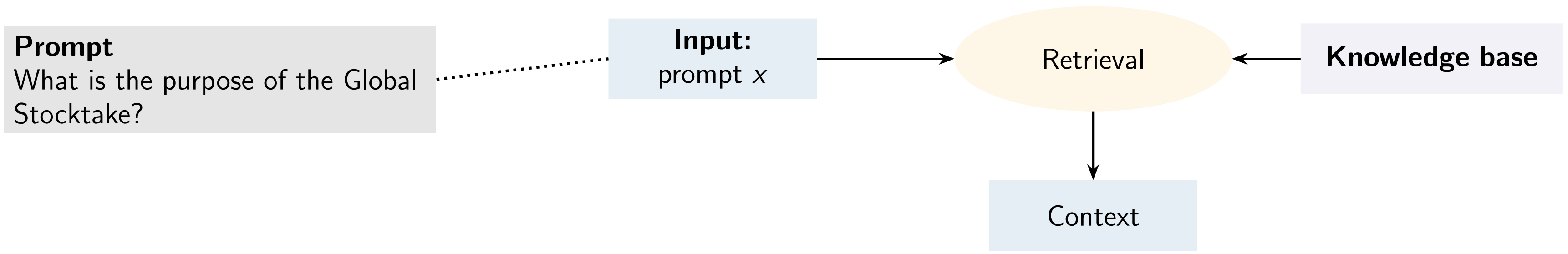
Assessing and Improving Faithfulness

Assessing Faithfulness



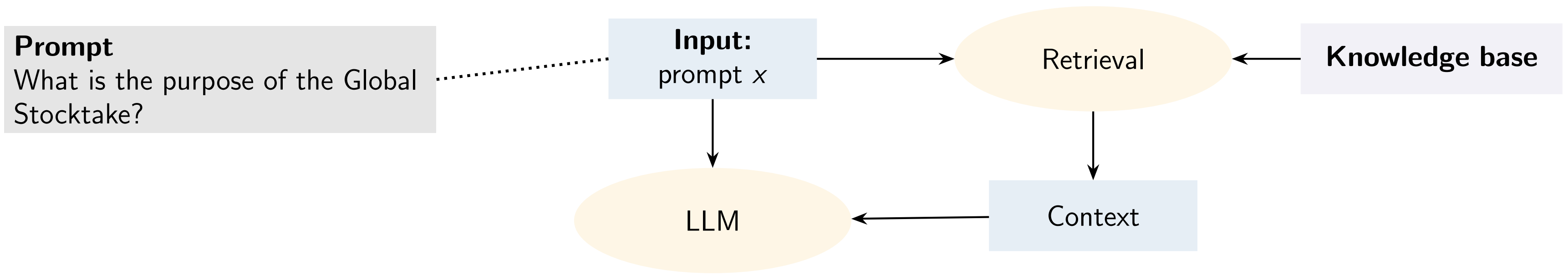
Assessing and Improving Faithfulness

Assessing Faithfulness



Assessing and Improving Faithfulness

Assessing Faithfulness

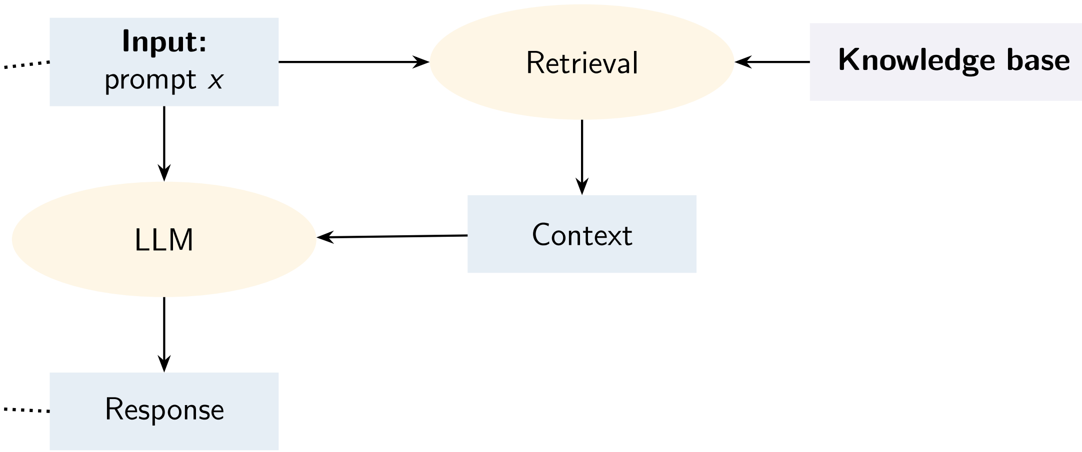


Assessing Faithfulness

Prompt
What is the purpose of the Global Stocktake?

Original response

The Global Stocktake is a key component of the Paris Agreement, scheduled to begin in 2023. It involves a comprehensive assessment of global progress in addressing climate change, considering scientific evidence, economic [...]. The Stocktake aims to [...]

