APPENDIX

This document contains a set of supplemental materials: (A) We conduct a trio of experiments to help validate the usefulness of PROMPTAID's human-in-the-loop approach, by demonstrating how various fully automated strategies can fail. (B) We report the performance of prompt instructions that were derived during Use Case #1 in the main paper. (C) We report the computational time required to compute the accuracy for the test samples used in our paper, to help demonstrate some of the backend computational demands for tools like PROMPTAID. (D) In addition to the user study reported in the main paper, we report a second user study that was conducted, which compares PROMPTAID against two "non-Chatbot-style" alternative baselines.

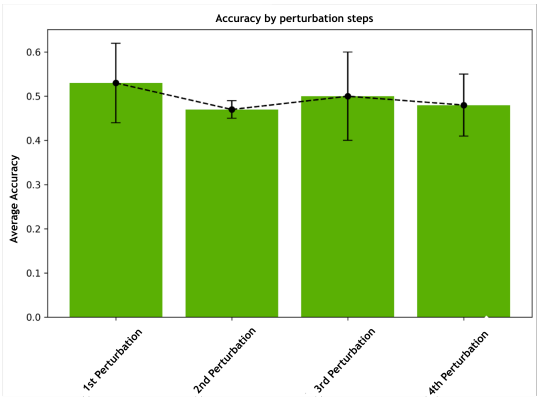

Fig. 8: Prompt vs Accuracy chart on performing keyword-based perturbation on each step

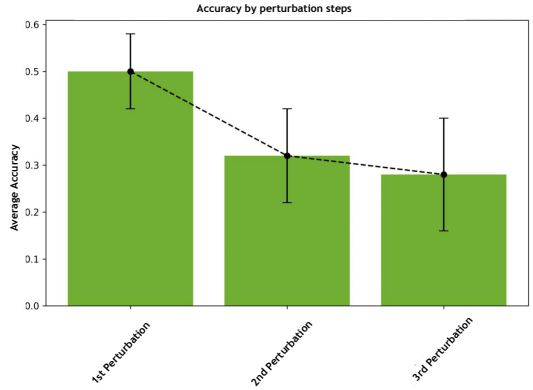

Fig. 9: Prompt vs Accuracy chart on performing paraphrasing based perturbation on each step. Initial Prompt: *What label best describes this news article?*

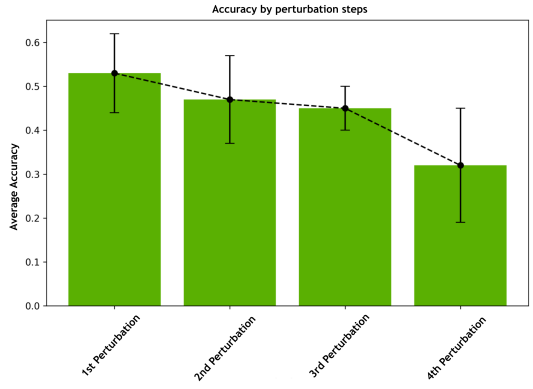

Fig. 10: Prompt vs Accuracy chart on performing alternate keyword + paraphrasing based perturbation on each step. We first start with keyword alteration. Initial Prompt: *What label best describes this news article?*

A THREE EXPERIMENTS TESTING AUTOMATIC PERTURBATIONS OF PROMPTS WITHOUT HUMAN INTERVENTION

To help motivate the need for PROMPTAID's human-in-the-loop workflow, and to help demonstrate that "automated only" solutions (i.e.., without human intervention) have limitations, we conducted a trio of experiments in automatically perturbing prompts.

Each experiment focused on the evaluation of specific perturbation types. These experiments were categorized as follows: (a) solely keyword-based perturbations, (b) solely paraphrase-based perturbations, and (c) a combination of alternating keyword and paraphrase-based alterations. The termination criteria for each experiment were delineated by four conditions: (a) modification of task semantics for all perturbed prompts, (b) introduction of grammatical errors in all the perturbed prompts, (c) achievement of an accuracy level below 35%, or (d) no recommendations were provided.

