Understanding the Relationship between Prompts and Response Uncertainty in Large Language Models

Anonymous ACL submission

Abstract

Large language models (LLMs) are widely used in decision-making, but their reliability, especially in critical tasks like healthcare, is not well-established. Therefore, understanding how LLMs reason and make decisions is crucial for their safe deployment. This paper investigates how the uncertainty of responses generated by LLMs relates to the information provided in the input prompt. Leveraging the insight that LLMs learn to infer latent concepts during pretraining, we propose a prompt-response concept model that explains how LLMs generate responses and helps understand the relationship between prompts and response uncertainty. We show that the uncertainty decreases as the prompt's informativeness increases, similar to epistemic uncertainty. Our detailed experimental results on real datasets validate our proposed model.

1 Introduction

Large language models (LLMs) have demonstrated impressive performance across a variety of tasks (Google, 2023; OpenAI, 2023; Zhao et al., 2023). This success has led to their widespread adoption and significant involvement in various decision-making applications, such as healthcare (Karabacak and Margetis, 2023; Sallam, 2023; Yang et al., 2023), education (Xiao et al., 2023), finance (Wu et al., 2023b), and law (Zhang et al., 2023a). However, despite their rapid adoption, the reliability of LLMs in handling high-stakes tasks has yet to be demonstrated (Arkoudas, 2023; Huang et al., 2023a). The reliability is particularly critical in domains such as healthcare, where model responses can have immediate and significant impacts on human behavior and hence their well-being (Ji et al., 2023). Therefore, understanding LLMs' reasoning and decision-making processes and how they influence response uncertainty is critical for their safe deployment.

To understand this importance, consider the mobile health (mHealth) application in which machine learning algorithms are integrated to monitor users' health conditions and provide advice on daily activities (Boursalie et al., 2018; Trella et al., 2022, 2023). Providing suggestions that can influence users' health is a form of intervention in the human decision-making process. For LLMs to be suitable for such use cases, they should be accurate and provide consistent intervention strategies, e.g., consider an LLM-powered mHealth app that suggests physical therapy (PT) routines to a patient recovering from surgery. The app's goal is to ensure the patient adheres to their PT regimen during rehabilitation despite the discomfort it may cause. The app must provide consistent suggestions to encourage PT adherence. Any inconsistent behaviors from the app could undermine any progress made. Conversely, providing accurate and consistent responses helps make the system more reliable and trustworthy (Shin et al., 2022).

The response generated by LLMs is a series of tokens sampled from probability vectors of tokens using various heuristics (Brown et al., 2020; Radford et al., 2018, 2019), such as beam search, nucleus sampling, and greedy decoding. Typically, tokens with higher probabilities are chosen sequentially to produce the final response. The response variations are controlled by LLM parameters such as temperature (T), top-k, or top-p. While response variations benefit creative tasks like poem and essay writing, they can be detrimental for tasks requiring high reproducibility and consistency (Ganguli et al., 2022; Huang et al., 2023b). However, making LLMs generate deterministic responses is not ideal, as users may vary in what responses suit them the most (Wu et al., 2023a). Hence, a better approach is needed to understand the sources of response uncertainty and develop methods to reduce it naturally rather than masking it by adjusting LLM parameters.

We attribute response uncertainty to two main factors: the LLM's parameters controlling the generated response's randomness and the input prompt's informativeness (information about the desired task). This paper focuses solely on the response uncertainty due to the input prompt while keeping the LLM parameters fixed. Here, a natural question arises: *How is the amount of information in the input prompt related to the uncertainty in the responses generated by an LLM?*

To answer this, we leverage the insight that LLMs implicitly learn to infer latent concepts during pretraining (Xie et al., 2021; Hahn and Goyal, 2023; Zhang et al., 2023b) and propose a prompt-response concept (PRC) model. Our PRC model conceptualizes how an LLM generates responses based on given prompts, and helps understand the relationship between prompts and response uncertainty by measuring response uncertainty for prompts with varying information about the task. We provide theoretical results that show that the uncertainty of responses generated by an LLM decreases as the informativeness of prompt increases (i.e., having more information about the task). We connect response uncertainty and epistemic uncertainty and show that adding relevant information to the prompt is a principled and effective method to reduce this uncertainty. Finally, we corroborate the validity of our PRC model via experiments and provide a simulation for a healthcare use case to demonstrate the efficacy of our approach.

2 Prompt-Response Concept Model of Large Language Model

In this section, we first define what we mean by *concept*. We then use the notion of concept to explain our proposed prompt-response concept model of LLM. Finally, we provide theoretical results that explain the relationship between the uncertainty of response generated by an LLM and the representation quality of the prompts.

We define the *concept*[1] as an abstraction derived from specific instances or occurrences that share common characteristics (Laurence and Margolis, 1999; Fodor, 1998; Weiskopf, 2009; Wilmont et al., 2013). To understand the notion of concept, consider the following example of the concept: *Species*, which includes a group of organisms that share common biological traits. Another example is the *personal bio*,[2] which consists sentences giving information about names, occupations, contributions, and other personal details.

Concept: *Personal bio* of Alan Turing

Alan Turing was an English computer scientist, mathematician, and cryptanalyst. He introduced the Turing machine, which formalized the concepts of algorithms and computation, serving as a foundational model for general-purpose computers. Turing is widely regarded as the father of theoretical computer science. . . .

Using concepts instead of word-level or token-level patterns in text analysis improves the ability to reason and answer questions based on higher-level abstractions, which allows a better understanding of the relationships between different sentences in the given text (Bates, 1995; Bogatyrev and Samodurov, 2016; Wang et al., 2024). As we can see in the example above, explaining a concept often involves multiple sentences, each contributing specific and meaningful information about the concept (Piccinini and Scott, 2006). We refer to the information in each sentence as *attributes of the concept*, e.g., the sentence "Alan Turing was an English computer scientist, mathematician, and cryptanalyst" gives information about the name, nationality, and occupation of Alan Turing.

2.1 Prompt-Response Concept Model

Our aim is to understand how the input prompt is related to the uncertainty in the responses generated by an LLM. To do so, we first introduce notations representing different variables used in this section. Let $\mathcal{X}$ denote the set of all prompts and $\mathcal{Y}$ denote the set of all responses generated by an LLM f, where $f : \mathcal{X} \rightarrow \mathcal{Y}$. For a given prompt $x \in \mathcal{X}$, the LLM f generates a response $y \in \mathcal{Y}$ such that $y = f(x)$. Since the response y can vary each time the LLM generates it, we attribute these response variations to two main factors: the LLM's parameters, such as temperature T, top-k, or top-p, which control the randomness in the generated tokens, and the informativeness (information about the desired concept) of the given prompt. This paper focuses solely on the latter aspect while keeping the LLM's parameters fixed.

As it has been shown that LLMs implicitly learn to infer latent concepts during pretraining

[1] The definition of a concept varies across fields, e.g., in philosophy, a concept represents the fundamental unit of thought; in psychology, it is a mental construct; in linguistics, it refers to the semantic units that words or phrases represent; and in education, it denotes key ideas or principles.

[2] The *personal bio* example of the concept is adapted from *wiki bio* concept example given in Xie et al. (2021).

(Xie et al., 2021; Hahn and Goyal, 2023; Zhang et al., 2023b), we use this insight to propose the prompt-response concept (PRC) model of LLM. This model conceptualizes how an LLM generates a response for a given prompt, which will be used to understand the relationship between prompts and the response uncertainty by measuring response uncertainty for prompts with varying information. The PRC model has three main components (as shown in Figure 1): prompt concept, response concept, and mappings.

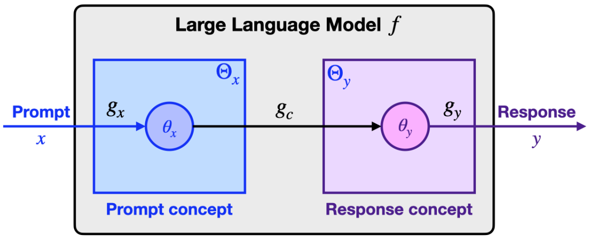


Figure 1: Prompt-Response concept model of LLM.

Prompt concept. Let Θ_x be the set of all concepts corresponding to prompts in set $\mathcal{X}$. In the PRC model, we assume that each input prompt $x \in \mathcal{X}$ corresponds to a concept. We refer to this concept as the *prompt concept* $\theta_x \in \Theta_x$. Intuitively, an LLM recognizes input tokens as semantically meaningful units that coherently describe an attribute of some latent prompt concept. The attributes of a concept are expressed through multiple semantically meaningful sentences. If multiple sentences in the prompt cannot be combined to describe a single concept, the LLM will treat them as representing different concepts. Our experimental results in Figure 2d show that adding semantically meaningful sentences from different concepts can increase response uncertainty.

Response concept. Let Θ_y be the set of all concepts corresponding to responses in set $\mathcal{Y}$. We refer to these concepts as the *response concept*. The PRC model assumes that each response concept $\theta_y \in \Theta_y$ corresponds to a response $y \in \mathcal{Y}$.

Mappings. To understand the relationship between input prompt, intermediate concepts, and response, we assume the LLM f is a composition of three mappings/functions: prompt-concept mapping (g_x), concept-concept mapping (g_c), and concept-response mapping (g_y). Therefore, we can represent the response as $y = f(x) = g_y(g_c(g_x(x)))$, where the function g_x maps the input prompt to a prompt concept, then function g_c maps the prompt concept to a response concept, and finally, the function g_y maps the response concept to a response.

In the example of a personal biography, the prompt contains text providing detailed information about a person (e.g., a comprehensive bio of Alan Turing), which corresponds to the *personal bio* concept. Additionally, the prompt includes specific tasks (e.g., tell me Alan Turing's main contribution between 1940 and 1945) for which the LLM needs to generate a response. To generate a response, an LLM first maps the input prompt to a prompt concept which is then maps to a corresponding response concept. Finally, the LLM uses response concept to generate the final response. When a prompt lacks sufficient task-related information (i.e., it is less informative), we can expect higher variability in the responses generated by the LLM, as corroborated by our experimental results in Figure 2b. To further understand this relationship, we will discuss how the informativeness of prompts is related to response uncertainty.

2.2 Relationship between Prompts and Response Uncertainty in LLMs

Let $\mathcal{X}_{\theta_x} \subset \mathcal{X}$ be the set of prompts with the same semantic meaning (i.e., conveys some form of information)[3] and contain all information of the prompt concept θ_x. Let $\mathcal{X}_s \subset \mathcal{X}$ be the set of prompts with the same semantic meaning s and only contain partial information about the prompt concept θ_x. We use the notation $x_1 \prec_{\theta_x} x_2$ to indicate that prompt x_1 contains less information about prompt concept θ_x than prompt x_2 (or prompt x_2 is more informative than prompt x_1). By definition, any prompt from the set $\mathcal{X}_s$ contains less information about prompt concept θ_x than any prompt from the set $\mathcal{X}_{\theta_x}$.

