Additional Plots of F1 Scores on HellaFresh

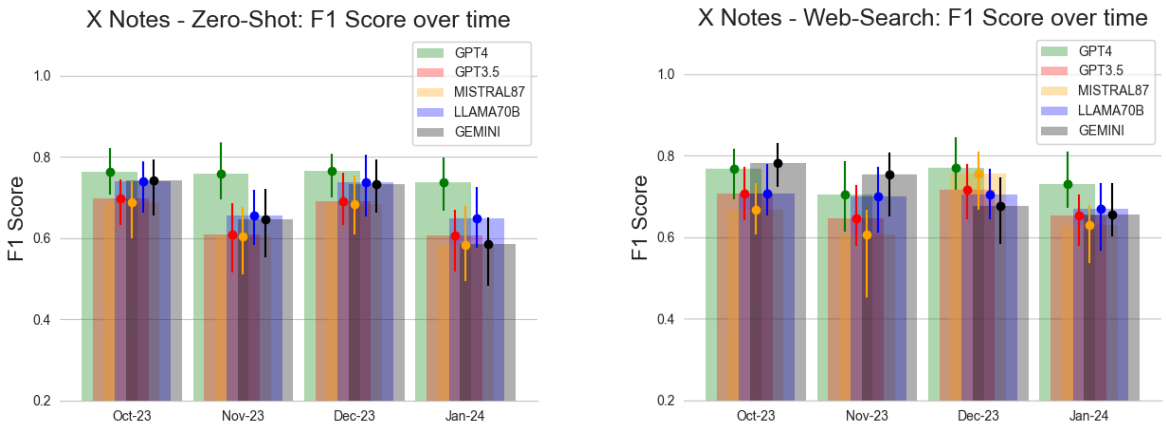

Figure 4: Results for F1 Scores on X notes: Zero-shot classification on the left and web-search agent on the right. We observe that GPT4 outperforms all models in zero-shot classification, while, e.g., Mistral8x7 consistently underperforms all other models as a web-search agent.

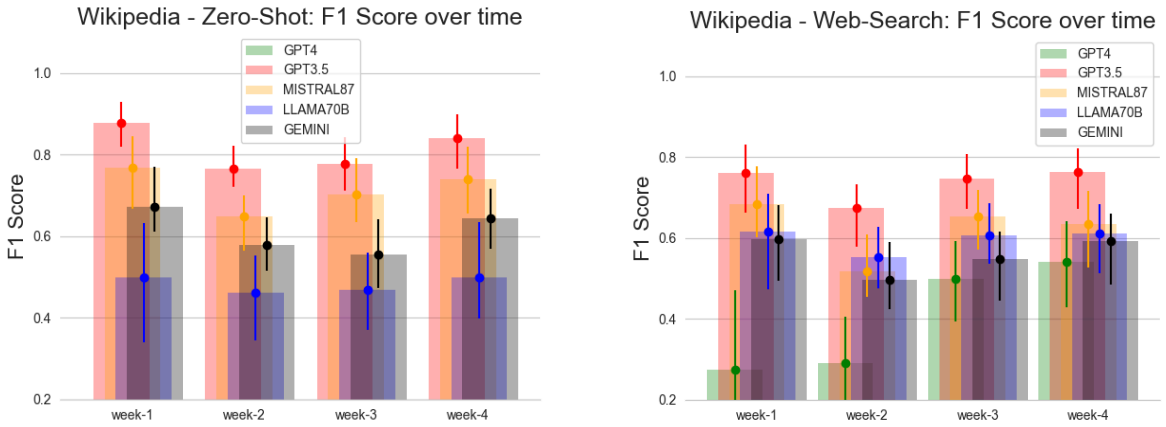

Figure 5: Results for F1 Scores on Wikipedia edits: Zero-shot classification on the left and web-search agent on the right. GPT3.5 and Mistral8x7 consistently outperform other models. Results for web-search agents are higher in variance.