The experimental evaluation entailed the utilization of prompt instructions extracted from the topic classification dataset: *What label best describes this news article?* which has a 51.3% accuracy on a 200 data point test set. These instructions were employed in testing the LLaMA-2 13b model, using a designated test dataset consisting of 200 data points.

In each step of the perturbation, we chose the best-performing prompt and averaged over all perturbations possible until the prompt which performed the best met one of the termination criteria was met.

The outcomes of our experiments are visually represented in Figures 8–10. The height of the bar chart indicates the average performance of all the prompt recommendations; the error bar shows the standard deviation. In almost all iterations, a significant decline is observed (the exception to this is the keyword-based perturbations in Figure 8). Moreover, while some prompt perturbations did have comparable accuracy to the initial prompt, a human is needed to ensure that the chosen prompt is not only high-performing but contains correct task semantics and lexical correctness.

This consideration is highlighted in the second experiment (shown in Figure 9) that perturbs prompts by paraphrasing them. At each step, the average accuracy drastically decreases due to changes in task semantics (for a majority of recommended prompts). For example, in the third step perturbation, some of the recommended prompts instructed the LLM to perform a summarization task instead of the desired topic classification task (the best-performing prompt from this set was *"Tell me the best description of this news article?"*).

Similarly, in the third experiment (Figure 10) that alternated keyword+paraphrasing perturbations, some prompts had comparable and even better accuracy than the initial prompt, but many of these perturbed prompts encountered a similar issue of changes in task semantics and introduced grammatical errors. The best performing perturbed prompt in the fourth step was, "*Which standard is the best for this news article?*" with 45% accuracy. Despite being lower than the initial prompt, this altered prompt seems to still perform well regardless of its semantics. Such performances are hugely dependent on the LLM's internal representation, which helps indicate the criticality of having human oversight and steering.

Interestingly, in the first experiment (Figure 8) which solely perturbed keywords, the average accuracy remained relatively flat (with slight dips and increases). This phenomenon can plausibly be attributed to the relatively limited contextual changes induced by keyword-based modifications, which typically involve single-word alterations, in contrast to the more extensive changes introduced by the paraphrasing perturbations used in the other two experiments.

To summarize, these three experiments help underscore the idea that, while automated techniques can certainly be employed, human oversight is necessary to ensure the desired outcomes are achieved effectively. This assertion is not only substantiated by the observed declines (or lack of any real increases) in prompt accuracy in the fully automated approaches tested in the experiments, but also by the need for nuanced and **contextually informed adjustments** that can easily elude such automated-only strategies.

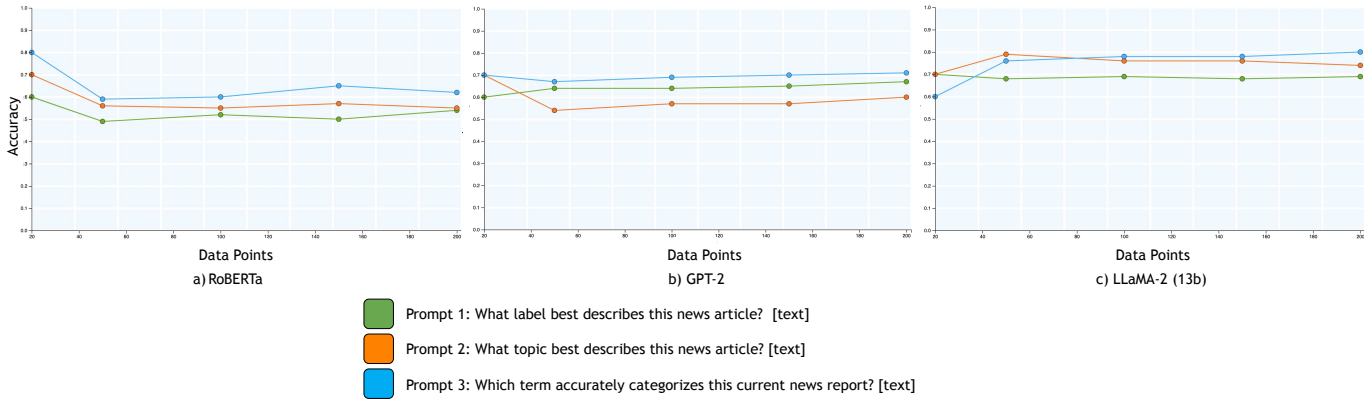

Fig. 11: Number of data points vs Accuracy scores for prompts in Use Case 1, calculated on LLaMA-2 13 billion on 200 data points. Prompt 1: *What label best describes this news article? [text]* Prompt 2: *What topic best describes this news article? [text]* Prompt 3: *Which term accurately categorizes this current news report? [text].* Prompt 3 consistently outperforms the prior prompts on all the 3 language models.