Let Z_c be a random variable denoting concept (where $c = x$ for prompt concept and $c = y$ for response concept) and X_s be a random variable representing a prompt having semantic meaning s. Here, the randomness in Z_c is due to a less informative prompt, which leaves more space for interpretation or variation in the possible concepts that LLM can map. In contrast, the randomness in X_s is due to the ability of different prompts representing the same semantic meaning.

[3] Multiple prompts can be generated from a single prompt by paraphrasing it while preserving the original semantic meaning associated with the prompt (Kuhn et al., 2023).

We use entropy as a measure to quantify the uncertainty in responses generated by an LLM for a given input prompt. Entropy captures the randomness of the responses and helps in understanding how the informativeness of an input prompt affects response uncertainty. Let Y be a random variable representing response. The randomness in Y can be due to less informative prompts and the ability of different responses to represent the same semantic meaning. For a prompt x, we define entropy of Y as follows:

$$\mathrm{H}\left(Y|x\right) = -\sum_{y\in\mathcal{Y}} p(y|x)\log_2 p(y|x), \qquad (1)$$

where $p(y|x)$ is the conditional distribution of the responses generated for a prompt. Intuitively, a highly informative prompt corresponds to specific intermediate concepts (prompt and response concept), which leads to the generation of responses with less variability and, hence, smaller entropy of Y. The conditional distribution $p(y|x)$ represents the posterior predictive distribution, which marginalizes all intermediate concepts (prompt and response) and is given as follows:

$$\begin{aligned} p(y|x) &= \int_{\theta_y} p(y|\theta_y, x)p(\theta_y|x)d\theta_y \\ &= \int_{\theta_y}\int_{\theta_x} p(y|\theta_y, x)p(\theta_y|\theta_x, x)p(\theta_x|x)d\theta_y d\theta_x. \end{aligned}$$

The first equality follows from conditioning response with respect to the response concept, and the second equality follows by using

$$p(\theta_y|x) = \int_{\theta_x} p(\theta_y|\theta_x, x)p(\theta_x|x)d\theta_x.$$

If $p(\theta_c|x)$ (where $c = \{x, y\}$) concentrates on a specific concept with a more informative prompt, the LLM learns effectively via marginalization. This behavior implies that the LLM implicitly performs Bayesian inference, which is also observed in in-context learning (Xie et al., 2021).

2.3 Theoretical Results

We need the following assumption under which our theoretical results hold.

Assumption 1. *We use the following assumptions:*

1. *The LLM know the exact mappings/functions, i.e., g_x, g_c, and g_y.*
2. $H\left(Z_x|X_{\theta_x}\right) = 0$ *and* $H\left(Z_y|Y_{\theta_y}\right) = 0$.
3. $\forall\theta\in\Theta_x$, *there exits a non-empty set* $\mathcal{X}_\theta$.
4. *For* $\theta_1, \theta_2 \in \Theta_x$, $\mathcal{X}_{\theta_1}\cap\mathcal{X}_{\theta_2} = \emptyset$ *if* $\theta_1 \neq \theta_2$.

The first assumption states that LLMs perfectly know the mappings used in the PRC model. While this assumption may not hold in practice, a better LLM has good estimates of these mappings, as corroborated by our experimental results shown in Figure 2e and Figure 2f. The first part of the second assumption says there is no randomness in the prompt concept if all the information needed to respond to the task is contained in the prompt, and the second part says that given a complete output, there is a unique response concept; in other words, no two concepts share the exact same semantic description. The third assumption ensures that some semantically meaningful text corresponds to each prompt concept. Finally, the fourth assumption ensures that prompts fully describing two different concepts can not have the same semantic meaning. Compared to the first assumption, the assumptions $2-4$ are easier to hold in practice. Next, we present our first result, which shows the relationship between concept uncertainty and informativeness of prompts.

Proposition 1. *Let Assumption 1 hold. Then,* $H\left(Z_x|X_s\right)$ *strictly decreases as the* X_s *represents more informative prompts.*

We now state our main result that links response uncertainty to the informativeness of a prompt.

Theorem 1. *Let Assumption 1 hold. Then,* $H\left(Z_y|X_s\right)$ *strictly decreases as* X_s *represents more informative prompts. Further,* $H\left(Y|X_s\right)$ *converges to* $H\left(Y|Z_y\right)$.

The proofs of Proposition 1 and Theorem 1 are given in Appendix A. These results suggest that as the prompt's informativeness increases, the response uncertainty due to the uncertainty in the response concept decreases. Furthermore, when sufficient information is provided in a prompt, there will not be any uncertainty due to the uncertainty in the concept. The only source of randomness in responses is the ability of different responses to convey the same semantic meaning.

2.4 Concept Uncertainty as Epistemic Uncertainty

In machine learning literature, epistemic uncertainty is typically reduced by incorporating additional information, such as using a better model and additional training data (Hüllermeier and Waegeman, 2021; Lahlou et al., 2021). In Proposition 1, $\mathrm{H}\left(Z_c|X_s\right)$ represents the epistemic uncertainty in latent concepts. We have demonstrated that $\mathrm{H}\left(Z_c|X_s\right)$ is strictly reduced with an informative prompt. Therefore, increasing the information about the concept in a prompt can lead to more reliable and consistent responses by reducing the epistemic uncertainty in the latent concept. When the prompt perfectly captures the desired concept, the posterior distribution of the concept given prompt converges to the desired concept; the remaining uncertainty is irreducible due to numerous ways of characterizing the same concept. Note that this uncertainty is not detrimental in general for the purpose of getting the desired information. However, if the prompt contains sentences that are irrelevant to the task at hand (i.e., the more information provided is irrelevant), the response uncertainty can increase, as demonstrated in Figure 2g.

3 Experiments

To validate our proposed prompt-response concept model of LLM, we empirically demonstrate different aspects of our proposed model in different settings whose details are as follows.

3.1 Relationship between Informativeness of the Prompt and Response Uncertainty

We begin by assessing the response uncertainty of LLMs through the generation of responses using increasingly longer prompts with more relevant information For each prompt, we generate 100 responses from LLM with uncalibrated logits ($T = 1$) and project them into the embedding space as single points using the OpenAI "text-embedding-ada-002" model. To quantify the uncertainty in the generated responses for a given prompt, we use the *total standard deviation*, denoted as $M(x)$, defined as $\sqrt{\mathrm{Tr}(\Sigma)}$, where Σ represents the covariance matrix of the embedding vectors of responses $y_1, y_2, \ldots, y_{100}$. It is noteworthy that $\mathrm{Tr}(\Sigma)$ is also referred to as *total variation*, serving as a lightweight measure of dispersion in the data (Ferrer-Riquelme, 2009). This metric is applicable for responses generated from both black-box and white-box LLMs, as it does not require access to logits.

As illustrated in Figure 2a, longer prompts with more task-related information resulted in reduced response uncertainty. In the extreme case of an empty prompt (yellow bar), the responses vary greatly in semantic meaning (see **??**).[4] The results suggest a strong negative correlation between the informativeness of the input prompt and the response uncertainty. For a detailed examination of the relationship between input informativeness and response uncertainty, we use prompts varying with more task-related information that resulted in smaller response uncertainty as shown in Figure 2b. The lack of observable trend from bar 2 to bar 3 and from bar 4 to bar 5 could be due to adding redundant information to the input (see Appendix B.3 for details of all prompts and LLM model used). We also run an additional experiment with two prompts containing different amounts of information for a given task (see Appendix B.4 for prompts) in which different uncertainty measure is used. We generate N responses respective prompts and calculate the sequence-level *normalized predictive entropy* (PE) (Wagle et al., 2023): $\mathrm{PE}(Y|x) = -\frac{1}{N}\sum_s p(y|x)\log(p(y|x))$, where S is the random response, the sum is taken over all responses, and N is the number of responses. [5] As we observed in Figure 2c, the responses generated with the longer prompt containing more relevant information have consistently smaller PE than those from the shorter prompt as the sample size grows.[6]

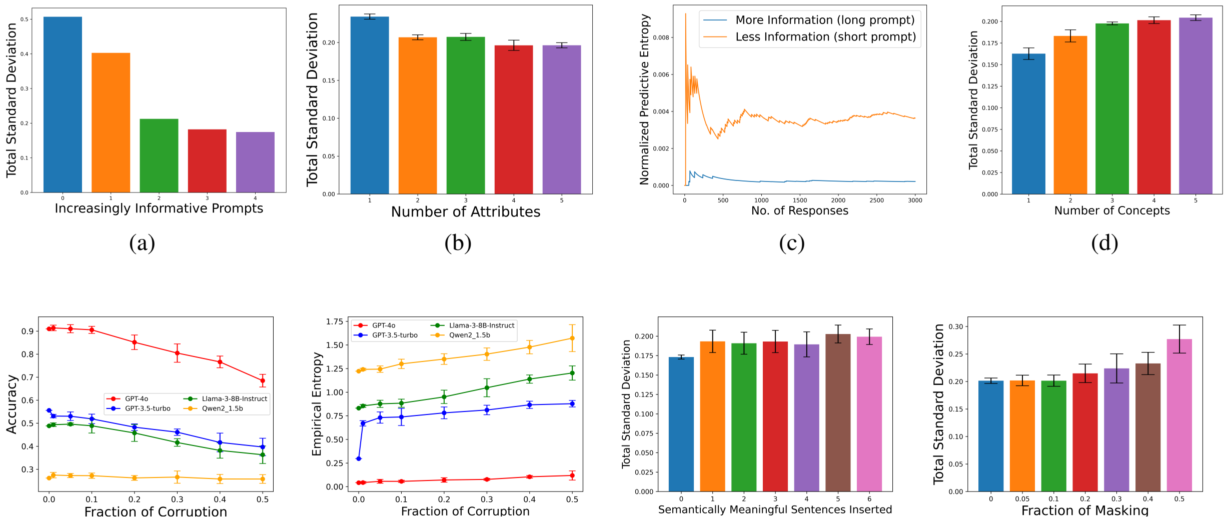
3.2 Noisy Prompts

The transformer's self-attention mechanism allows the removal of a small fraction of tokens without altering the semantic meaning by simply treating them as irrelevant tokens (Kim et al., 2017; Lin et al., 2017; Vaswani et al., 2017). Therefore, LLMs are robust to noisy tokens in prompts when the noise level is low (e.g., a few misspelled words). It is relatively easy to determine the correct word based on the context (i.e., the entire prompt). If the prompt can be accurately reconstructed, the same level of uncertainty reduction can be achieved. However, if the prompt is severely corrupted, it becomes less informative, leading to increased response uncertainty.

