

Robust Temporal QA

Dan Schumacher Gagana Uday Kumar Tara Grey Niharika Bandlamudi Nupoor Karnik

Abstract

This study investigates the efficacy of leveraging irrelevant contexts to mitigate temporal blinds. Using data sourced from the GPT-3.5 turbo baseline model and Gemma 2b instruction tuned and non-instruction tuned baseline models, we examine four types of contexts: relevant, random(irrelevant), wrong data, and no contexts. We compare the model's performance against the instruction tuned Gemma model after training on these diverse contexts, employing metrics such as exact match, contains accuracy, and BEM scores for evaluation. Our analysis reveals that the baseline CHAT GPT 3.5 turbo model outperforms others, particularly excelling in handling relevant contexts. This underscores the robustness of the model against irrelevant contexts.

Introduction

In recent years, large language models (LLMs) have revolutionized natural language processing (NLP) with their remarkable ability to understand and generate human-like text. These models, such as GPT-3.5 turbo and GEMMA, are trained on vast amounts of text data, allowing them to learn complex patterns and relationships within language. However, while LLMs excel in many NLP tasks, they can struggle when faced with temporal aspects of text, such as understanding historical events or reasoning about the order of events over time. This limitation raises questions about how LLMs prioritize and utilize temporal information, particularly in the presence of irrelevant or conflicting context.

This study focuses on investigating the temporal understanding abilities of these LLMs, specifically exploring how these models handle temporal information. Understanding how LLMs process temporal aspects of text is crucial for improving their performance in tasks that require temporal reasoning, such as question-answering and summarization.

The aim of this study is to develop a model that is robust and can temporally reason about returned contexts to understand whether it should be used or not.

Methodology/Approach

The research employed a dataset comprising 500 observations, wherein contexts were derived from CHAT GPT 3.5 turbo and Gemma 2b instruction tuned (IT) and non-instruction tuned (NIT). Each observation featured a temporal question alongside its corresponding answer. These elements served as inputs for both models to generate contexts, which were subsequently categorized into four distinct types:

- No_context: (zero shot) Models generated responses without any provided context.
- Relevant_context: Models furnished context directly pertinent to the question at hand.
- Random_context: Models produced context that was unrelated to the question or answer.
- Wrong_date_context: Models provided context identical to relevant_context, except for a deliberately incorrect date.

Example of CHAT GPT 3.5 turbo and Gemma 2B baselines:

CHAT GPT 3.5 turbo			
Question- Who was the winner of American Idol in 2006?		Answer- Taylor Hicks	
Relevant_Context	Relevant_context output	Wrong_date_context	Wrong_date context_output

The winner of American Idol in 2006, which was the show's fifth season, was Taylor Hicks. His soulful performances and unique style, often described as "Modern Whomp," captured the hearts of viewers and earned him a dedicated fan base known as the "Soul Patrol." Hicks's victory on American Idol helped launch his music career, allowing him to release studio albums and singles, as well as to perform in live shows across the United States.	Yes, Taylor Hicks was indeed the winner of American Idol in 2006. His win marked a significant moment in the show's history and helped propel his music career forward.	The winner of American Idol in 1896, which was the show's fifth season, was Taylor Hicks. His soulful performances and unique style, often described as "Modern Whomp," captured the hearts of viewers and earned him a dedicated fan base known as the "Soul Patrol." Hicks's victory on American Idol helped launch his music career, allowing him to release studio albums and singles, as well as to perform in live shows across the United States.	I'm sorry, but there seems to be some confusion in your question. The winner of American Idol in 2006 was Taylor Hicks. He won the show's fifth season, not in 1974.
--	---	--	--

Table1. Chat GPT Context Example

Gemma 2B IT			
Question- Which country ratified the Berne convention on copyright in March 1989?		Answer- The United States of America	
Relevant_Context_provided	Relevant_context_output	Wrong_date_context_provided	Wrong_date_context_output
The United States of America ratified the Berne Convention on copyright in March 1989.	The passage does not specify which country ratified the Berne Convention on copyright in March 1989, so I cannot answer this question from the provided context.	The United States of America ratified the Berne Convention on copyright in March 1969.	The country that ratified the Berne Convention on copyright in March 1989 was China.

Table 2. Gemma IT Context Example

From the above examples it can be said that CHAT GPT 3.5 turbo was able to ignore the wrong context provided and focused on the relevant context and gave the output accordingly. However, in Gemma 2b IT gave an output with the wrong answer when wrong context was provided. It was not able to provide correct information.