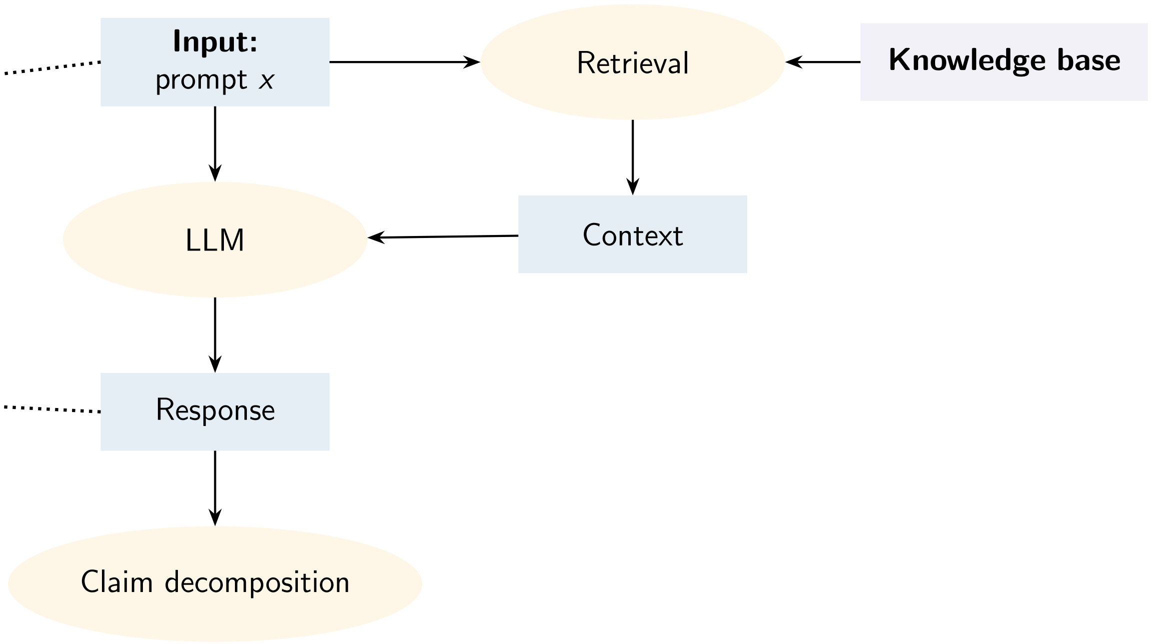


Assessing Faithfulness

Prompt
What is the purpose of the Global Stocktake?