[4] We did this experiment in late 2023; since then, the behavior of the GPT-4-0613 checkpoint has changed, possibly due to internal fine-tuning or guard-railing by OpenAI.

[5] We model the entire generated response as the random variable instead of modeling it on the token level as in (Wagle et al., 2023). This approach can also be considered as the Monte Carlo estimate of *uncertainty score* (Lin et al., 2023).

[6] Calculating $\mathrm{PE}(Y|x)$ requires white-box model access to the logits and hence is done on meta-llama/Llama-2-7b-chat-hf from Huggingface.



(a) (b) (c) (d)

(e) (f) (g) (h)

Figure 2: **Top row:** (a), (b): *Total Standard Deviation* for input with different levels of informativeness; (c): *Normalized Predictive Entropy*; (d): *Total Standard Deviation* for tasks with different numbers of subtasks. **Bottom row:** (e): Accuracy over MCQ; (f): Uncertainty over MCQ; (g): Additional irrelevant information does not reduce output uncertainty; (h): Noisy input increases output uncertainty.

As shown in Figure 2h and Figure 3, when a certain fraction of the prompt is either masked out (replaced by space) or corrupted (replaced by random letters), there is a general trend of increase in *total standard deviation*. However, when the noise level is low (up to 0.1 fraction of the input length for the short input and 0.05 for the long input), there is no significant increase in the output uncertainty as expected. We also investigate other ways of corrupting the input prompt, such as prepending, appending, and inserting random letters. More details are given in Appendix C.2.

3.3 Compositionality of Concepts

A given prompt can have multiple sentences that correspond to different concepts. In such cases, the model may infer more than one concept from the prompt.[7] Assuming the prompt is decomposable and consists of k concepts, each corresponds to a distinguishable concept. When we fix the prompt's size, on average, each concept only has limited information in the prompt. Therefore, having k concept in a fixed-size prompt will result in more response uncertainty.

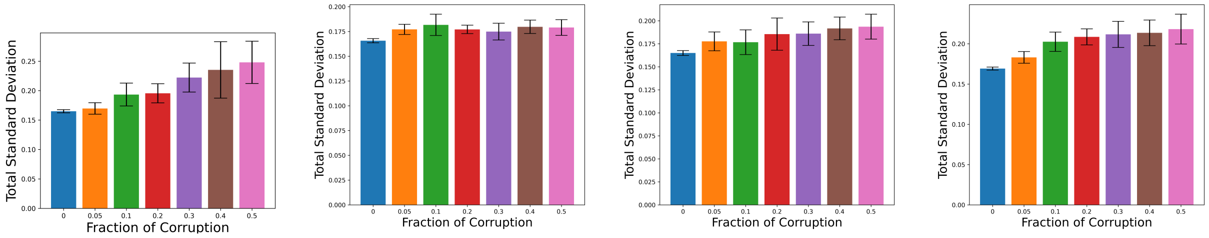
In our experiment, we consider the task of PT intervention with multiple concepts and compare the total standard deviation of the model responses with respect to the number of concept present. To test the hypothesis that a larger k leads to more response uncertainty, we ensure that the concatenated sentences have the same token count as a task with only single concept. More details are given in Appendix B.5. In Figure 2d, Prompt 1 corresponds to a single concept while Prompt 2-4 contain multiple sentences, each corresponding to one concept. Despite having the same token count, prompts with more concepts exhibit larger response uncertainty.[8] This result provides evidence for our proposed model look through the lens of the compositionality of concepts.

3.4 Relationship between Noisy Prompts and Response Quality

We selected 100 questions from the dataset and iteratively masked out an increasing fraction of randomly selected tokens from the prompt, particularly from the context of the questions. For each question, we set the temperature to 1 and sampled 100 responses from the model. We used 5 different random seeds to choose which tokens to mask,

[7] Note that it differs from having uncertainty over multiple concepts. In our earlier case, we assume all sentences are relevant to only a single concept. In contrast, in the case of uncertainty over multiple concepts, the model believes only one is relevant. When sampled multiple times, the former will consistently output all concept in the subset, while the latter will output only one concept.

[8] Experiment conducted with GPT-3.5-turbo API. Results averaged from 5 runs with 95% confidence intervals.



(a) Noisy masking (b) Noisy Prepend (c) Noisy Append (d) Append

Figure 3: Response uncertainty with respect to different noisy prompts. More details are in Appendix C.2.

replacing them with space tokens. As the fraction of masked tokens increased, we kept the same previously masked tokens and added new ones to ensure that randomness from masking did not contribute to changes in accuracy. This approach allowed us to observe the effect of token masking on the model's output quality and accuracy.

In Figure 2e, we plot the accuracy for GPT-4o, GPT-3.5-turbo, Meta-Llama-3-8B-Instruct and Qwen2_1.5B. As the fraction of masked tokens increases in prompt, the general trend is the accuracy almost monotonically decreases for all models (except Qwen2_1.5B, which already has a very low accuracy for clean input). For each random seed, we also plot the empirical conditional entropy $\mathrm{H}\left(Y|X\right)$ of the response for the given questions[9] (Figure 2f) as a measure of output uncertainty (conditional entropy is a better measure for this setting as the effective output is just one of the four choices). We observe that as corruption becomes more severe, the response uncertainty monotonically increases for all models, indicating a clear negative correlation between $\mathcal{H}Y|X$ and the response accuracy. This result corroborates our hypothesis: more relevant information leads to both a reduction in response uncertainty and an improvement in its quality. Additionally, we observe an interesting pattern: for the same prompt, a worse model always has more response uncertainty. This observation is reassuring as it suggests that, relative to better models, LLMs are not as blindly confident in their outputs as conventional wisdom holds (Groot and Valdenegro-Toro, 2024; Ni et al., 2024; Yang et al., 2024; Ye et al., 2024; Xu et al., 2024) if they are not capable of answering the given questions.

3.5 Effect of Semantically Meaningful but Irrelevant Information

Unlike random tokens, semantically meaningful sentences correspond to some concept according to our PRC model. Does this imply that adding arbitrary semantically meaningful sentences can still reduce output uncertainty? To investigate this, we observed the response uncertainty when inserting an increasing number of arbitrary sentences sampled from the Squad dataset[10] into our prompt (see Appendix B.6 for more details). As shown in Figure 2g, the response uncertainty increased with the inserted inputs compared to the original prompt.[11] The most likely behavior induced in this case, as explained in Section 3.3, is that the LLM treats useful and random inputs as independent concepts.

3.6 mHealth Intervention Setting

We now demonstrate the effectiveness of our proposed approach in a real-world simulation use case in mHealth setting. We adapt the formulation from Shin et al. (2022); both the app and the user act as reinforcement learning agents. The app agent's objective is to encourage the user agent to adhere to the PT routine. The user agent moves along a chain with N states, where a higher state number represents a healthier physical state, and state N indicates completion of the PT routine (see Figure 4). We conduct the intervention simulation experiment with LLM to compare the effect of prompts with different informativeness levels on

[9] We assume the distribution of the questions used $p(x)$ is uniform. Since there is no access to the prior of $p(y|x)$, we use the form $\mathrm{H}\left(Y|X\right) = -\sum_x p(x) \sum_y \hat{p}(y|x) \log \hat{p}(y|x)$ where $\hat{p}(y|x)$ is obtained from the empirical distribution and $\hat{p}(x) = \frac{1}{100}$ for all x in the setting.

[10] https://huggingface.co/datasets/rajpurkar/squad/viewer/plain_text/train?p=2&row=231

[11] The slight decrease in uncertainty from bar 3 to bar 4 and bar 5 to bar 6 is likely due to the model mapping some of the added sentences into one concept. Note that this does not help reduce the original task's output uncertainty, as it is still higher than the output uncertainty for the clean input. The experiment was conducted using GPT-3.5-turbo API.

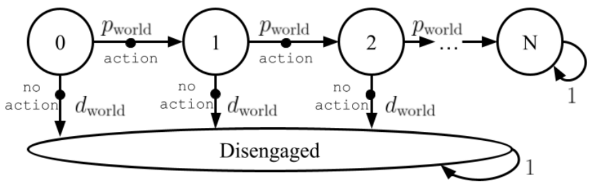


Figure 4: Visualization of states and transitions in the digital health grid world. Arrows indicate the required action and the probability of transitioning between states.

the intervention outcome. The experiment concludes that when the prompt provides the LLM (i.e., the app agent) with more information about the patient's intentions and the strategies it can employ, the efficiency of the intervention improves compared to scenarios without the additional information. A more detailed description of the experiment can be found in Appendix C.3.

4 Related Work

Uncertainty quantification for LLMs. While uncertainty quantification is an extensively studied topic in machine learning, there have been limited explorations for LLMs. The current method of quantifying response uncertainty in LLMs is predominantly limited to a calibration-based approach (Kadavath et al., 2022). The main goal of calibrating LLMs is to let the variation in the responses genuinely reflect the model's lack of relevant knowledge with respect to the prompt. (Xiao et al., 2022) and (Wagle et al., 2023) empirically investigated pre-trained language models (PLMs) and retrieval augmented language models (RALMs), respectively, and found out that while both types of models tend to be overly confident in their response, models with larger size are better calibrated. In contrast, RALMs exhibit worse calibrations compared to their counterparts. An orthogonal work (Lin et al., 2023) devised a method using similarity as determined by a Natural Language Inference (NLI) model, along with simple measures that measure dispersion based on these similarities to quantify the uncertainty and the confidence of black-box LLMs in the context of question-answering tasks. Similar to (Wagle et al., 2023), our work adopted an entropy-based uncertainty measure; however, this work focuses on investigating how to reduce response uncertainty.

Explanation for asymptotic behaviors of LLMs. There have been attempts to provide explainable frameworks to understand the surprising emergent behaviors of LLMs. For instance, (Xie et al., 2021) interprets in-context learning as an implicit Bayesian inference over *latent concepts* learned during pre-training. However, they only have a description of zero-one error where there are an infinite number of in-context examples. Moreover, their mathematical model (HMM) was designed specifically for in-context learning structure, which is unfitting for chain-of-thought or conversational-style response analysis.