B Prompt Evaluation for Use Case #1

This section reports the efficacy of the prompt instructions derived from Use Case #1 (see Figure 3). An evaluation of these prompt instructions in the Use Case was conducted using RoBERTa with a test set consisting of 20 samples. The overarching objective was to ascertain how well the prompt instructions from the use case generalize when applied to a more extensive test dataset. To this end, the prompt instructions were also systematically tested across three distinct language models: RoBERTa, GPT-2, and LLaMA-2, across a range of data points spanning from 20 to 200. Detailed results of this evaluation are presented in Figure 11.

The principal finding from this investigation is that the ultimate prompt instruction denoted as "Prompt 3," consistently exhibited superior performance in comparison to its antecedent prompt iterations across all the aforementioned language models.

C Time taken for LMs to produce outputs

Table 1 presents the computational time, measured in seconds, required by RoBERTa, GPT-2, and LLaMA-2 13 billion, to compute accuracy metrics for the test samples discussed in Use Case #1, for 20 data points and 200 data points. As the size of the testing set grows, the computational workload also increases, resulting in extended inference times. Such escalation in computational demands (and potentially an escalated feedback loop with a human user) can be addressed by future software engineering efforts on the software side.

Language Model	Data Points	Topic classification	Sentiment analysis	CSQA
RoBERTA	20	12.64	20.32	18.34
	200	99.86	120.76	132.49
GPT-2	20	15.05	16.22	13.87
	200	112.84	108.37	120.92
LLaMA-2	20	20.62	23.29	25.22
	200	135.19	129.62	147.67

Table 1: Time in seconds taken by language models on a test set of 20 vs 200 data points

D Testing PromptAid Against Non-Chatbot-Style Interfaces

In addition to the user study reported in Section 7, which compared PromptAid against two Chatbot-style basleine interfaces (shown in Figure 6), we also ran second user study comparing PromptAid against two "non-Chatbot-style" interfaces (shown in Figures 12 and 13). For posterity, we report that study here; at a high level, the general results and findings are similar to the study in the main paper.

Fig. 12: Alternative Baseline #1: HuggingFace-based interface

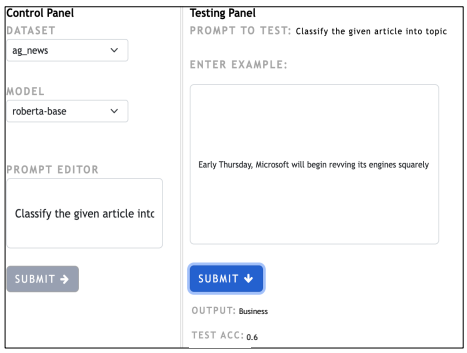

Fig. 13: Alternative Baseline #2

Baselines. For the first alternative interface (AltBaseline 1, shown in Figure 12), we used a prompting interface provided in Huggingface, as it represents a commercially available prompting interface (albeit one that omits the Chatbot-style user experience that shows the prompt history). For the second alternative interface (AltBaseline 2, shown in Figure 13), we designed a straightforward interface that features an input area for prompt templates, a box for users to input examples and additionally an accuracy output on the test dataset, to see if this metric had an impact on prompt intuitions among users. AltBaseline 2 represents a midpoint between PromptAid and the HuggingFace commercial interface (AltBaseline 1), by providing feedback which might influence a user's prompts

Data Domains, Model, and Study Design, Participants, Apparatus. We used the same datasets, language models, study design, and testing apparatus as what was done for the main user study reported in Section 7.1. The participant recruitment strategy was also identical (ten participants who were non-experts in NLP and language models), though we recruited a separate cohort for this study (average age = 24.6, SD = 1.42, 6 males, 4 females, all computer science graduate students from the Arizona State University). Study durations lasted approximately 30–45 minutes. The data collection and analysis methodology