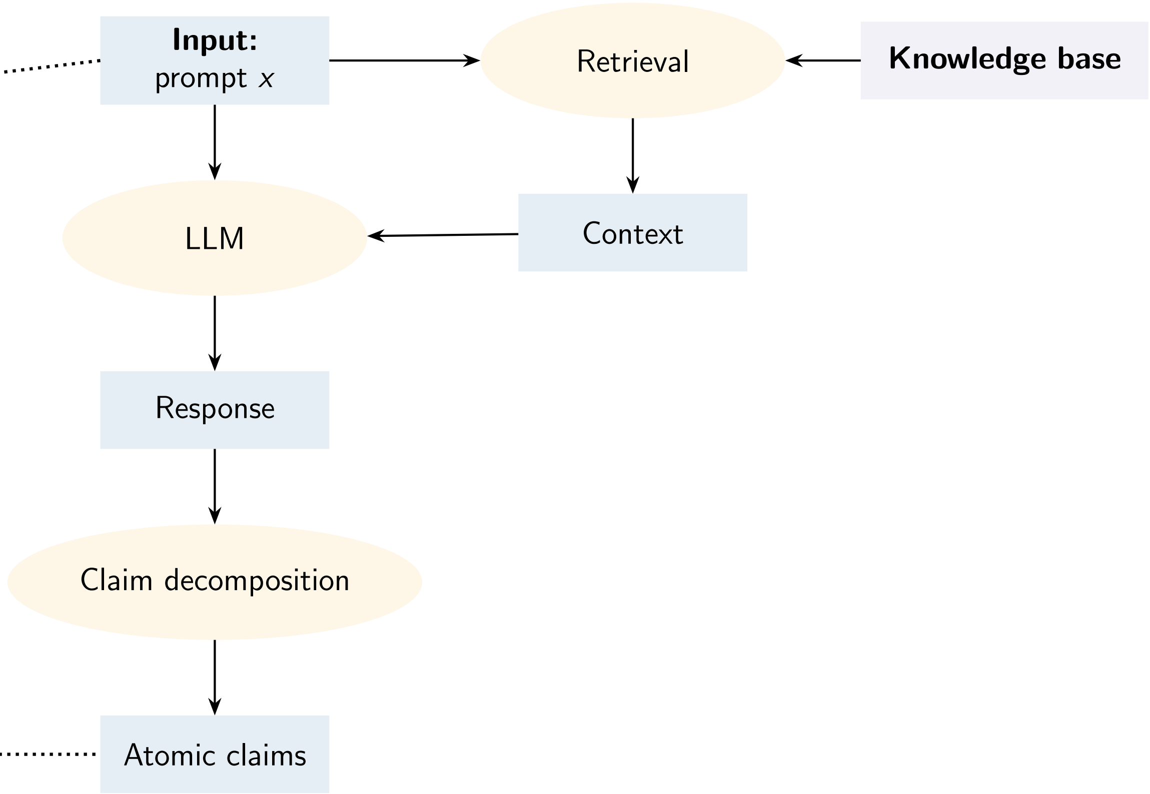
Original response

The Global Stocktake is a key component of the Paris Agreement, scheduled to begin in 2023. It involves a comprehensive assessment of global progress in addressing climate change, considering scientific evidence, economic [...]. The Stocktake aims to [...]



Assessing Faithfulness

Prompt
What is the purpose of the Global Stocktake?



Atomic claims

- The Global Stocktake is a key component of the Paris Agreement.
- The Global Stocktake is scheduled to begin in 2023.
- The Global Stocktake involves a comprehensive assessment of global progress in addressing climate change.
- The Global Stocktake considers scientific evidence, economic implications, and societal impacts.

...

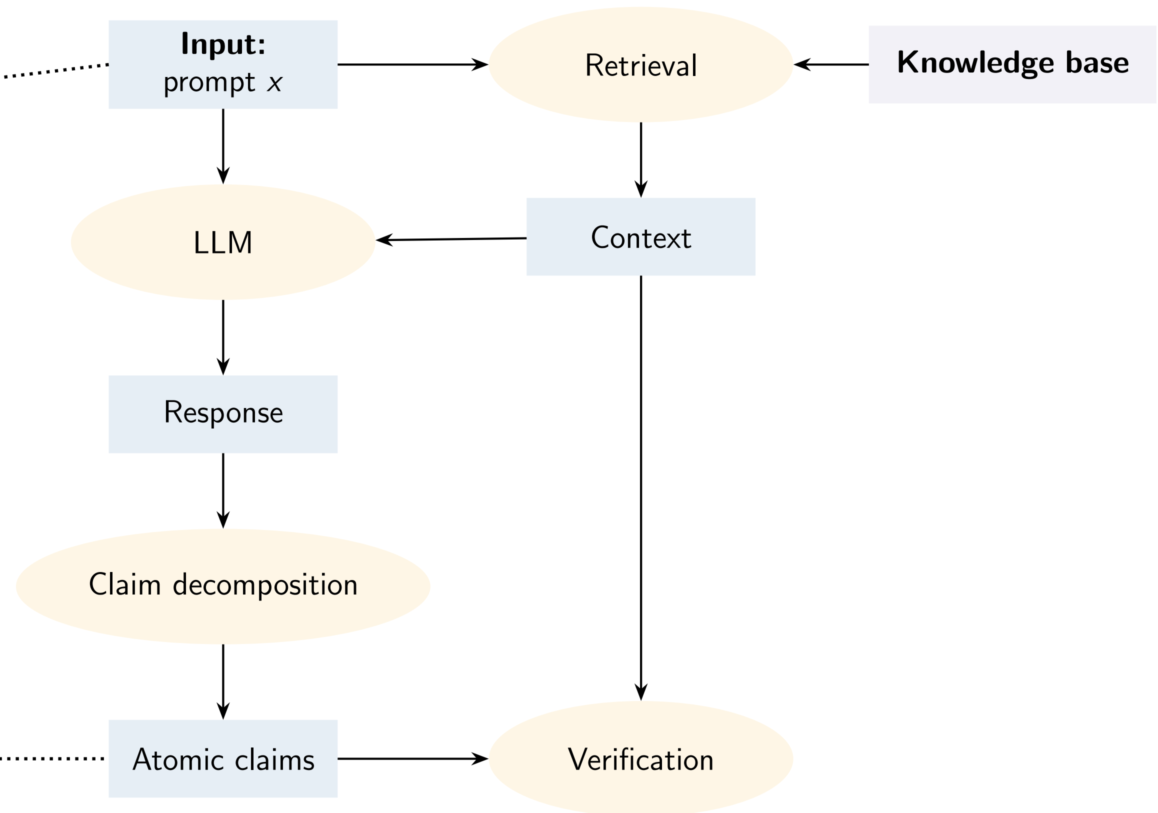
Assessing Faithfulness

Prompt
What is the purpose of the Global Stocktake?

Atomic claims (Verified)

- The Global Stocktake is a key component of the Paris Agreement.
- The Global Stocktake is scheduled to begin in 2023.
- The Global Stocktake involves a comprehensive assessment of global progress in addressing climate change.
- The Global Stocktake considers scientific evidence, economic implications, and societal impacts.

...