Example of the Trained Gemma 2b IT outputs:

	Question	answer	prededctions			
			no_context	rel_context	rand_context	wd_context
no_context	January 1st, 2002, saw which currency being used in several European countries including France, Germany and Ireland?	The Euro	The Eurocurrency was being used in several European countries including France, Germany and Ireland on January 1st, 2002.	The euro (€) became the official currency for 12 member states of the European Union, including France, Germany, and Ireland.	The Eurocurrency was being used in several European countries including France, Germany and Ireland.	The euro (€) became the official currency for 12 member states of the European Union, including France, Germany, and Ireland on October 1st.
rel_context	In 1998, Exxon and Mobil merged and form what new company?	Exxon mobil	Sure, in 1998, Exxon and Mobil merged to form the new company, ExxonMobil.	Sure, In 1998, Exxon and Mobil merged to form the new company known as ExxonMobil.	The passage does not mention Exxon or Mobil, so I cannot answer this question from the provided context.	Sure, in 1998, Exxon and Mobil merged and formed the new company known as ExxonMobil.
rand_context	Chinese troops kill hundreds of pro-democracy students in which city in June 1989?	Beijing	The premise of the question is incorrect. There is no evidence to support the claim that Chinese troops killed hundreds of pro-democracy students in June 1989.	Sure, in Beijing, China. The Tiananmen Square Massacre was a major event in June 1989 in which hundreds of pro-democracy students were killed.	The context does not provide any information about the city in which Chinese troops killed hundreds of pro-democracy students in June 1989, so I cannot answer.	The context does not provide any information about the city where Chinese troops killed hundreds of pro-democracy students in June 1989, so I cannot answer.
wd_context	In 1998 120 countries agreed to set up what?	The International Criminal Court	Sure, in 1998 120 countries agreed to set up the "G8", a group of developed countries that meet regularly to discuss	Sure, in 1998 120 countries agreed to set up the International Criminal Court (ICC). The ICC was created as a permanent tribunal	The passage does not specify what 120 countries agreed to set up, so I cannot answer this question from the provided context.	Sure, In 1893, 120 countries agreed to establish the International Criminal Court (ICC). The ICC was created as a permanent tribunal

Table 3. Gemma NIT Context Outputs

Based on the data presented in the table, it's evident that the trained Gemma 2b model performs better than the non-trained model. When presented with out-of-place context, the trained model appropriately identifies the context as incorrect and refrains from providing an erroneous answer. This improvement indicates the effectiveness of training in enhancing the model's performance. However, despite this improvement, the trained Gemma 2b model still lags behind CHAT GPT in certain aspects. Specifically, it struggles to confidently disregard wrong date contexts and deliver accurate responses. This suggests that the Gemma 2b model encounters challenges, particularly in handling temporal information, such as dates, where it exhibits some degree of confusion.

After categorizing the contexts into four distinct types, baseline metrics such as Exact Match F1, Contains Accuracy, and Contains BEM score were computed for each context for the GPT 3.5 Turbo model, and the IT, and NIT Gemma 2b models. These baseline metrics served as a benchmark for evaluating the performance of the fine-tuned Gemma Models.

Subsequently, Gemma 2b underwent fine-tuning using all four contextual types. Fine-tuning involved adjusting model parameters to better align with the different contextual scenarios, aiming to improve performance on temporal questions answering tasks. The fine-tuning process included training over multiple epochs, adjusting learning rates, and potentially employing specialized techniques to adapt the models to the diverse contextual types.

Throughout the fine-tuning process, key evaluation metrics, including 'exact match', 'contains accuracy', and 'BEM' scores, were used to assess both baseline and fine-tuned model performance. Higher scores across these metrics denoted a greater alignment between the generated answer and its corresponding context.

This methodological approach was devised to offer a comprehensive assessment of model efficacy across varied contextual scenarios, thus contributing to a nuanced understanding of model performance in temporal question answering tasks.

Findings/Evaluation

After finding the metrics, it was seen that CHAT GPT 3.5 turbo model generated the best outputs with 'contains accuracy' of 89% followed by Gemma 2b NIT with 79% and lastly by Gemma 2b IT model by 69%. The BEM scores were highest for CHAT GPT viz; 95.9%

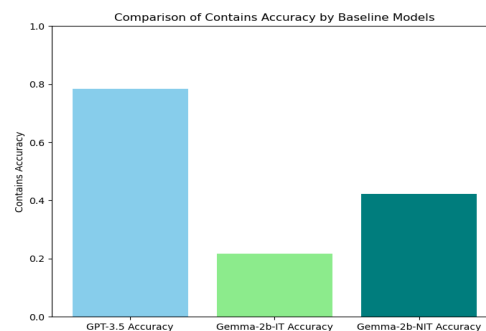


Chart 1. Baseline Model's Average Accuracy

After fine tuning the Gemma 2b IT it was seen that it performed very well in generating 'relevant_context' compared to other contexts with an approximate 'contains accuracy' of 84% . The BEM score was high as well with an average of 87.3% for relevant context.