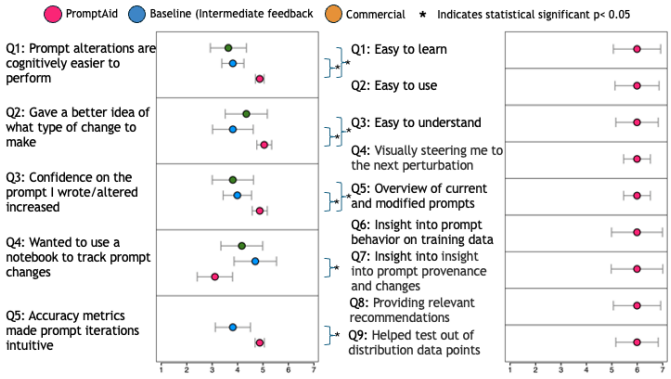

Fig. 14: Participant ratings from the second user study; median ratings are indicated in gray.

was also the same as what was done in the main paper's study.

D.1 Study Results

D.1.1 Task Stage Performance.

As in the main study, we first report on collected survey ratings about the cognitive effort required, tracking abilities and confidence of acquired prompt templates. Where applicable, we report Mann-Whitney U tests to indicate if there is a statistical difference between PROMPTAID and an alternative baseline (using a threshold of $p = 0.05$) in terms of ease of generating good performing prompts by providing U and p values.

Figure 14 summarizes survey responses. For survey questions about the cognitive effort required, knowledge about the change being made, need to track prompt template changes, and confidence in the prompt templates they reach (see Figure 14(Qn1–Qn4), PROMPTAID performed significantly better in terms of the cognitive effort required while prompting ($U = 12.5, p < 0.005$), tracking the kind of change being made ($U = 6, p < 0.005$), ease of tracking ($U = 23.5, p < 0.05$) and confidence acquired in the prompt template reached ($U = 1, p < 0.005$). That is, like the study in the main paper, compared to the alternative baselines we found that PROMPTAID was helpful for aiding prompt changes and tracking, comparing, and analyzing prompt templates over iterations. Also similar to the main study, we saw that participants using PROMPTAID were able to elevate their performance to 80% within two perturbation steps, while they were generally unable to do so with the alternative baselines.

D.1.2 Freeform Stage: User Comments and Survey Ratings.

We next report comments and feedback collected during and after the Freeform Analysis Stage. Figure 14(Qn5 and Qn1–Qn9 in the right column) shows survey feedback about the system during this stage. PROMPTAID's functionality and interface feature was highly rated by almost all the participants (as the alternative baselines were not used in this stage, they do not have corresponding ratings for these questions). Similar to the main study analysis reported in Section 7.2, we performed an open coding on participant verbalizations and discussed both positive feedback as well as some suggested system improvements below, in the context of the PROMPTAID's design goals.

(G3, G4) PROMPTAID was preferred over the alternative baselines. Similar to the main study, all 10 participants preferred PROMPTAID's visual aids for prompting, compared to the alternative baselines, and obtained better-performing prompt templates with lower cognitive effort. Similar to the main study, several participant comments emphasized this: "*The visualization panels are basically pruning my search in the prompt space. I know what change to make in the next step to make my prompt template better*" (u7). "*The recommendation panel was not only helpful in giving me new words or paraphrases, but it actually helped me think newer words which I normally wouldn't*" (u4).

Some participants explicitly described that the alternative baselines required more cognitive processing: "*I could think of one word for the first change but then it gets harder to think of more changes to the template over steps*" (u10). "*Prompting in baseline was harder as after a point I couldn't think of more changes but the interface was still giving me new suggestions for the same prompt template*" (u9). "*Thinking in baseline to come up with synonyms was still fine but paraphrasing was hard. It was a bit uncomfortable*" (u2). These comments echo the Likert scale ratings in Figure 14.