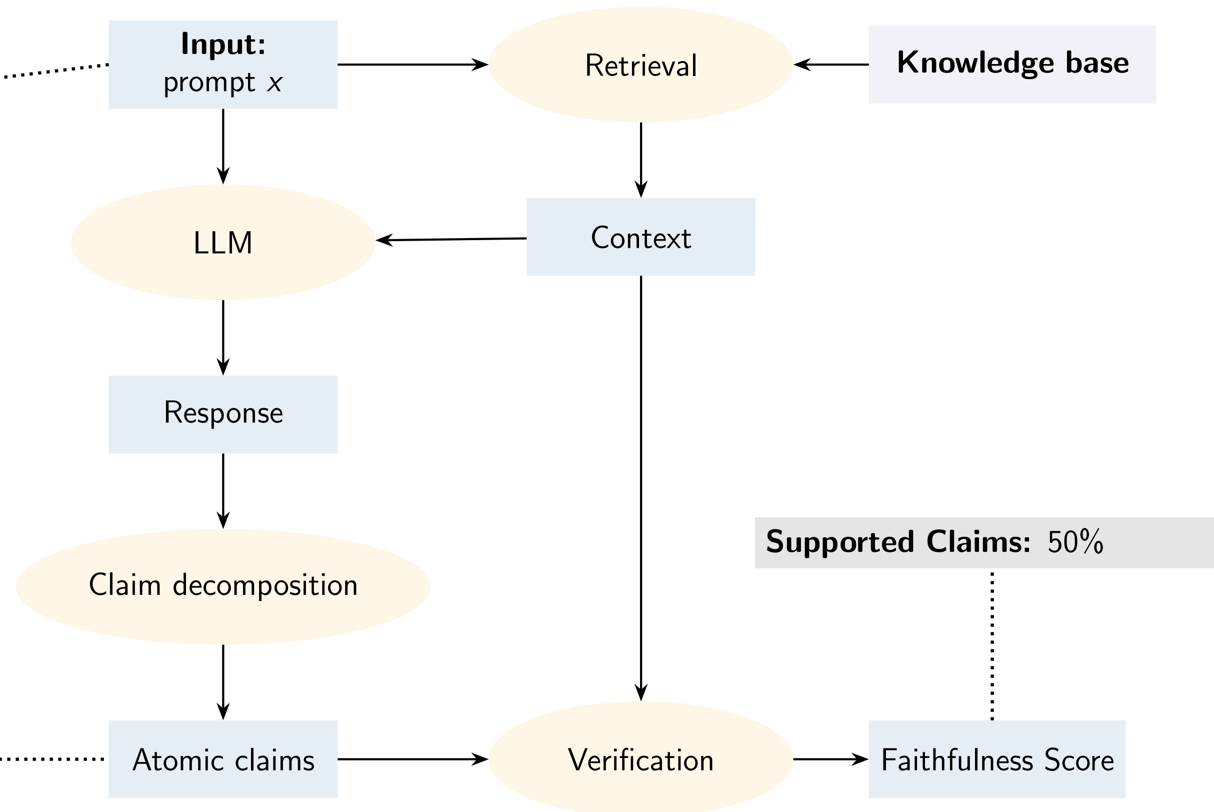
Assessing Faithfulness

Prompt
What is the purpose of the Global Stocktake?

Atomic claims (Verified)

- The Global Stocktake is a key component of the Paris Agreement.
- The Global Stocktake is scheduled to begin in 2023.
- The Global Stocktake involves a comprehensive assessment of global progress in addressing climate change.
- The Global Stocktake considers scientific evidence, economic implications, and societal impacts.

...