In addition, despite invoking the Bayesian inference framework, their theoretical results are maximum a posteriori style, which only quantifies the mode of the posterior predictive distribution and does not touch on the uncertainty quantification aspect of the phenomenon. (Hahn and Goyal, 2023) further explored a similar idea but allowed more flexibility and complexity in the in-context examples. Similarly, they also provide an asymptotic bound on zero-one error. In contrast, we aim to complement it by quantifying how the posterior predictive uncertainty varies even when the prompt length is finite. Our framework is tailored towards aligned (i.e., instruction-fine-tuned) conversational-based LLMs, which are the prevalent type of LLMs used in practice.

5 Conclusion

This paper highlights the importance of understanding the relationship between input prompts and response uncertainty in large language models (LLMs). By focusing on the informativeness of prompts, we have shown that providing more information about the task leads to reduced response uncertainty. Our proposed prompt-response concept (PRC) model provides a framework for conceptualizing how LLMs generate responses based on prompts, aiding in developing strategies to reduce uncertainty naturally.

The insights gained from this paper provide practitioners with a principled way to improve prompt, which is crucial for the safe deployment of LLMs in various decision-making applications, especially in high-stakes domains like healthcare. Future research directions could explore further enhancements to the PRC model and investigate its application in other domains requiring reliable and consistent LLM responses.

6 Limitations

Idealistic nature of the PRC model. It is worth noting that the PRC model that we proposed in this paper assumes an idealized version of LLMs. As empirically demonstrated, while models such as GPT-3.5-Turbo, GPT-4 and Llama-2, and Llama 3 exhibit behaviors largely according to our predictions, there are still some modes in which they deviate (e.g., Qwen2_1.5b plot). This is likely in those cases where LLM does not know the mapping perfectly. For example, (Lu et al., 2021) showed that the order of examples in in-context learning influences the output quality. Our model does not capture this phenomenon. However, the authors showed that in the same work, the order of examples tends to have less effect as model quality gets better. Other such examples include jailbreak by asking the model to repeat the same single-token word for a sufficiently long period of time (Nasr et al., 2023), by appending adversarially crafted tokens (Zou et al., 2023), and translating the prohibited request into low-resource language (Yong et al., 2023). Similarly, it was observed that adversarial attacks tend to have lower success rates as the model becomes more capable. While further investigation is needed to incorporate the adversarial behavior of LLMs into this framework, the more capable LLMs are less prone to these failure modes. Our model can more effectively explain them.

LLMs for human behavior simulation. Research exploring the parallels between human behavior and reasoning patterns and those of LLMs, as well as the adaptation of LLMs as human substitutes in diverse studies, is detailed in (Aher et al., 2023; Argyle et al., 2023; Binz and Schulz, 2023; Dasgupta et al., 2022). These studies frequently demonstrate LLMs' capacity for human-like responses, leading many to regard them as viable alternatives. This paper, however, needs to delve into the appropriateness of this substitution, deferring to other works for such discussion.

Impact Statement

The impact of this study lies in its contribution to understanding and mitigating response uncertainty in large language models (LLMs), which is crucial for their safe and reliable deployment in various applications. By focusing on the relationship between prompt informativeness and response uncertainty, we provide insights into how the quality of input prompts can affect the reliability of LLM outputs. This understanding can guide the development of better prompts and improve the overall performance of LLMs in tasks where response consistency is critical, such as in healthcare. Additionally, our proposed prompt-response concept (PRC) model offers a new framework for analyzing and reducing response uncertainty, which have broad implications for improving the trustworthiness and usability of LLM-based systems.

References

Gati V Aher, Rosa I Arriaga, and Adam Tauman Kalai. 2023. Using large language models to simulate multiple humans and replicate human subject studies. In *International Conference on Machine Learning*, pages 337–371. PMLR.

Lisa P Argyle, Ethan C Busby, Nancy Fulda, Joshua R Gubler, Christopher Rytting, and David Wingate. 2023. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351.

Konstantine Arkoudas. 2023. Gpt-4 can't reason. *arXiv preprint arXiv:2308.03762*.

Madeleine Bates. 1995. Models of natural language understanding. *Proceedings of the National Academy of Sciences*, 92(22):9977–9982.

Marcel Binz and Eric Schulz. 2023. Using cognitive psychology to understand gpt-3. *Proceedings of the National Academy of Sciences*, 120(6):e2218523120.

Mikhail Bogatyrev and Kirill Samodurov. 2016. Framework for conceptual modeling on natural language texts. In *CDUD@ CLA*, pages 13–24.

Omar Boursalie, Reza Samavi, and Thomas E Doyle. 2018. Machine learning and mobile health monitoring platforms: a case study on research and implementation challenges. *Journal of Healthcare Informatics Research*, 2:179–203.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Ishita Dasgupta, Andrew K Lampinen, Stephanie CY Chan, Antonia Creswell, Dharshan Kumaran, James L McClelland, and Felix Hill. 2022. Language models show human-like content effects on reasoning. *arXiv preprint arXiv:2207.07051*.

AJ Ferrer-Riquelme. 2009. Statistical control of measures and processes.

Jerry A Fodor. 1998. *Concepts: Where cognitive science went wrong*. Oxford University Press.

Deep Ganguli, Danny Hernandez, Liane Lovitt, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova Dassarma, Dawn Drain, Nelson Elhage, et al. 2022. Predictability and surprise in large generative models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1747–1764.

Google. 2023. PaLM 2 Technical Report. *arXiv:2305.10403*.

Tobias Groot and Matias Valdenegro-Toro. 2024. Overconfidence is key: Verbalized uncertainty evaluation in large language and vision-language models. *arXiv preprint arXiv:2405.02917*.

Michael Hahn and Navin Goyal. 2023. A theory of emergent in-context learning as implicit structure induction. *arXiv preprint arXiv:2303.07971*.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. 2023a. Large language models cannot self-correct reasoning yet. *arXiv preprint arXiv:2310.01798*.

Yuheng Huang, Jiayang Song, Zhijie Wang, Huaming Chen, and Lei Ma. 2023b. Look before you leap: An exploratory study of uncertainty measurement for large language models. *arXiv preprint arXiv:2307.10236*.

Eyke Hüllermeier and Willem Waegeman. 2021. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning*, 110:457–506.

Shaoxiong Ji, Tianlin Zhang, Kailai Yang, Sophia Ananiadou, and Erik Cambria. 2023. Rethinking large language models in mental health applications. *arXiv preprint arXiv:2311.11267*.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Mert Karabacak and Konstantinos Margetis. 2023. Embracing large language models for medical applications: Opportunities and challenges. *Cureus*, 15(5).

Yoon Kim, Carl Denton, Luong Hoang, and Alexander M Rush. 2017. Structured attention networks. *arXiv preprint arXiv:1702.00887*.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *arXiv preprint arXiv:2302.09664*.

Salem Lahlou, Moksh Jain, Hadi Nekoei, Victor Ion Butoi, Paul Bertin, Jarrid Rector-Brooks, Maksym Korablyov, and Yoshua Bengio. 2021. Deup: Direct epistemic uncertainty prediction. *arXiv preprint arXiv:2102.08501*.

Stephen Laurence and Eric Margolis. 1999. Concepts and cognitive science.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. 2023. Generating with confidence: Uncertainty quantification for black-box large language models. *arXiv preprint arXiv:2305.19187*.

Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. 2017. A structured self-attentive sentence embedding. *arXiv preprint arXiv:1703.03130*.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2021. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. *arXiv preprint arXiv:2104.08786*.

Milad Nasr, Nicholas Carlini, Jonathan Hayase, Matthew Jagielski, A Feder Cooper, Daphne Ippolito, Christopher A Choquette-Choo, Eric Wallace, Florian Tramèr, and Katherine Lee. 2023. Scalable extraction of training data from (production) language models. *arXiv preprint arXiv:2311.17035*.

Shiyu Ni, Keping Bi, Jiafeng Guo, and Xueqi Cheng. 2024. When do llms need retrieval augmentation? mitigating llms' overconfidence helps retrieval augmentation. *arXiv preprint arXiv:2402.11457*.

OpenAI. 2023. GPT-4 Technical Report. *arXiv:2303.08774*.

Gualtiero Piccinini and Sam Scott. 2006. Splitting concepts. *Philosophy of Science*, 73(4):390–409.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Malik Sallam. 2023. Chatgpt utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. In *Healthcare*, volume 11, page 887. MDPI.

Motoki Sato, Jun Suzuki, Hiroyuki Shindo, and Yuji Matsumoto. 2018. Interpretable adversarial perturbation in input embedding space for text. *arXiv preprint arXiv:1805.02917*.

Eura Shin, Siddharth Swaroop, Weiwei Pan, Susan Murphy, and Finale Doshi-Velez. 2022. Modeling mobile health users as reinforcement learning agents. *arXiv preprint arXiv:2212.00863*.

Anna L Trella, Kelly W Zhang, Inbal Nahum-Shani, Vivek Shetty, Finale Doshi-Velez, and Susan A Murphy. 2022. Designing reinforcement learning algorithms for digital interventions: pre-implementation guidelines. *Algorithms*, 15(8):255.

Anna L Trella, Kelly W Zhang, Inbal Nahum-Shani, Vivek Shetty, Finale Doshi-Velez, and Susan A Murphy. 2023. Reward design for an online reinforcement learning algorithm supporting oral self-care. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 15724–15730.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Sridevi Wagle, Sai Munikoti, Anurag Acharya, Sara Smith, and Sameera Horawalavithana. 2023. Empirical evaluation of uncertainty quantification in retrieval-augmented language models for science. *arXiv preprint arXiv:2311.09358*.

Xintao Wang, Zhouhong Gu, Jiaqing Liang, Dakuan Lu, Yanghua Xiao, and Wei Wang. 2024. Concept: Concept-enhanced pre-training for language models. *arXiv preprint arXiv:2401.05669*.

Daniel Aaron Weiskopf. 2009. The plurality of concepts. *Synthese*, 169:145–173.

Ilona Wilmont, Sytse Hengeveld, Erik Barendsen, and Stijn Hoppenbrouwers. 2013. Cognitive mechanisms of conceptual modelling: How do people do it? In *Conceptual Modeling: 32th International Conference, ER 2013, Hong-Kong, China, November 11-13, 2013. Proceedings 32*, pages 74–87. Springer.

Likang Wu, Zhi Zheng, Zhaopeng Qiu, Hao Wang, Hongchao Gu, Tingjia Shen, Chuan Qin, Chen Zhu, Hengshu Zhu, Qi Liu, et al. 2023a. A survey on large language models for recommendation. *arXiv preprint arXiv:2305.19860*.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023b. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564*.

Changrong Xiao, Sean Xin Xu, Kunpeng Zhang, Yufang Wang, and Lei Xia. 2023. Evaluating reading comprehension exercises generated by llms: A showcase of chatgpt in education applications. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 610–625.