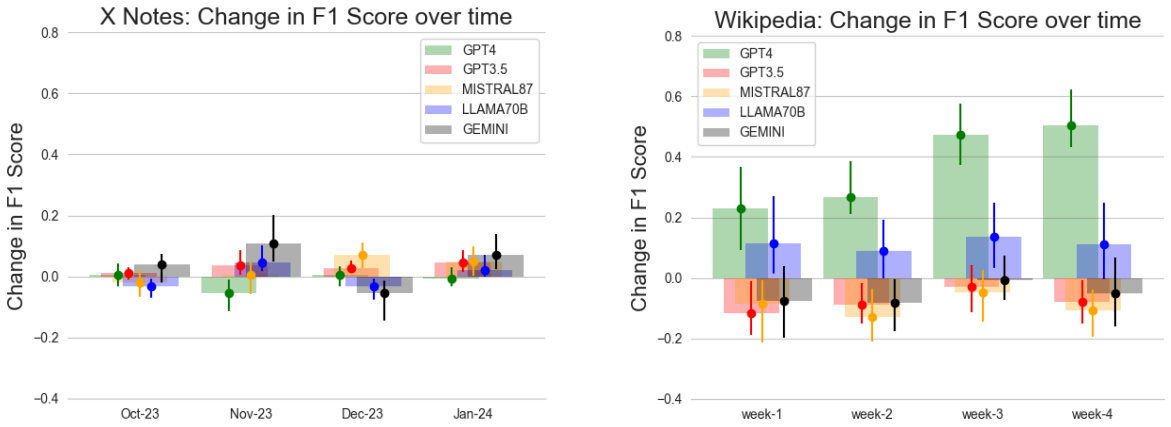

Figure 6: Overview of the difference in F1 Score between the web-search agent and the zero-shot classifier (left: X notes, right: Wikipedia edits). Web-search improves the F1 Score of most LLMs on X, with some exceptions. However, for Wikipedia edits, web-search significantly decreases the performance of GPT3.5 and Mistral8x7, while significantly increasing the F1 Score of GPT4.

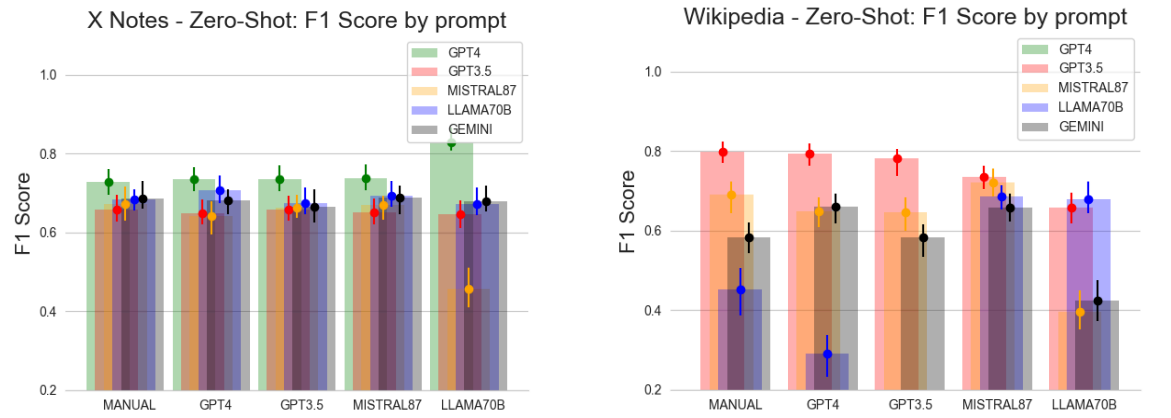

Figure 7: F1 Score (averaged over all time intervals) of all LLMs for different prompts (x-axis). Performance is highly sensitive to prompt wording. For example, changing from the manual prompt to the LLAMA70B prompt decreases the F1 Score of Mistral8x7 by 19% while increasing the F1 Score of GPT4 by 11%.