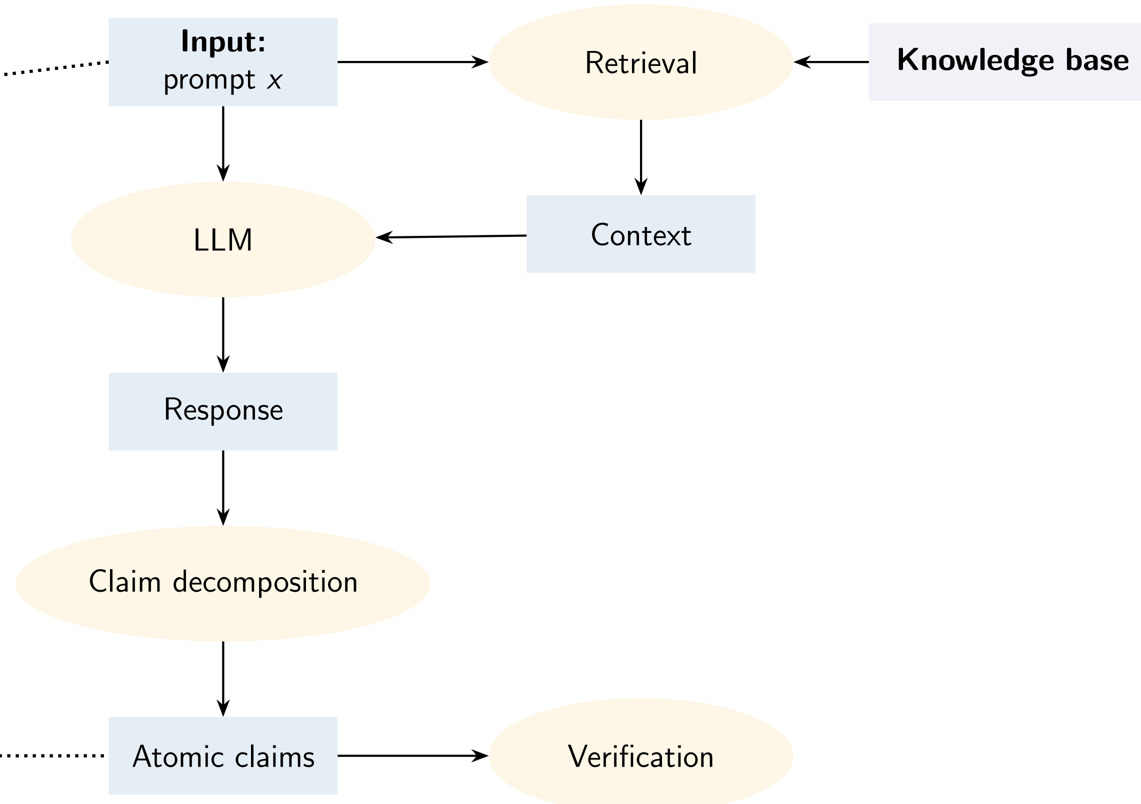
Assessing Faithfulness

Prompt
What is the purpose of the Global Stocktake?

Atomic claims (Verified)

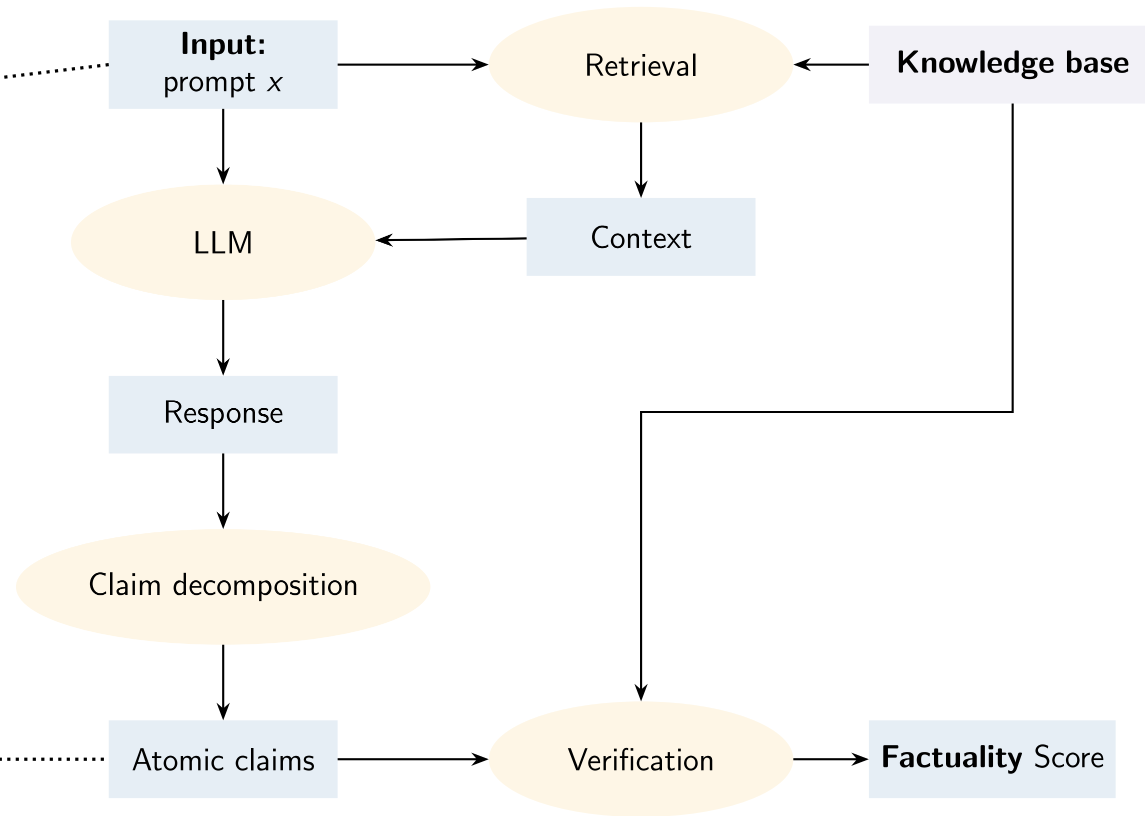
- The Global Stocktake is a key component of the Paris Agreement.
- The Global Stocktake is scheduled to begin in 2023.
- The Global Stocktake involves a comprehensive assessment of global progress in addressing climate change.
- The Global Stocktake considers scientific evidence, economic implications, and societal impacts.

...



Assessing Faithfulness

Prompt
What is the purpose of the Global Stocktake?