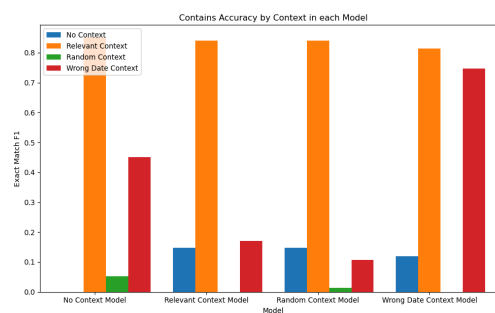


Chart 2. Trained Model's Contains Accuracy for Each Context

The Gemma model trained on wrong date context achieved the highest average accuracy score among the fine-tuned models as shown below.

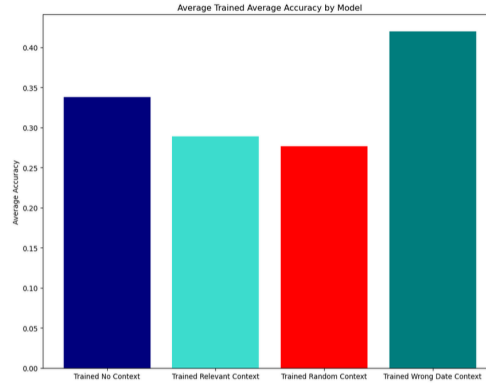


Chart 3. Average Trained Models Accuracy Score

This indicates that exposure to incorrect temporal information during training has rendered the model robust to erroneous data, thereby contributing to its superior performance.

Artifact Discussion

In this section we delve deeper into the implication of our findings and discuss their significance in the context of the current research of training a model to be robust against irrelevant context to help the model use information already in its memory. Furthermore, there were performance discrepancies between the GPT 3.5 model and the Gemma 2b models as mentioned earlier. These discrepancies help underscore the importance of model selection and architecture in temporal understanding tasks. While the CHAT GPT 3.5 turbo model demonstrated superior performance, particularly in handling relevant contexts, the Gemma 2b models, despite fine-tuning, lagged behind. This raises questions about the underlying mechanisms governing temporal reasoning in these models and warrants further investigation into model design and training methodologies.

Our study highlights the challenges encountered in fine-tuning the Gemma 2b models to improve their temporal understanding abilities. Despite efforts to adapt the models to diverse contextual scenarios, including irrelevant and wrong data contexts, the performance gains were limited. This suggests potential limitations in the fine-tuning process or the inherent architecture of the Gemma models, prompting a deeper exploration into the factors influencing model adaptability and generalization.

Moreover, it's important to emphasize the significance of training models to be robust against irrelevant context. This will enhance the models' ability to filter out noise and focus on relevant information, improving the accuracy, efficiency, and reliability of natural language processing systems. Models that have the ability to discern between relevant and irrelevant information are better equipped to provide accurate responses and make informed decisions. Beyond the specific models and contexts investigated in this study, the insights gained have a broader applicability. The models can inform the development and refinement of other large language models and temporal reasoning tasks across various domains and applications. Understanding how models prioritize and utilize temporal information is crucial for advancing the field of natural language processing and improving the performance of AI systems in real-world scenarios.

Our findings contribute to the ongoing discourse on model interpretability and explainability in AI systems. By understanding how models process temporal information, researchers can develop methods to make their decision-making processes more transparent and interpretable. This is particularly important in the applications where reasoning behind model predictions needs to be understood and justified, such as in healthcare or legal domains. There are also ethical considerations and implications to consider when leveraging irrelevant contexts in machine learning models. Propagating misinformation or biased decision-making can have harmful consequences if models are not adequately trained to discern between relevant and irrelevant information. Ethical considerations should be integrated into the design, development and deployment of AI systems to mitigate potential risks and ensure responsible AI practices.

Lastly, a couple limitations were faced when designing and fine-tuning our model, one limitation was the dataset size and diversity. The size and diversity of the dataset was a limitation

because even though we tried to collect a comprehensive dataset consisting of 500 observations, the dataset may not fully represent the complexity and variability of real-world temporal question answering tasks. Additionally, the dataset is limited in its coverage of diverse context and linguistic nuances, which could impact the generalizability of the findings. We would expect the model trained on relevant context to perform better if more question and answer data as it would allow the model to learn from even more information. Another limitation was the Evaluation Metrics, the metrics used in this study may present limitations down the road as it does not capture the full spectrum of the models performance. While the metrics, such as exact match, contain accuracy and BEM scores provide valuable insights into the models accuracy and alignment with contextual information, they may not fully capture the nuances in semantic understanding or contextual relevance. For example the BEM score is a BERT-based evaluation metric and the BERT model only processes input sequences of a fixed length, usually up to only 512 tokens. Due to the fixed length limitation, if the model outputs a answer to the question that is longer then 512 tokens and the answer to the question is at the end, the BEM score will be very low or a score of zero. Therefore, as we increase the dataset it is good to be aware of the limitation of BEM scores and adjust accordingly or even consider an alternative evaluation metric.