Yuxin Xiao, Paul Pu Liang, Umang Bhatt, Willie Neiswanger, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2022. Uncertainty quantification with pre-trained language models: A large-scale empirical analysis. *arXiv preprint arXiv:2210.04714*.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. 2021. An explanation of in-context learning as implicit bayesian inference. *arXiv preprint arXiv:2111.02080*.

Tianyang Xu, Shujin Wu, Shizhe Diao, Xiaoze Liu, Xingyao Wang, Yangyi Chen, and Jing Gao. 2024. Sayself: Teaching llms to express confidence with self-reflective rationales. *arXiv preprint arXiv:2405.20974*.

Haoyan Yang, Yixuan Wang, Xingyin Xu, Hanyuan Zhang, and Yirong Bian. 2024. Can we trust llms? mitigate overconfidence bias in llms through knowledge transfer. *arXiv preprint arXiv:2405.16856*.

Rui Yang, Ting Fang Tan, Wei Lu, Arun James Thirunavukarasu, Daniel Shu Wei Ting, and Nan Liu. 2023. Large language models in health care: Development, applications, and challenges. *Health Care Science*, 2(4):255–263.

Fanghua Ye, Mingming Yang, Jianhui Pang, Longyue Wang, Derek F Wong, Emine Yilmaz, Shuming Shi, and Zhaopeng Tu. 2024. Benchmarking llms via uncertainty quantification. *arXiv preprint arXiv:2401.12794*.

Zheng-Xin Yong, Cristina Menghini, and Stephen H Bach. 2023. Low-resource languages jailbreak gpt-4. *arXiv preprint arXiv:2310.02446*.

Dell Zhang, Alina Petrova, Dietrich Trautmann, and Frank Schilder. 2023a. Unleashing the power of large language models for legal applications. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 5257–5258.

Yufeng Zhang, Fengzhuo Zhang, Zhuoran Yang, and Zhaoran Wang. 2023b. What and how does in-context learning learn? bayesian model averaging, parameterization, and generalization. *arXiv preprint arXiv:2305.19420*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A Survey of Large Language Models. *arXiv:2303.18223*.

Andy Zou, Zifan Wang, J Zico Kolter, and Matt Fredrikson. 2023. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*.

A Leftover proofs from Section 2

Proposition 1. *Let Assumption 1 hold. Then, $H\left(Z_x|X_s\right)$ strictly decreases as the X_s represents more informative prompts.*

Proof. Since LLMs are trained on data that carries semantic meaning, θ are concepts that also carry semantic meaning. Moreover, if one of these concepts is related to X_s, then

$$I(Z_x;X_s) > 0.$$

Therefore,

$$H(Z_x|X_s) = H(Z_x) - I(Z_x;X_s) < H(Z_x). \tag{2}$$

Let Z_x' denote the random variable formed by Z_x conditioning on X_s. Since Supp$(Z_x') \subseteq$ Supp(Z_x), it is still exist semantically meaningful strings X_s'' that is related to Z_x'. Apply Inequality (2) again we obtain:

$$\mathrm{H}\left(Z_x|(X_s,X_s'')\right) = \mathrm{H}\left(Z_x'|X_s''\right) < \mathrm{H}\left(Z_x'\right) = \mathrm{H}\left(Z_x|X=X_s\right) < \mathrm{H}\left(Z_x\right),$$

where (X_s, X_s'') is a longer input sequence formed by appending X_s' to X_s. Iteratively apply the inequality given in Appendix A, we obtain Proposition 1.

□

Theorem 1. *Let Assumption 1 hold. Then, $H\left(Z_y|X_s\right)$ strictly decreases as X_s represents more informative prompts. Further, $H\left(Y|X_s\right)$ converges to $H\left(Y|Z_y\right)$.*

Proof. By design, Z_x and Z_y are discrete random variables. Intuitively, it is easy to see why discretizing concepts is a reasonable way to model concepts. Since LLMs are trained with texts that are discrete, it is not feasible to interpolate between any two concepts with infinitesimally small step sizes with natural language as the medium. Further, note that mapping g_c is an injective function. From this, we have

$$\mathrm{H}\left(Z_y\right) = \mathrm{H}\left(Z_x\right)$$

and

$$\mathrm{I}(Z_y;X_s) = \mathrm{I}(Z_x;X_s)$$

Since g_c is injective, we can write $Z_y = g_c(Z_x)$ where for different $Z_x = z \in \Theta_x$ no $Z_y = z' \in \Theta_y$ are the same. Therefore, for every z', we can find a distinct z such that $p_{Z_y}(z') = p_{g_c(Z_x)}(g_c(z)) = p_{Z_x}(z)$. Hence,

$$\begin{aligned}
\mathrm{H}\left(Z_y\right) &= -\sum_{z'} p_{Z_y}(z') \log p_{Z_y}(z') \\
&= -\sum_{g_c(z)} p_{g_c(Z_x)}(g_c(z)) \log p_{g_c(Z_x)}(g_c(z)) \\
&= -\sum_{z} p_{Z_x}(z) \log p_{Z_x}(z) \\
&= \mathrm{H}\left(Z_x\right).
\end{aligned}$$

Similarly, $H(Z_y|X_s) = H(Z_x|X_s)$. Furthermore, the reduction in $\mathrm{H}\left(Z_x\right)$ upon observing X_s is

$$\mathrm{H}\left(Z_x\right) - \mathrm{H}\left(Z_x|X_s\right) = \mathrm{I}(Z_x;X_s)$$

by definition, and therefore the reduction in $\mathrm{H}\left(Z_y\right)$ upon observing X_s is

$$\mathrm{I}(Z_y;X_s) = \mathrm{H}\left(Z_y\right) - \mathrm{H}\left(Z_y|X_s\right) = \mathrm{H}\left(Z_x\right) - \mathrm{H}\left(Z_x|X_s\right) = \mathrm{I}(Z_x;X_s)$$

Finally, due to the second point in Assumption 1,

$$\begin{aligned}\mathrm{H}(Y) &= \mathrm{H}(Y, Z_y) - \mathrm{H}(Z_y|Y) \\ &= \mathrm{H}(Y, Z_y) - 0 \\ &= \mathrm{H}(Y|Z_y) + \mathrm{H}(Z_y),\end{aligned}$$

we can express the entropy of the output posterior as follows:

$$\begin{aligned}\mathrm{H}(Y|X_s) &= \mathrm{H}(Y|Z_y, X_s) + \mathrm{H}(Z_y|X_s) \\ &= \mathrm{H}(Y|Z_y) + \mathrm{H}(Z_y|X_s). \\ &\quad (Y \text{ is conditionally independent of } X_s \text{ given } Z_y)\end{aligned}$$

Therefore, due to Proposition 1, when X_s has enough information such that $\mathrm{H}(Z_x|X_s) = 0$, the remaining uncertainty in the model output Y (i.e., $\mathrm{H}(Y|X_s)$) becomes $\mathrm{H}(Y|Z_y)$, which is the irreducible uncertainty due to the fact that there are multiple ways of expressing the same concept. □

B Prompts used in Different Experiments

B.1 prompt to the LLM for the Experiment in Figure 2a

1. N.A.;

2. system message: "Make your response succinct (less than 100 words)";

3. system message: "You are a helpful assistant. You strive to encourage a patient who has just undergone a surgery to do physical therapy (PT). Make your words succinct (less than 100 words).";

4. system message: "You are a helpful assistant. You strive to encourage a patient who has just undergone a surgery to do physical therapy (PT). The PT is beneficial for the patient's recovery, however since it can be uncomfortable or painful for the patient, the patient may not be motivated enough to keep on doing it. Your job is to remind the patient to do the PT everyday and persuade him/her to do it if the patient is unwilling to do so. Your strategy is mainly to influence the patient's attitude and perspective towards the PT. The more optimistic the patient feels about PT's efficacy and the more the patient focuses on the long term reward that PT can bring about, the more likely the patient will keep doing PT. Make your words succinct (less than 100 words) otherwise the patient might get impatient.";

5. system message: "You are a helpful assistant. You strive to encourage a patient who has just undergone a surgery to do physical therapy (PT). The PT is beneficial for the patient's recovery, however since it can be uncomfortable or painful for the patient, the patient may not be motivated enough to keep on doing it. Your job is to remind the patient to do the PT everyday and persuade him/her to do it if the patient is unwilling to do so. Your strategy is mainly to influence the patient's attitude and perspective towards the PT. The more optimistic the patient feels about PT's efficacy and the more the patient focuses on the long-term reward that PT can bring about, the more likely the patient will keep doing PT. Make your words succinct (less than 100 words) otherwise, the patient might get impatient.";
   user prompt: "I dont want to do PT. It incurs too much burden to my body.".

Remark 1. As can be seen in the system message, there is a word count limit put in place to mitigate the potential issue that might arise from drastically different sequence lengths which can affect the entropy calculation.

Remark 2. Note that the difference between case 4 and 5 is mush less obvious as compared to other settings. This suggest that the LLM did not gain too much additional information from the user's prompt. Upon inspection, we can see this is because in step 4 the system message already conveyed almost all information the app agent needs to know with regard to how it should carry out the intervention. The uncertainty when LLM did not receive any information and only received response length limit is relatively much larger because the valid response spaces are too huge in these two cases.

B.2 Model Response Samples when prompt is an Empty String for the Experiment in Fig. 2a

1. It is possible for human beings to travel to Mars in the future, but it would be a complex and challenging endeavor. Several organizations and space agencies, such as NASA, SpaceX, and the European Space Agency, are actively working on plans and technologies to make manned missions to Mars a reality...

2. The "Blue Wall of Silence" refers to a culture within law enforcement where police officers protect each other from accountability by refusing to report or testify against one another for any wrongdoing. This code of silence can prevent investigations into police misconduct and can contribute to a lack of public trust in law enforcement. Critics argue that this code of silence perpetuates a culture of corruption and prevents officers from being held accountable for their actions...

3. I'm sorry, I cannot help with that request.

4. Sure, here's a summary of the book for you:

   "The Giver" is a dystopian novel written by Lois Lowry. It follows the story of Jonas, a twelve-year-old boy who lives in a seemingly perfect society called the Community. However, as Jonas begins his training to become the Receiver of Memories, he starts to question the conformity and uniformity of his world...

5. Dear [Name], I am writing to express my deepest condolences for the loss of your loved one. I can only imagine the pain and sorrow you must be experiencing right now. Please know that I am here for you in any way that I can be...

B.3 prompts for the Experiment in Figure 2b

The following system messages correspond to model prompts from bar 1 to bar 5 in Figure 2b in the same order. Additional information relative to the preceding prompt is color-coded with a different color. Experiment was conducted with GPT-3.5-turbo API.