Atomic claims (Verified)

The Global Stocktake is a key component of the Paris Agreement.

The Global Stocktake is scheduled to begin in 2023.

The Global Stocktake involves a comprehensive assessment of global progress in addressing climate change.

The Global Stocktake considers scientific evidence, economic implications, and societal impacts.

...

Assessing and Improving Faithfulness

Assessing Faithfulness

Prompt
What is the purpose of the Global Stocktake?

Atomic claims (Verified)

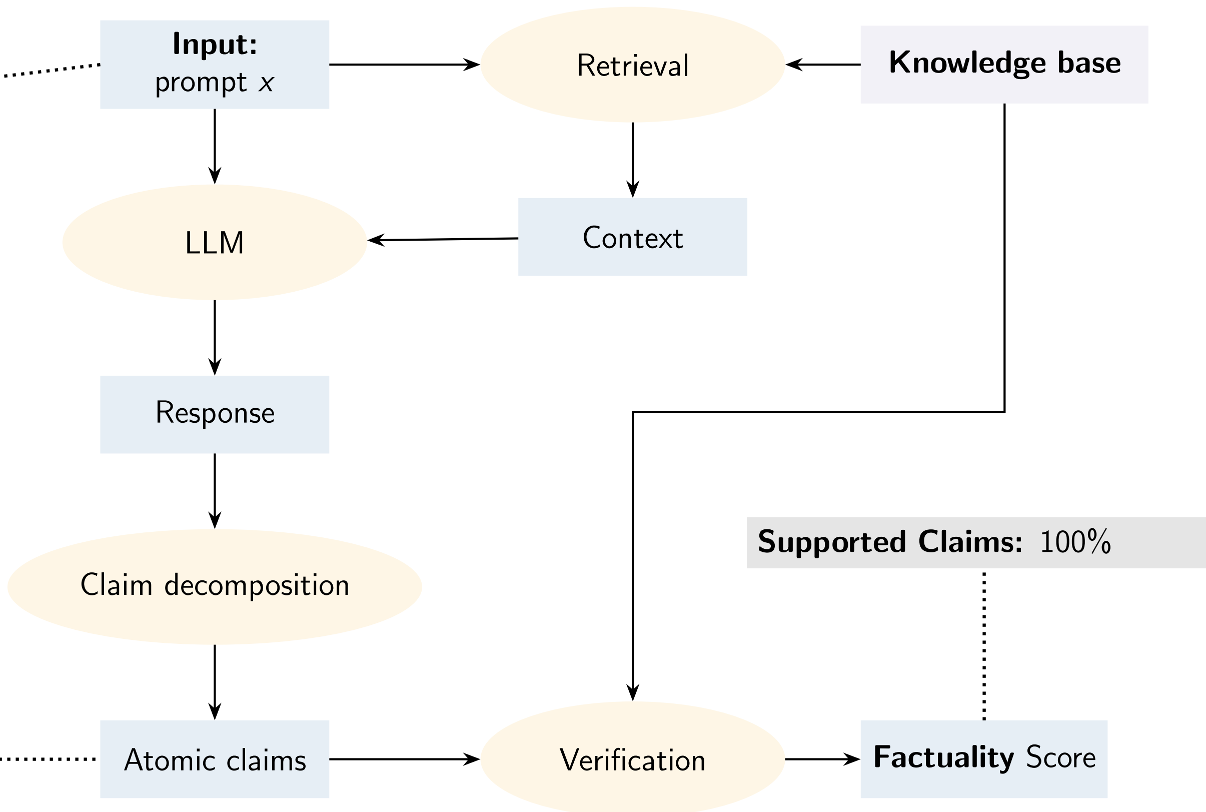
The Global Stocktake is a key component of the Paris Agreement.

The Global Stocktake is scheduled to begin in 2023.

The Global Stocktake involves a comprehensive assessment of global progress in addressing climate change.

The Global Stocktake considers scientific evidence, economic implications, and societal impacts.

...



Results

Faithfulness Evaluated on a held-out part of the Instruction Fine-Tuning data:

Model	#Tokens in Trillion	#Parameters in Billion	RAG	Avg. #Claims	Faithfulness	Factuality[4]
LLama 2 Chat	2	7	-	23.3	-	60
			✓	21.2	48	65
ClimateGPT	2	7	-	21.6	-	59
			✓	21.1	30	61

Results for faithfulness, and factuality as measured by RAGAs for different large language models with and without RAG.

[4]Might be underestimated due to small size of knowledge base.

Assessing and Improving Faithfulness

Improving Faithfulness - Analysing the IFT Data

Ablation experiment showing influence of parts of non-expert data:

	Other	**Open-End.**	**Closed-End.**	**Grounded**	**Avg.**	
Size	65,000	8,503	1,160	3,328	**#Claims**	**Faithfulness**
ClimateGPT 7B	✓	✓	✓	✓	21.1	30
	✓	✓	✓	-	19.2	57
	✓	✓	-	-	18.9	49
	✓	-	✓	-	20.1	58
	✓	-	-	-	20.4	53

Ablation study results showing test-time claim support for different training data combinations.