Conclusion

The study compared the performance of large language models (LLMs) like GPT-3.5 turbo and Gemma 2b when trained on different types of contexts, including irrelevant ones, to improve their temporal understanding abilities. Despite training on diverse contexts, the baseline GPT-3.5 turbo model performed better than Gemma 2b, especially in handling relevant contexts. This indicates that while LLMs can ignore irrelevant contexts, they do not necessarily benefit from them in temporal tasks. Future research could explore other methods to enhance LLMs' temporal reasoning, such as using larger models or refining training data.

Future Work

It's intriguing that despite our efforts to refine the model with pertinent context, the enhancement in relevant context predictions remains elusive. Perhaps experimenting with Gemma-7b or LLama-3 could yield better results; after all, bigger might be better in this case! Additionally, augmenting the training dataset with more diverse and extensive data could provide the model with richer context, potentially enhancing its ability to make accurate predictions on relevant contexts.. Instead of focusing solely on instruction tuning (IT), which is tailored for generalized tasks, we could explore the benefits of incorporating non-instruction tuning (NIT) as well. This approach might offer a more comprehensive understanding of context and lead to more accurate predictions. Expanding the training dataset, both concatenated and independent, could also be advantageous. More data might offer a richer understanding of context, leading to improved predictions.

Regarding context manipulation, employing techniques like shifted context could be beneficial. For instance, consider the following:

- 'Question': "The first African American President to hold the office wrote a memoir called Dreams of My Father. Who was the president of the United States in 2001?"
- 'Answer': George Bush

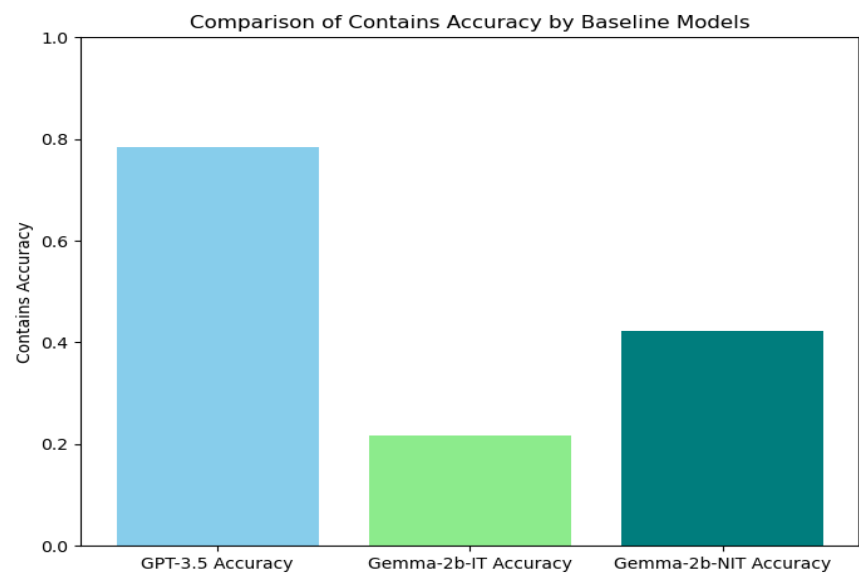
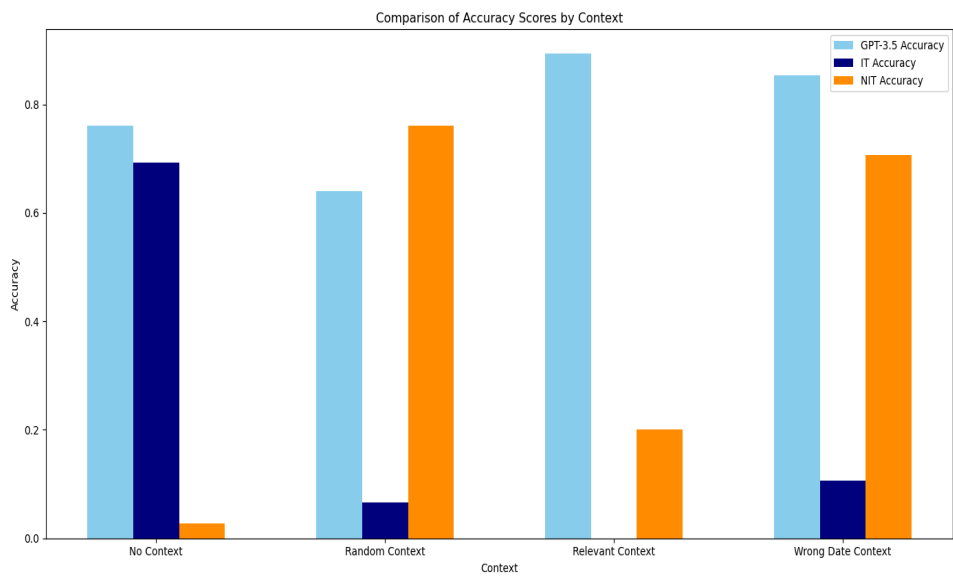
Crafting responses in sentence format might facilitate clearer communication and comprehension. Exploring such techniques could provide valuable insights into how the models handle shifted or manipulated contexts, which will hopefully contribute to a deeper understanding of their capabilities and limitations.