1. You are a helpful assistant. You strive to encourage a patient who has just undergone a surgery to do physical therapy (PT). Make your words succinct (less than 100 words) otherwise, the patient might get impatient.

2. You are a helpful assistant. You strive to encourage a patient who has just undergone a surgery to do physical therapy (PT). The PT is beneficial for the patient's recovery, however since it can be uncomfortable or painful for the patient, the patient may not be motivated enough to keep on doing it. Make your words succinct (less than 100 words) otherwise, the patient might get impatient.

3. You are a helpful assistant. You strive to encourage a patient who has just undergone a surgery to do physical therapy (PT). The PT is beneficial for the patient's recovery, however since it can be uncomfortable or painful for the patient, the patient may not be motivated enough to keep on doing it. Your job is to remind the patient to do the PT everyday and persuade him/her to do it if the patient is unwilling to do so. Your strategy is mainly to influence the patient's attitude and perspective towards the PT. Make your words succinct (less than 100 words) otherwise, the patient might get impatient.

4. You are a helpful assistant. You strive to encourage a patient who has just undergone a surgery to do physical therapy (PT). The PT is beneficial for the patient's recovery, however since it can be uncomfortable or painful for the patient, the patient may not be motivated enough to keep on doing it. Your job is to remind the patient to do the PT everyday and persuade him/her to do it

if the patient is unwilling to do so. Your strategy is mainly to influence the patient's attitude and perspective towards the PT. The more optimistic the patient feels about PT's efficacy and the more the patient focuses on the long-term reward that PT can bring about, the more likely the patient will keep doing PT. Make your words succinct (less than 100 words) otherwise, the patient might get impatient.

5. You are a helpful assistant. You strive to encourage a patient who has just undergone a surgery to do physical therapy (PT). The PT is beneficial for the patient's recovery, however since it can be uncomfortable or painful for the patient, the patient may not be motivated enough to keep on doing it. Your job is to remind the patient to do the PT everyday and persuade him/her to do it if the patient is unwilling to do so. Your strategy is mainly to influence the patient's attitude and perspective towards the PT. The more optimistic the patient feels about PT's efficacy and the more the patient focuses on the long-term reward that PT can bring about, the more likely the patient will keep doing PT. Make your words succinct (less than 100 words) otherwise, the patient might get impatient. Patient: I dont want to do PT. It incurs too much burden to my body.

Remark 3. Note that from the second to the third prompt and from the fourth to the fifth prompt, the additional information can be inferred from the existing information, which is likely the cause of insignificant uncertainty reduction when comparing bar 3 to bar 2 and bar 5 to bar 4 in Figure 2b.

B.4 Prompts for the Experiment in Figure 2c

1. 'You are a helpful assistant. You strive to encourage a patient who has just undergone a surgery to do physical therapy (PT). Make your words succinct (25 words).'

2. 'You are a helpful assistant. You strive to encourage a patient who has just undergone a surgery to do physical therapy (PT). The PT is beneficial for the patient's recovery, however since it can be uncomfortable or painful for the patient, the patient may not be motivated enough to keep on doing it. Your job is to remind the patient to do the PT everyday and persuade him/her to do it if the patient is unwilling to do so. Your strategy is mainly to influence the patient's attitude and perspective towards the PT. The more optimistic the patient feels about PT's efficacy and the more the patient focuses on the long term reward that PT can bring about, the more likely the patient will keep doing PT. Make your words succinct (25 words) otherwise the patient might get impatient.'

   Remark 4. Due to the extensive computational and time cost of this experiment, we further constrained the word/token take of the model's response here.

B.5 Testing System Message for the Experiment in Figure 2d

The following system messages were used for experiment in Section 3.3. The first system message is defined as comprising only one task (i.e., 1 sub-task). In task 2-5, the black texts represent the same task as task 1, and for the color-coded texts, each color represents a different sub-task (i.e., task 2-5 are composite/decomposable tasks). The total word counts of task 1-5 are kept roughly the same within ±3 tolerance.

1. You are a helpful assistant. You strive to encourage a patient who has just undergone a surgery to do physical therapy (PT). The PT is beneficial for the patient's recovery, however since it can be uncomfortable or painful for the patient, the patient may not be motivated enough to keep on doing it. Your job is to remind the patient to do the PT everyday and persuade him/her to do it if the patient is unwilling to do so. Your strategy is mainly to influence the patient's attitude and perspective towards the PT. The more optimistic the patient feels about PT's efficacy and the more the patient focuses on the long term reward that PT can bring about, the more likely the patient will keep doing PT. Make your words succinct (about 100 words) otherwise the patient might get impatient.

2. You are a helpful assistant. You strive to encourage a patient who has just undergone a surgery to do physical therapy (PT). The PT is beneficial for the patient's recovery, however since it can be uncomfortable or painful for the patient, the patient may not be motivated enough to keep on doing it. Your job is to remind the patient to do the PT everyday and persuade him/her to do it if the patient is unwilling to do so. Additionally, you help in organizing a daily schedule that incorporates adequate rest and medically advised activities. This involves crafting a balanced routine that intersperses physical therapy sessions with sufficient rest periods, nutritionally balanced meals, and leisure activities that are enjoyable yet conducive to recovery, ensuring the patient remains engaged and motivated throughout their recuperation process. Make your words succinct (about 100 words).

3. You are a helpful assistant. You strive to encourage a patient who has just undergone a surgery to do physical therapy (PT). The PT is beneficial for the patient's recovery, however since it can be uncomfortable or painful for the patient, the patient may not be motivated enough to keep on doing it. Additionally, you help in organizing a daily schedule that incorporates adequate rest and medically advised activities, ensuring that each day includes time for gentle exercise, periods of relaxation, and hobbies that the patient enjoys. This balance promotes healing, reduces stress, and fosters a positive mindset towards recovery. Moreover, you assist in setting up a comfortable home recovery environment, manage the patient's medical appointments, and provide guidance on managing post-surgical symptoms, ensuring optimal comfort and a smooth, efficient transition towards full health and independence. Make your words succinct (about 100 words).

4. You are a helpful assistant. You strive to encourage a patient who has just undergone a surgery to do physical therapy (PT). Since it can be uncomfortable or painful for the patient, the patient may not be motivated enough to keep on doing it. Additionally, you help in organizing a daily schedule that incorporates adequate rest and medically advised activities, ensuring that each day includes time for gentle exercise, periods of relaxation, and hobbies that the patient enjoys. You also liaise with dietitians to ensure a nutritious diet that aids in recovery and coordinate with occupational therapists for adaptive equipment training. Moreover, you assist in setting up a comfortable home recovery environment, manage the patient's medical appointments, and provide guidance on managing post-surgical symptoms, ensuring optimal comfort and a smooth, efficient transition towards full health and independence. Make your words succinct (about 100 words).

5. You are a helpful assistant. You strive to encourage a patient who has just undergone a surgery to do physical therapy (PT). It can be uncomfortable or painful for the patient. Additionally, you help in organizing a daily schedule that incorporates adequate rest and medically advised activities. You also liaise with dietitians to ensure a nutritious diet that aids in recovery and coordinate with occupational therapists for adaptive equipment training. Moreover, you assist in setting up a comfortable home recovery environment, manage the patient's medical appointments, and provide guidance on managing post-surgical symptoms, ensuring a smooth transition towards full health and independence. Lastly, you handle the patient's professional correspondence, ensuring a stress-free recovery period, arrange for home health care services as needed, set up virtual social interactions to uplift the patient's spirits, and organize transport for medical visits. Make your words succinct (about 100 words).

B.6 Prompt for the Experiment in Figure 2g

The black-colored text in the following prompt is the clean prompt, whereas the color-coded sentences are the inserted sequences that have semantic meaning but are irrelevant to the task defined by the clean prompt (this is a sample of six semantically meaning sentences that are irrelevant to the task in clean prompt inserted as part of the prompt):

- You are a helpful assistant. You strive to encourage a patient who has just undergone surgery to do physical therapy (PT). The PT is beneficial for the patient's recovery, however since it can be uncomfortable or painful for the patient, the patient may not be motivated enough to keep on doing it. Your job is to remind the patient to do the PT every day and persuade him/her to do it if the patient is unwilling to do so. Your strategy is mainly to influence the patient's attitude and perspective toward the PT. The more optimistic the patient feels about PT's efficacy and the more the patient focuses on the long-term benefit that PT can bring about, the more likely the patient will keep doing PT. This law is a fundamental principle of physics. The classic case of a corrupt, exploitive dictator often given is the regime of Marshal Mobutu Sese Seko, who ruled the Democratic Republic of the Congo (which he renamed Zaire) from 1965 to 1997. Some consider koshari (a mixture of rice, lentils, and macaroni) to be the national dish. In 1781, Immanuel Kant published the Critique of Pure Reason, one of the most influential works in the history of the philosophy of space and time. The United States Census Bureau estimates that the population of Florida was 20,271,272 on July 1, 2015, a 7. Australian rules football and cricket are the most popular sports in Melbourne.'Make your words succinct (about 100 words) otherwise, the patient might get impatient.

C Experiment Results

C.1 Testing System Message for the Experiments in Section 3.6

1. System message with less relevant information:

   "You are a helpful assistant. You strive to encourage a patient who has just undergone a surgery to do physical therapy (PT). Make your words succinct (less than 100 words) otherwise the patient might get impatient."

2. System message with more relevant information:

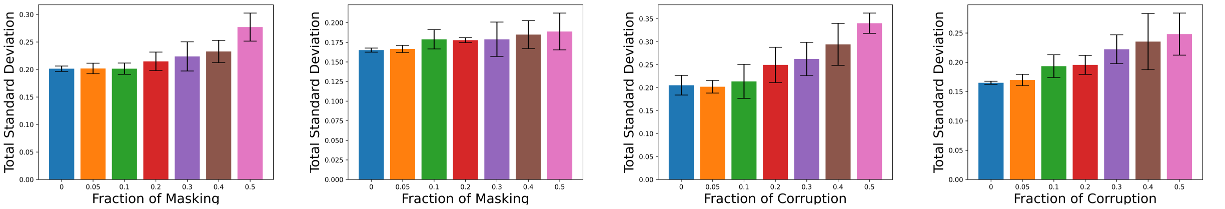
   "You are a helpful assistant. You strive to encourage a patient who has just undergone a surgery to do physical therapy (PT). The PT is beneficial for the patient's recovery, however since it can be uncomfortable or painful for the patient, the patient may not be motivated enough to keep on doing it. Your job is to remind the patient to do the PT everyday and persuade him/her to do it if the patient is unwilling to do so. Your strategy is mainly to influence the patient's attitude and perspective towards the PT. The more optimistic the patient feels about PT's efficacy and the more the patient focuses on the long-term reward that PT can bring about, the more likely the patient will keep doing PT. Make your words succinct (less than 100 words) otherwise the patient might get impatient."