Overall expert data has higher claim support and thus we keep it.

Assessing and Improving Faithfulness

Improving Faithfulness - Analysing the IFT Data

Ablation experiment showing influence of parts of non-expert data:

		Other	Open-End.	Closed-End.	Grounded	Avg.	
	Size	65,000	8,503	1,160	3,328	**#Claims**	**Faithfulness**
	ClimateGPT 7B	✓	✓	✓	✓	21.1	30
⇒		✓	✓	✓	-	19.2	57
		✓	✓	-	-	18.9	49
		✓	-	✓	-	20.1	58
		✓	-	-	-	20.4	53

Ablation study results showing test-time claim support for different training data combinations.

Overall expert data has higher claim support and thus we keep it.

ClimateGPT Faithful+

Model	#Tokens in Trillion	#Parameters in Billion	RAG	Avg. #Claims	Faithfulness	Factuality[5]
LLama 2 Chat	2	7	-	23.3	-	60
			✓	21.2	48	65
ClimateGPT	2	7	-	21.6	-	59
			✓	21.1	30	61

Results for faithfulness, and factuality as measured by RAGAs for different large language models with and without RAG.

[5] Might be underestimated due to small size of knowledge base.

ClimateGPT Faithful+

Model	#Tokens in Trillion	#Parameters in Billion	RAG	Avg. #Claims	Faithfulness	Factuality[5]
LLama 2 Chat	2	7	-	23.3	-	60
			✓	21.2	48	65
ClimateGPT	2	7	-	21.6	-	59
			✓	21.1	30	61
ClimateGPT Faithful+	2	7	-	20.2	-	57
			✓	19.2	<u>57</u>	<u>69</u>

Results for faithfulness, and factuality as measured by RAGAs for different large language models with and without RAG.

[5] Might be underestimated due to small size of knowledge base.

ClimateGPT Faithful+

Model	#Tokens in Trillion	#Parameters in Billion	RAG	Avg. #Claims	Faithfulness	Factuality[5]
LLama 2 Chat	2	7	-	23.3	-	60
			✓	21.2	48	65
ClimateGPT	2	7	-	21.6	-	59
			✓	21.1	30	61
ClimateGPT Faithful+	2	7	-	20.2	-	57
			✓	19.2	57	69
LLama 3.1 Instruct	15	8	-	22.7	-	59
			✓	17.3	67	72

Results for faithfulness, and factuality as measured by RAGAs for different large language models with and without RAG.

[5] Might be underestimated due to small size of knowledge base.

Example

Prompt and Context from Climate Policy Radar (Juhasz et al., 2024)
Parts in red are not supported by the context.

Prompt: What is the purpose of the Global Stocktake?

Example

Prompt and Context from Climate Policy Radar (Juhasz et al., 2024)
Parts in red are not supported by the context.

Prompt: What is the purpose of the Global Stocktake?

Context: Recalling Article 14 of the Paris Agreement, the objective of the Global Stocktake is to assess the collective progress towards achieving the purpose of the Agreement [...]
Paragraph 34 of 19/CMA.1 decides that the outputs of this component of the global stocktake should: [...]
The SBSTA and the SBI invited submissions on views on the approach to the consideration of outputs component of the first [...]

Example

Prompt and Context from Climate Policy Radar (Juhasz et al., 2024)
Parts in red are not supported by the context.

Prompt: What is the purpose of the Global Stocktake?

Context: Recalling Article 14 of the Paris Agreement, the objective of the Global Stocktake is to assess the collective progress towards achieving the purpose of the Agreement [...]
Paragraph 34 of 19/CMA.1 decides that the outputs of this component of the global stocktake should: [...]
The SBSTA and the SBI invited submissions on views on the approach to the consideration of outputs component of the first [...]

ClimateGPT 7B
The Global Stocktake is a key component of the Paris Agreement, scheduled to begin in 2023. It involves a comprehensive assessment of global progress in addressing climate change, considering scientific evidence, economic implications, and societal impacts. The Stocktake aims to inform the updates to countries' climate commitments, known as "nationally determined contributions" (NDCs), to ensure they are in line with the agreement's objectives and targets. This process seeks to enhance international cooperation and commitment to mitigating climate change on a global scale.

Assessing and Improving Faithfulness

Example

Prompt and Context from Climate Policy Radar (Juhasz et al., 2024)
Parts in red are not supported by the context.

Prompt: What is the purpose of the Global Stocktake?

Context: Recalling Article 14 of the Paris Agreement, the objective of the Global Stocktake is to assess the collective progress towards achieving the purpose of the Agreement [...]
Paragraph 34 of 19/CMA.1 decides that the outputs of this component of the global stocktake should: [...]
The SBSTA and the SBI invited submissions on views on the approach to the consideration of outputs component of the first [...]