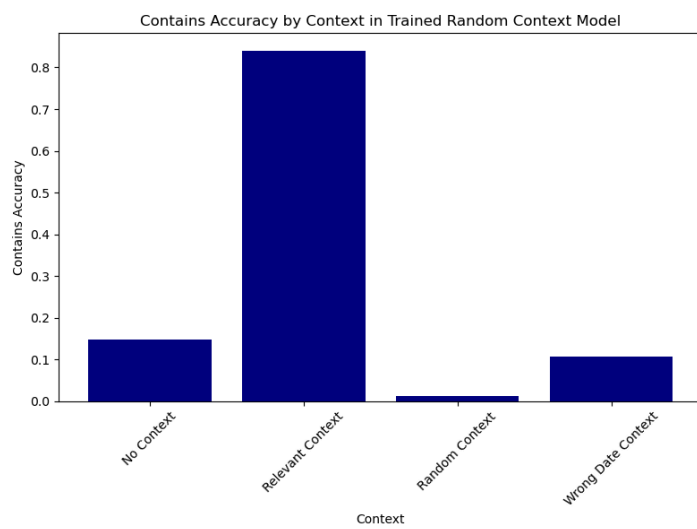
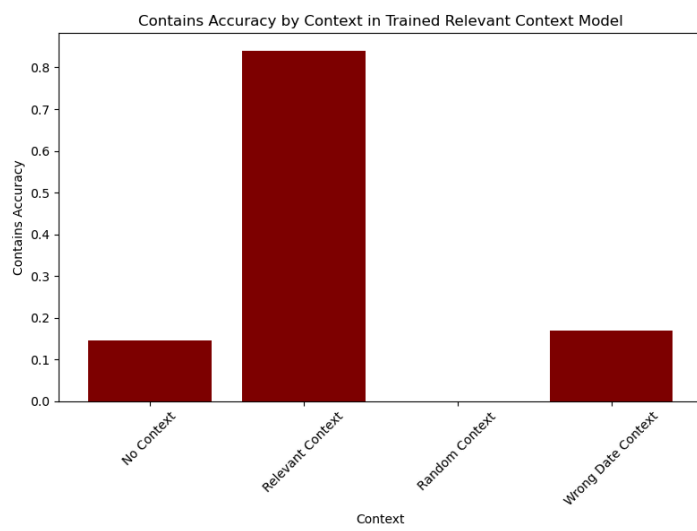
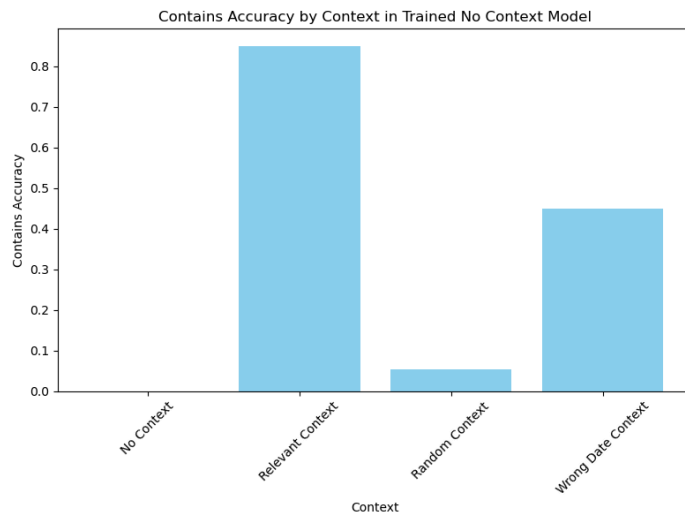
References