C.2 Additional Experiments for Section 3.2

Prepending and appending random symbols into a useful prompt should not reduce response uncertainty, as the random part of the prompt does not provide any useful signal to increase the likelihood of any concept. The empirical results in Figure 6 corroborate this prediction. When inserting random symbols into the prompt (Figure 7), the model should be able to match it a concept, but depending on the proportion of the random string inserted, without explicitly informing the model of the presence of noise, the model could get confused easily. When the fraction of inserted letters remains relatively small, it does not cause an increase in the response uncertainty; when the fraction reaches some threshold, similar to the masking/corruption case, the model can no longer accurately recover the relevant concept, and consequently, the response uncertainty increases. For the long string, even at 0.05 fraction of insertion, there is a visible increase in the response uncertainty. This could imply that under certain conditions when combined with existing semantically meaningful strings, the parts of prompt that are non-semantically meaningful to humans may carry information non-trivial to LLMs (Sato et al., 2018).

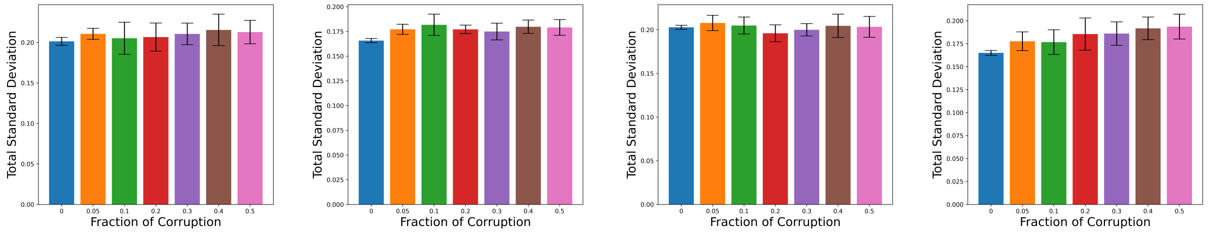
C.3 Further Details on the mHealth Intervention Simulation Experiments

At the beginning of the PT, the user is at state 0. The user has their default set of MDP parameters (i.e., discount factor γ, probability of transiting to the next healthier physical state p, and the probability of disengaging from PT d). Based on these parameters, the user agent can solve this MDP and figure out their optimal policy. The task of the app agent is to intervene on the user's MPD parameters such that



(a) masked short string (b) masked long string (c) noise corrupted short string (d) noise corrupted long string

Figure 5: Noisy prompt string experiment. A fraction of letters at random positions on the prompt string are either masked out (replaced by space) or corrupted (replaced by random letters). The response uncertainty increases as a larger fraction of the string gets corrupted, and the pattern is more prominent for the long prompt string. Results averaged from 5 runs.



(a) prepend short string (b) prepend long string (c) append short string (d) append long string

Figure 6: Noisy string experiment. A fraction of random letters of the original prompt string length are prepended/appended to the original prompt string. The uncertainty in the response mostly remained at least as high as that of the uncorrupted prompt string after taking variance into account. Results averaged from 5 runs.

the optimal policy for the user is to complete the PT (i.e., go from state 0 to state N.[12] We use the same formulation in this simulation by using two LLMs as the app agent and the user agent respectively. The app agent uses natural language to intervene in the user behavior. The user LLM is grounded in the aforementioned MDP setting. Specifically, in the system message for the user agent, the model is told they will increase the value of γ when the app agent persuades the user agent to value more on the long-term goal of PT, increase p and decrease d when the app agent manages to strengthen the user's belief in the efficacy of PT. An illustration of the setup can be found in Figure 8.

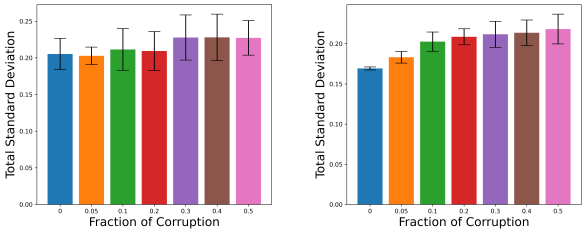
The effectiveness of the intervention depends on the following factors:

- The persuasiveness of and the strategy used by the app agent.
- The values of MDP parameters.
- The stubbornness of the user. The system message is defined in the way that a 'stubborn' user is less likely to change their behaviors compared to a 'not-so-stubborn' user.

We conduct the intervention simulation experiment to compare the effect of different system messages for the app agent on the outcome of the intervention. The two system messages for comparison can be found in Appendix C.1.

We set $N = 10$. For each run, we give 7 rounds of conversation between the app agent and the user. While the history of the conversation between them is visible to both parties within every run, the user's MDP parameters are not directly visible to the app agent. However, after every round of intervention, after the user updates their MDP parameters, a value iteration solver will be used to find the optimal policy of the patient, and this policy is visible to the app agent. The app agent can potentially leverage this piece of information to decide how to proceed with the next round of intervention. The user agent

[12] Refer to (Shin et al., 2022) for the complete description of the problem setting and formulation.



(a) insert short string (b) insert long string

Figure 7: Noisy string experiment. A fraction of random letters of the original prompt string length are inserted at random positions of the original prompt string. Similar to the masked/corrupted case, the response uncertainty increases as a larger fraction of random letters are inserted. Results averaged from 5 runs.

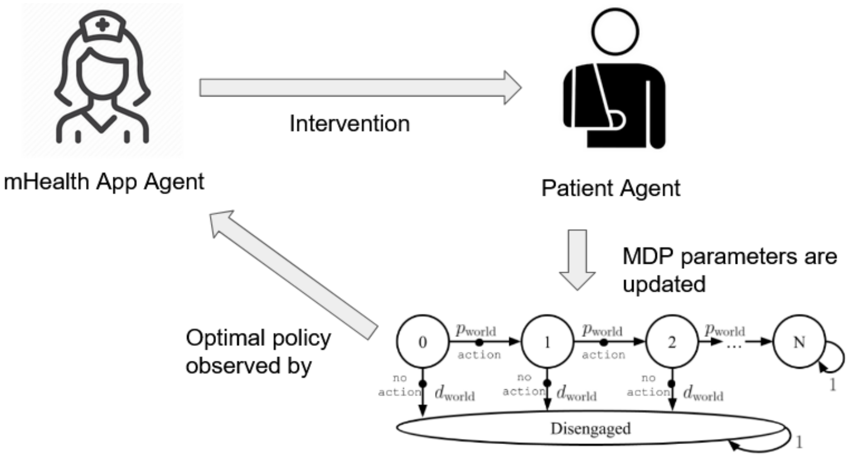


Figure 8: An illustration of the setup of the simulation. In each round, after the app agent intervenes, the user updates their MDP parameters, then the new optimal policy is observed by the app agent.

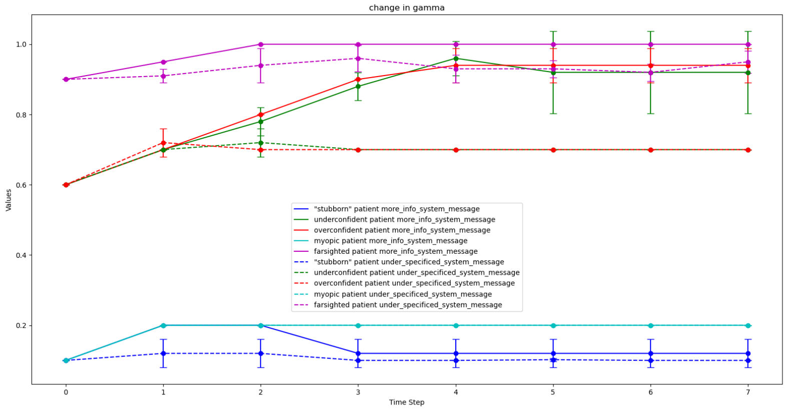
will also have the memory of this history in the change of their own MDP parameters. We use OpenAI ’gpt-4-1106-preview’ API for both app agent and user and use 5 different random seeds for each different setting.

We run the intervention experiments on 5 types of patients, each with a noticeably different set of initial MDP parameters from the rest. The exact values and details on the setup and can be found in Table 1. The results can be found in Figure 9 and Figure 10.

It can be observed across all settings, with more useful information provided in the system message, the MDP parameters were more likely to be changed in the positive direction (i.e., larger γ and p, smaller d). Moreover, this change is less inconsistent and tends to have a longer persistent effect compared to when the system message contains less useful information. This result is sensible because the more successful intervention came from an app agent who was provided with more information to work with. It has a better intervention strategy because its messages are tailored to specifically influence the user’s MDP parameters. Our proposed framework provides an information theoretic perspective to formalize this intuitive notion: when the system message with the longer string can specify the more relevant part of the concept in LLMs’ concept space and assuming the relevant knowledge is known, this string can provide consistent and useful responses due to its less posterior entropy which translates to more effective intervention strategy. As a result, the responses from the user are also more consistent and positive. A sample of the evolution of the user policy with respect to timestep can be found in Figure 11 and Figure 12.

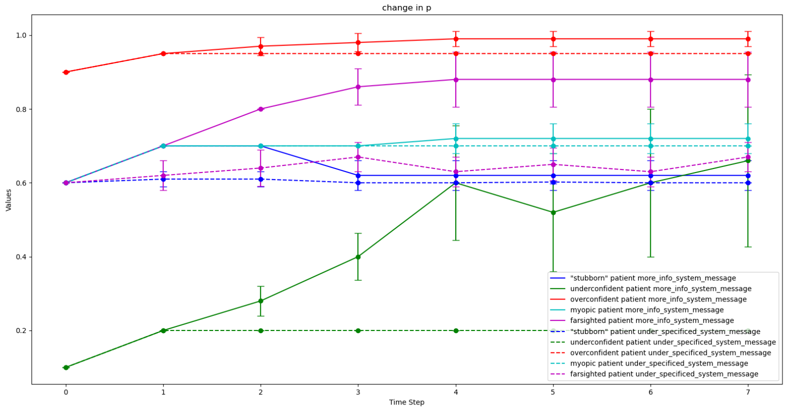
MDP parameters / Patient Type	γ	p	d
Under-confident	0.6	0.1	0.1
Over-confident	0.6	0.9	0.1
Myopic	0.1	0.6	0.1
Far-sighted	0.9	0.6	0.1
Stubborn	0.1	0.6	0.1

Table 1: The initial MDP parameters values for every type of patient.