(G1, G2, G3, G4, G5) PROMPTAID supported improving and tracking prompt templates via linked panels. Also similar to the main study, participants used all six of PROMPTAID's linked panels to iterate and track prompt template performances. Several participants here also mentioned the usefulness of the Prompt Canvas Panel as an overview and comparison strategy: "*In baseline I couldn't make sure if my prompt was doing well on just one point or on a bunch of other points, Prompt canvas was really useful*" (u2). "*[The] prompt canvas was really useful and helpful to track the changes I made to the template, it sort of acted like a global view unlike baseline where I didn't know how well my prompt was performing*" (u1).

Four participants (u4, u5, u7, u9) especially liked the Perturbation Sensitivity Panel. "*The sensitivity panel was acting like a direction to let me know which way or change I need to make. I didn't have to think of it, unlike baseline*" (u7). They also found the Recommendation Panel to be useful: "*The Recommendation Plot was useful for me. I could think of newer words and also in my perturbations, I found words or paraphrases farther from the red dot perform way better than recommendations near the red dot*" (u6). "*Since I did the visual interface first I knew which word to change for the prompt template in the baseline interface but if I hadn't known this it would have taken me longer to think of this for the baseline*" (u5).

Regarding the Data and Testing Panels, all participants provided feedback that it was beneficial for analyzing and validating prompt templates. "*I used the Data Panel to see if the logits were biased to a certain class or not and used it to compare new iterations of the same prompt.*" (u6). "*The confusion matrix is very useful, all I need to keep in mind is make the diagonals the darkest in color*" (u8). "*This is similar to baseline, but testing data points along with all these other panels is more useful than just blindly testing like I was doing in the base interface, and now I am more sure of my prompt when I test*" (u8).

(G3) More user controls for in-context examples for experimentation purposes. All ten participants found the in-context example recommendations useful to improve the performance of the prompt template, and many commented that it increased their creativity: "*I never knew adding examples can increase the accuracy of the same prompt by so much, and the fact that I don't have to think of those examples is very convenient*" (u10). "*I trust the backend with these optimal K values and the examples appended to the test data point, I would have never thought of those on my own.*"(u4).

One suggestion, made by four participants (u1, u2, u8, u10), was to have a panel that allowed users to pick the k-shot examples themselves from a larger set to determine the best-performing prompt. "*I actually trust that the system has come up with good examples for the k-shot setting, the accuracy here increased by 40% but it wouldn't hurt to have an option where I too can play with examples I want to enter instead of just accepting the suggestions*" (u10). "*I like the k-shot example suggestions and would definitely use it. It performs the best among all the changes I made to the prompt. But if I enter fake news to see if the prompt template does well or not I would want an option to enter fake examples for them as well just to play around*" (u8). We considered the approach during PROMPTAID's prototyping (see Section 4.2), but it was omitted as we wanted to not overwhelm non-technical users. However, even without controls to choose their own examples, participants were satisfied with the examples recommended for the test data point.

Peaking inside the LLM to contextualize behavior. One suggestion, made by three participants (u2, u3, u8) was to provide explainability in terms of the saliency of the text being entered into the LLM, to understand what words were focused on more during predictions. "*I wish there was a panel that showed words that were more important for the [LLM] while giving out a prediction. The outputs match my expectations, but I would want to see something like text highlights as well*" (u2). "*Knowing why the model works on one prompt and doesn't*

on another is something I would like to see" (u8). While we agree that a gradient-based saliency would aid users, computing such saliency for LLMs is computationally prohibitive under current computational methods; if this is addressed in future work, it would make sense as an additional to this type of tool.

On-demand training to improve usability. Similar to the main study, participants found PROMPTAID easy to learn and use, though two participants (u3, u8) had similar feedback that additional training time and reference guides would improve their abilities to use the system.

D.2 Study Takeaways.

This study was omitted from the main paper as the Chatbot-style baselines were considered more "relevant" to everyday usage and thus served as a better realistic comparison against PROMPTAID. Nevertheless, the results of both studies were highly similar in terms of the findings, insights, and overall takeaways. Succinctly, in both studies, PROMPTAID's human-in-the-loop and visualization-based approach provided significant advantages over the baselines in terms of supporting users to improve prompt performance and understand how to variate, experiment with, and modify prompts.