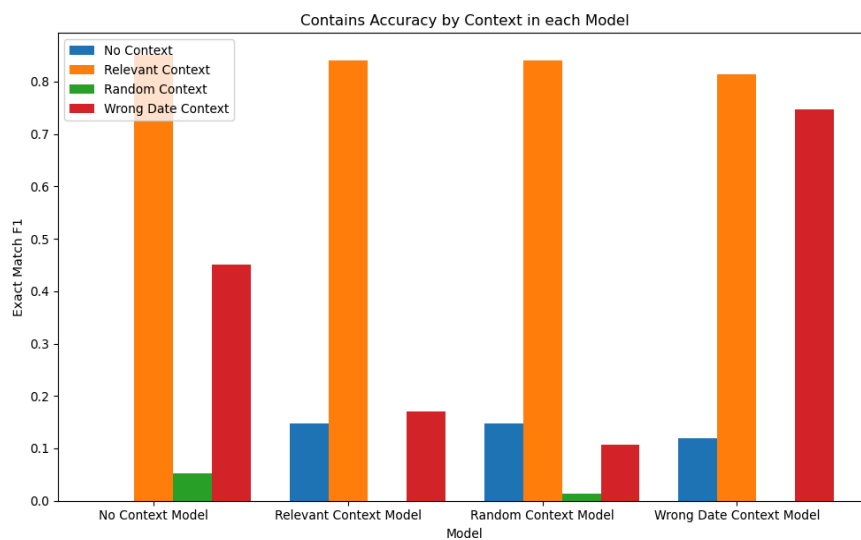
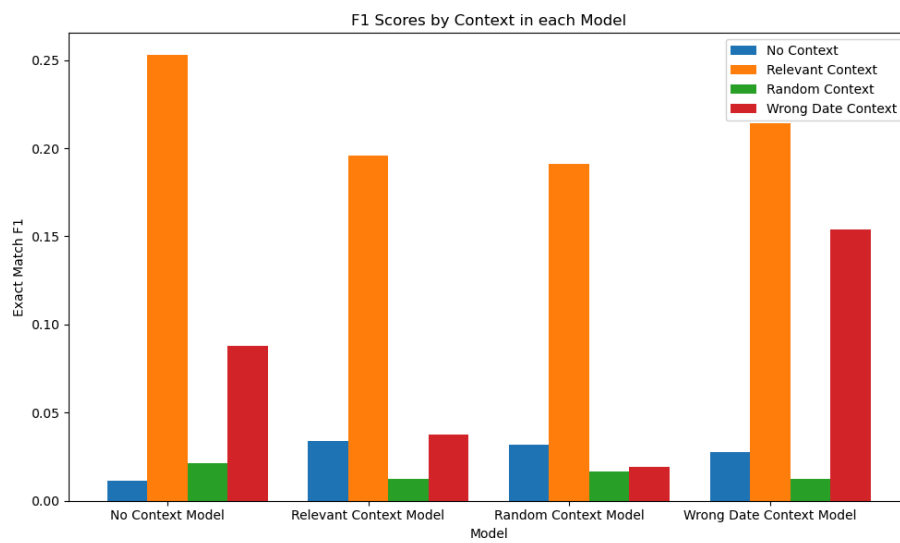
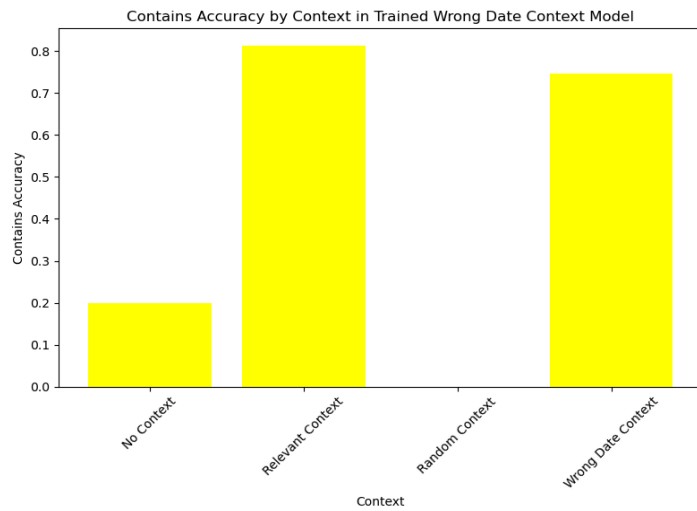
- [1] Wallat, J., Jatowt, A., & Anand, A. (2024, March). Temporal Blind Spots in Large Language Models. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining* (pp. 683-692).
- [2] S. Althammer, M. Buckley, S. Hofstätter, and A. Hanbury. Linguistically informed masking for representation learning in the patent domain. CoRR, abs/2106.05768, 2021. [3] A. Anand, S. J. Bedathur, K. Berberich.