ClimateGPT 7B
The Global Stocktake is a key component of the Paris Agreement, scheduled to begin in 2023. It involves a comprehensive assessment of global progress in addressing climate change, considering scientific evidence, economic implications, and societal impacts. The Stocktake aims to inform the updates to countries' climate commitments, known as "nationally determined contributions" (NDCs), to ensure they are in line with the agreement's objectives and targets. This process seeks to enhance international cooperation and commitment to mitigating climate change on a global scale.

ClimateGPT 7B Faithful+
The Global Stocktake's purpose, as outlined in Article 14 of the Paris Agreement, is to assess the collective progress made towards achieving the Agreement's purpose and long-term goals, which encompass mitigation, adaptation, and means of implementation. This assessment is conducted considering equity and the best available science. The Global Stocktake process spans two years, beginning with information collection and preparation and culminating in a consideration of outputs. The outputs of the latter component are decided upon in paragraph 34 of 19/CMA.1, which states that they should identify opportunities for enhancing action and support for collective progress in relation to thematic areas of the global stocktake and possible measures and good practices.

Assessing and Improving Faithfulness

Conclusion

- Faithfulness is critical for trustworthy RAG systems especially in the climate-domain
- Automated evaluation can quantify how well models adhere to retrieved context.
- ClimateGPT's original instruction data contained unfaithful RAG examples, limiting faithfulness
- Excluding unfaithful non-expert grounded data improved faithfulness from 30% to 57%
- **Key Insight:** High-quality, faithful training examples drive better faithfulness in RAG settings

- Future Work and Open Challenges:
  - Separation of factual information and opinions
  - Improving automatic verification of complex climate-related claims
  - Further improve instruction fine-tuning data for faithfulness

Questions?

dthulke@apptek.com

Literature

Ainslie, J., Lee-Thorp, J., de Jong, M., Zemlyanskiy, Y., Lebrón, F., and Sanghai, S. (2023). Gqa: Training generalized multi-query transformer models from multi-head checkpoints. *ArXiv preprint*, abs/2305.13245.

Bulian, J., Schäfer, M. S., Amini, A., Lam, H., Ciaramita, M., Gaiarin, B., Huebscher, M. C., Buck, C., Mede, N., Leippold, M., and Strauss, N. (2023). Assessing large language models on climate information.

Chen, Z., Cano, A. H., Romanou, A., Bonnet, A., Matoba, K., Salvi, F., Pagliardini, M., Fan, S., Köpf, A., Mohtashami, A., Sallinen, A., Sakhaeirad, A., Swamy, V., Krawczuk, I., Bayazit, D., Marmet, A., Montariol, S., Hartley, M.-A., Jaggi, M., and Bosselut, A. (2023). Meditron-70b: Scaling medical pretraining for large language models.

Es, S., James, J., Espinosa Anke, L., and Schockaert, S. (2024). RAGAs: Automated evaluation of retrieval augmented generation. In Aletras, N. and De Clercq, O., editors, *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 150–158, St. Julians, Malta. Association for Computational Linguistics.

Juhasz, M., Dutia, K., Franks, H., Delahunty, C., Mills, P. F., and Pim, H. (2024). Responsible retrieval augmented generation for climate decision making from documents.

Leippold, M., Vaghefi, S. A., Stammbach, D., Muccione, V., Bingler, J., Ni, J., Colesanti-Senni, C., Wekhof, T., Schimanski, T., Gostlow, G., Yu, T., Luterbacher, J., and Huggel, C. (2024). Automated fact-checking of climate change claims with large language models.

Li, N., Zahra, S., de Brito, M. M., Flynn, C. M., Görnerup, O., Koffi, W., Kurfali, M., Meng, C., Thiery, W., Zscheischler, J., Messori, G., and Nivre, J. (2024). Using LLMs to build a database of climate extreme impacts. In *Natural Language Processing meets Climate Change @ ACL 2024*.

Shazeer, N. (2020). Glu variants improve transformer. *ArXiv preprint*, abs/2002.05202.

Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., and Liu, Y. (2023). Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, page 127063.

Taylor, R., Kardas, M., Cucurull, G., Scialom, T., Hartshorn, A., Saravia, E., Poulton, A., Kerkez, V., and Stojnic, R. (2022). Galactica: A large language model for science.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. (2023a). Llama: Open and efficient foundation language models. *ArXiv preprint*, abs/2302.13971.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. (2023b). Llama 2: Open foundation and fine-tuned chat models. *ArXiv preprint*, abs/2307.09288.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Literature

Wu, S., Irsoy, O., Lu, S., Dabravolski, V., Dredze, M., Gehrmann, S., Kambadur, P., Rosenberg, D., and Mann, G. (2023). Bloomberggpt: A large language model for finance. *ArXiv preprint: https://arxiv. org/pdf/2303.17564.*

Xiong, R., Yang, Y., He, D., Zheng, K., Zheng, S., Xing, C., Zhang, H., Lan, Y., Wang, L., and Liu, T. (2020). On layer normalization in the transformer architecture. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 10524–10533. PMLR.