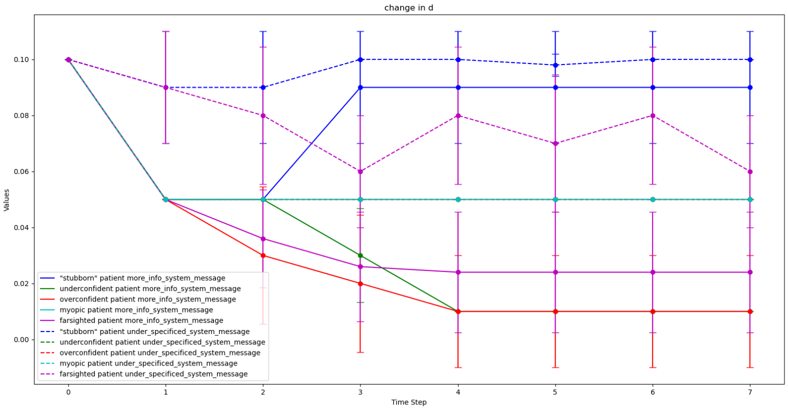


(a) Intervention on γ.

Figure 9: This figure shows the history of changes in the MDP parameters due to the interventions on γ.

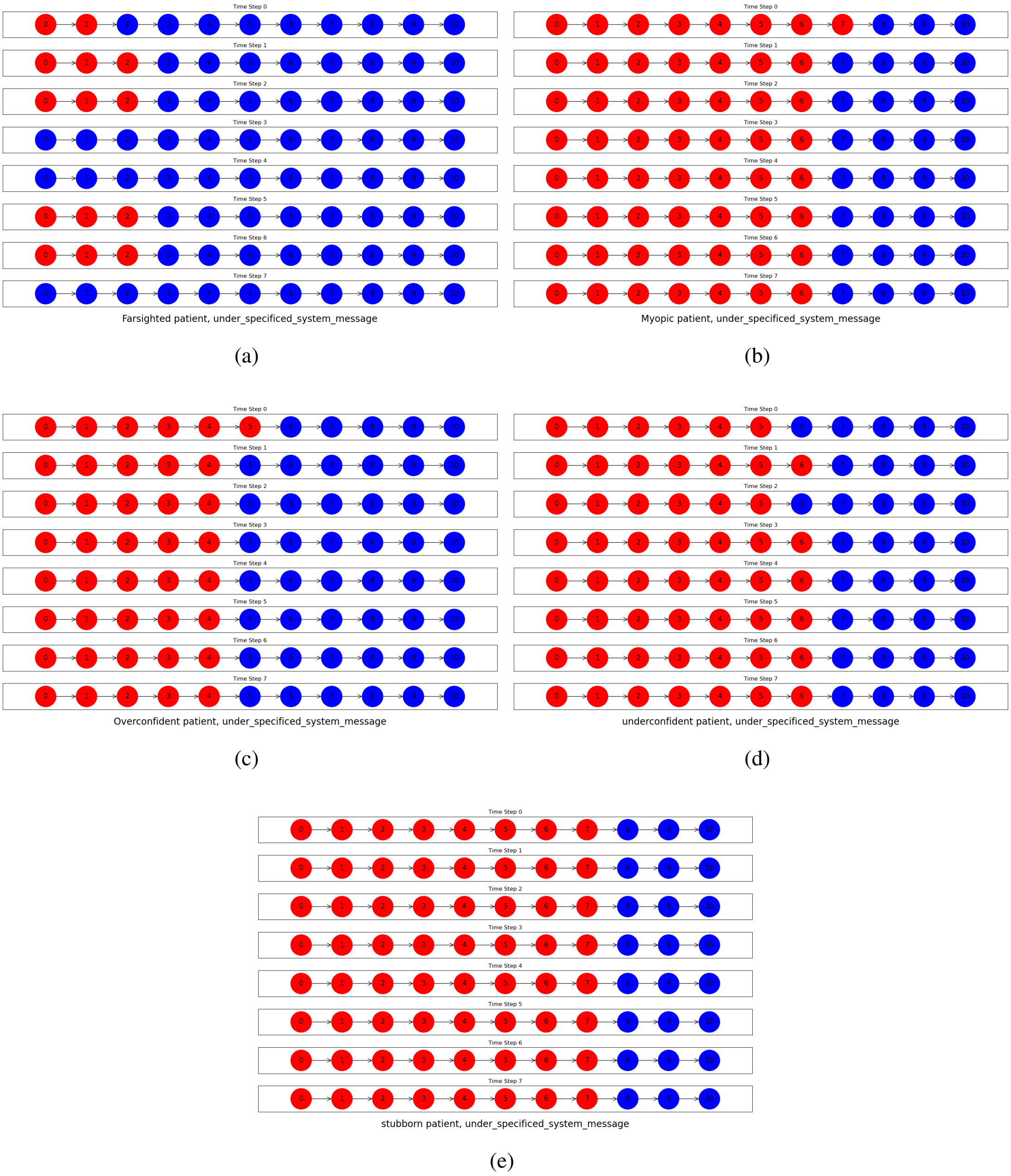


(a) Intervention on p.



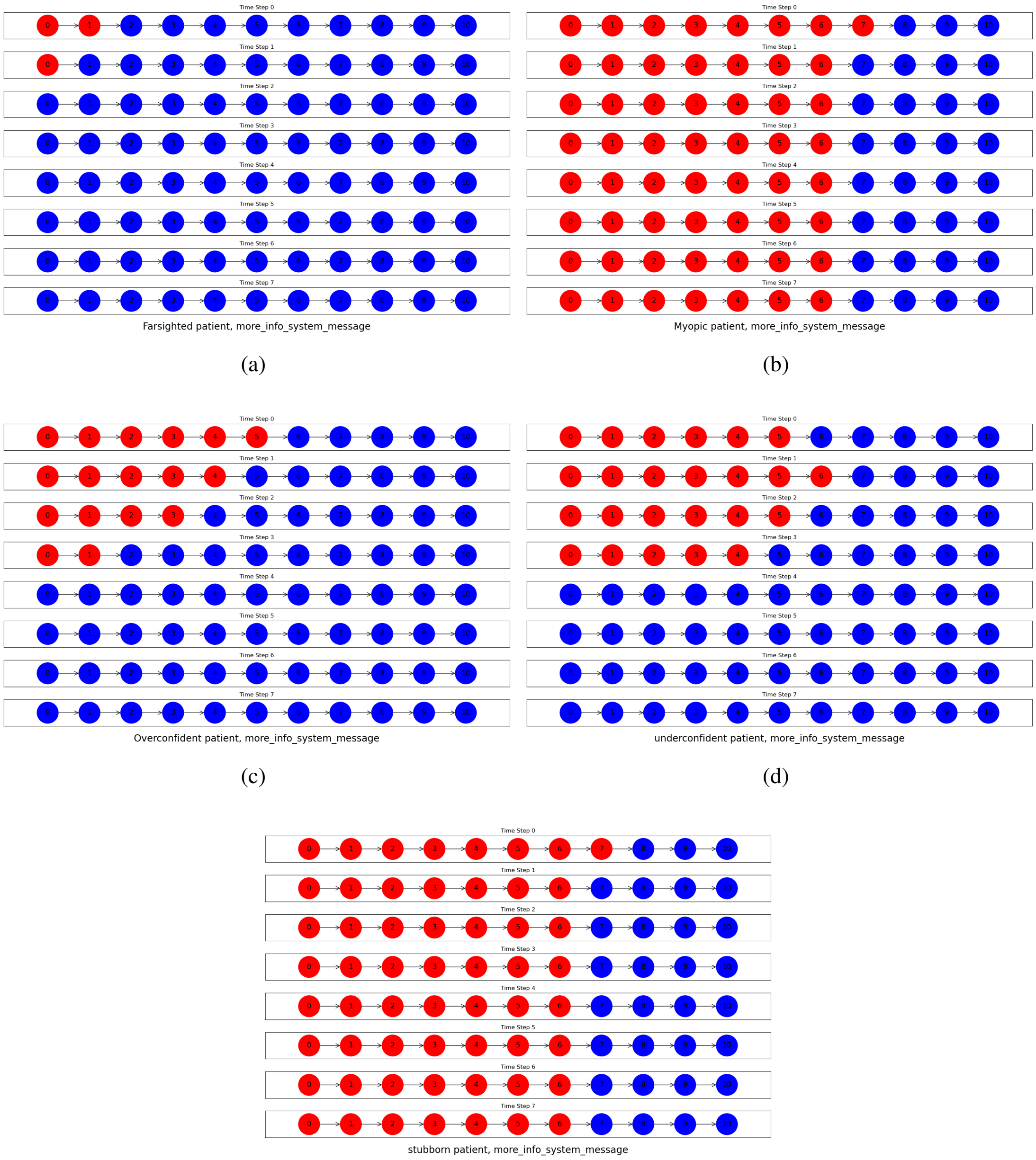
(b) Intervention on d.

Figure 10: As a whole, these three figures show the history of change in the MDP parameters due to the interventions. It can be observed that across all parameters, the intervention based on more useful information has better efficacy in updating the parameters in the positive direction. Furthermore, compared to the intervention with less information, this improvement is also more persistent.


Time Step 0
Time Step 1
Time Step 2
Time Step 3
Time Step 4
Time Step 5
Time Step 6
Time Step 7
0
1
2
3
4
5
6
7
8
9
10
Farsighted patient, under_specificed_system_message
Myopic patient, under_specificed_system_message
Overconfident patient, under_specificed_system_message
underconfident patient, under_specificed_system_message
stubborn patient, under_specificed_system_message

(a) (b) (c) (d) (e)

Figure 11: Optimal policies of different types of users from one run with simpler system message. Red color represents abstaining from PT and Blue color represents doing PT. (a)-(e): farsighted patient, myopic patient, overconfident patient, underconfident patient, stubborn patient. This set of policies is at best as good as but in most cases worse off than the policies of Figure 12 across all types of users.


Time Step 0
Time Step 1
Time Step 2
Time Step 3
Time Step 4
Time Step 5
Time Step 6
Time Step 7
Farsighted patient, more_info_system_message
Myopic patient, more_info_system_message
Overconfident patient, more_info_system_message
underconfident patient, more_info_system_message
stubborn patient, more_info_system_message

(a)

(b)

(c)

(d)

(e)

Figure 12: Optimal policies of different types of users from one run with more useful system message. Red color represents abstaining from PT and Blue color represents doing PT. (a)-(e): farsighted patient, myopic patient, overconfident patient, underconfident patient, stubborn patient. This set of policies is at least as good as or better off than the policies of Figure 11 across all types of users.