APPENDIX A

Baseline Model Comparisons:

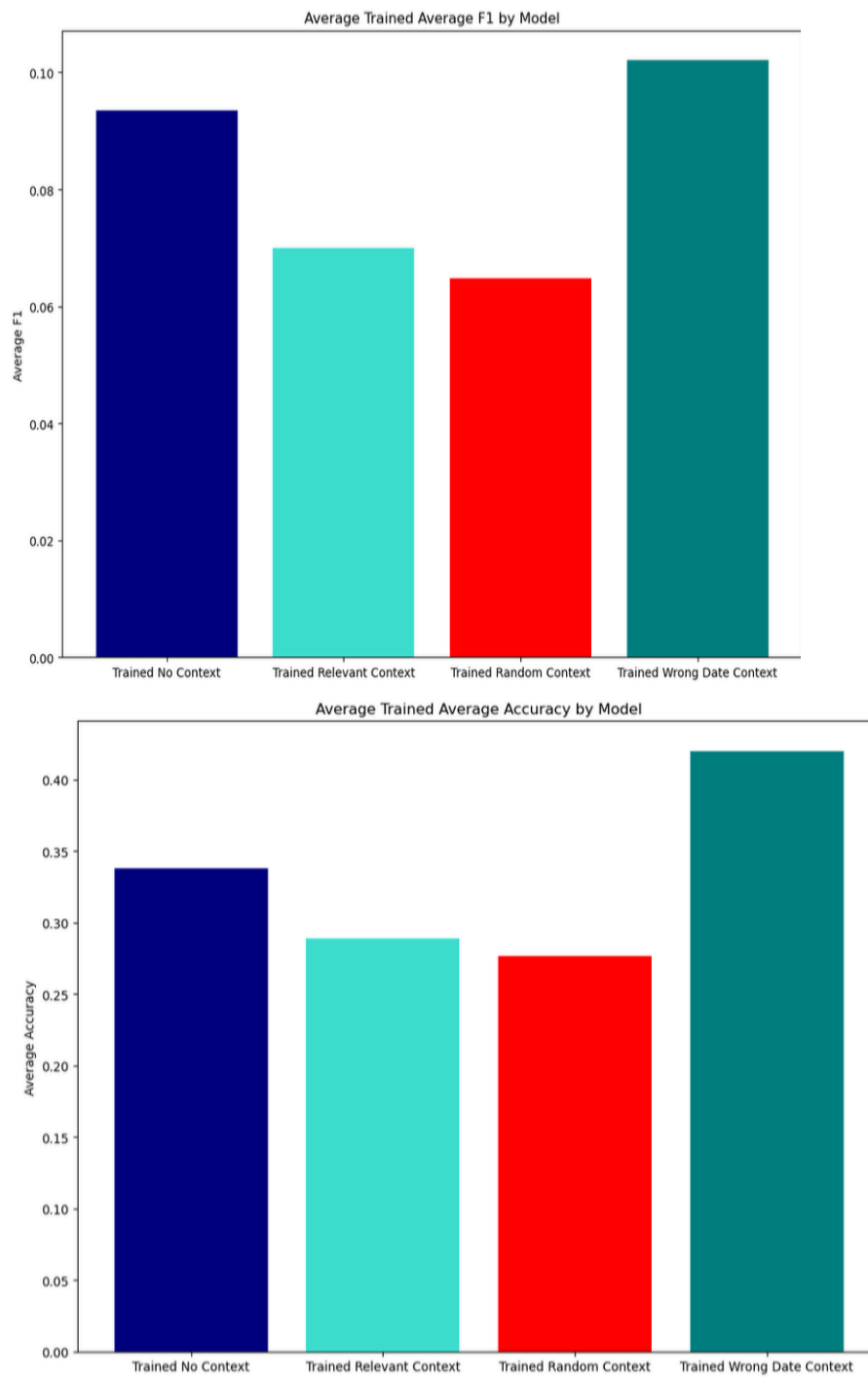






APPENDIX B

Trained Models Average Comparison:



APPENDIX C

Baseline Results:

GPT-3.5	No Context	Relevant Context	Random Context	Wrong Date Context	AVG
Contains Accuracy	0.76	0.89	0.64	0.85	0.785
F1	0.1528	0.1465	0.097	0.1191	0.12885
BEM	0.869	0.959	0.739	0.923	0.8725

IT	No Context	Relevant Context	Random Context	Wrong Date Context	AVG
Contains Accuracy	0.067	0.69	0	0.107	0.216
F1	0.0175	0.1488	0.0137	0.0191	0.049775
BEM	0.318	0.806	0.339	0	0.36575

NIT	No Context	Relevant Context	Random Context	Wrong Date Context	AVG
Contains Accuracy	0.027	0.76	0.2	0.707	0.4235
F1	0.0179	0.1374	0.0404	0.1559	0.0879
BEM	0.167	0.818	0.341	0.735	0.51525

Trained Models Results :

No Context	No Context	Relevant Context	Random Context	Wrong Date Context
Contains Accuracy	0	0.85	0.053	0.45
Exact Match F1	0.01155	0.2527	0.0215	0.0881
BEM	0.046	0.92	0.469	0.691
Relevant Context	No Context	Relevant Context	Random Context	Wrong Date Context
Contains Accuracy	0.1467	0.84	0	0.17
Exact Match F1	0.0338	0.1958	0.0124	0.03758
BEM	0.349	0.883	0.336	0.444
Random Context	No Context	Relevant Context	Random Context	Wrong Date Context
Contains Accuracy	0.147	0.84	0.0133	0.107
Exact Match F1	0.0319	0.191	0.01686	0.01911

BEM	0.385	0.835	0.339	0.43
Wrong Date Context	No Context	Relevant Context	Random Context	Wrong Date Context
Contains Accuracy	0.12	0.813	0	0.747
Exact Match F1	0.0274	0.2144	0.0126	0.1538
BEM	0.335	0.857	0.341	0.713

Examples:

Outputs of baseline models of CHAT GPT 3.5 turbo and Gemma 2B IT

CHAT GPT 3.5 turbo			
Question- Who was the winner of American Idol in 2006?		Answer- Taylor Hicks	
Relevant_Context	Relevant_context_output	Wrong_date_context	Wrong_date_context_output
The winner of American Idol in 2006, which was the show's fifth season, was Taylor Hicks. His soulful performances and unique style, often described as "Modern Whomp," captured the hearts of viewers and earned him a dedicated fan base known as the "Soul Patrol." Hicks's victory on American Idol helped launch his music career, allowing him to release studio albums and singles, as well as to perform in live shows across the United States.	Yes, Taylor Hicks was indeed the winner of American Idol in 2006. His win marked a significant moment in the show's history and helped propel his music career forward.	The winner of American Idol in 1896, which was the show's fifth season, was Taylor Hicks. His soulful performances and unique style, often described as "Modern Whomp," captured the hearts of viewers and earned him a dedicated fan base known as the "Soul Patrol." Hicks's victory on American Idol helped launch his music career, allowing him to release studio albums and singles, as well as to perform in live shows across the United States.	I'm sorry, but there seems to be some confusion in your question. The winner of American Idol in 2006 was Taylor Hicks. He won the show's fifth season, not in 1974.

Gemma 2B IT			
Question- Which country ratified the Berne convention on copyright in March 1989?		Answer- The United States of America	
Relevant_Context_provided	Relevant_context_output	Wrong_date_context_provided	Wrong_date_context_output
The United States of America ratified the Berne Convention on copyright in March 1989.	The passage does not specify which country ratified the Berne Convention on copyright in March 1989, so I cannot answer this question from the provided context.	The United States of America ratified the Berne Convention on copyright in March 1969.	The country that ratified the Berne Convention on copyright in March 1989 was China.

Trained Gemma 2B output

			prededctions			
	Question	answer	no_context	rel_context	rand_context	wd_context
no_context	January 1st, 2002, saw which currency being used in several European countries including France, Germany and Ireland?	The Euro	The Eurocurrency was being used in several European countries including France, Germany and Ireland on January 1st, 2002.	The euro (€) became the official currency for 12 member states of the European Union, including France, Germany, and Ireland.	The Eurocurrency was being used in several European countries including France, Germany and Ireland.	The euro (€) became the official currency for 12 member states of the European Union, including France, Germany, and Ireland on October 1st,
rel_context	In 1998, Exxon and Mobil merged and form what new company?	Exxon mobil	Sure, in 1998, Exxon and Mobil merged to form the new company, ExxonMobil.	Sure, In 1998, Exxon and Mobil merged to form the new company known as ExxonMobil.	The passage does not mention Exxon or Mobil, so I cannot answer this question from the provided context.	Sure, in 1998, Exxon and Mobil merged and formed the new company known as ExxonMobil.
rand_context	Chinese troops kill hundreds of pro-democracy students in which city in June 1989?	Beijing	The premise of the question is incorrect. There is no evidence to support the claim that Chinese troops killed hundreds of pro-democracy students in June 198	Sure, in Beijing, China. The Tiananmen Square Massacre was a major event in June 1989 in which hundreds of pro-democracy students	The context does not provide any information about the city in which Chinese troops killed hundreds of pro-democracy students in June 1989, so I cannot	The context does not provide any information about the city where Chinese troops killed hundreds of pro-democracy students in June 1989, so I cannot answer
wd_context	In 1998 120 countries agreed to set up what?	The Intern	Sure, in 1998 120 countries agreed to set up the **G8**, a group of developed countries that meet regularly to discuss	Sure, in 1998 120 countries agreed to set up the International Criminal Court (ICC). The ICC was created as a permanent tribunal	The passage does not specify what 120 countries agreed to set up, so I cannot answer this question from the provided context.	Sure, In 1893, 120 countries agreed to establish the International Criminal Court (ICC). The ICC was created as a permanent tribunal