Generating Mental Health Transcripts with SAPE (Spanish Adaptive Prompt Engineering)

Daniel Cabrera Lozoya[1], Alejandro Casar Berazaluce[1],
Juan Alberto Barajas Perches[2], Eloy Hernández Lúa[2],
Mike Conway[1], and Simon D'Alfonso[1]
[1]The University of Melbourne, Australia
[2]ITESM, Mexico
{dcabreralozo, acasarberaza}@student.unimelb.edu.au
{jbperches, eloyhl}@exatec.tec.mx
{mike.conway, dalfonso}@unimelb.edu.au

Abstract

Large language models have become valuable tools for data augmentation in scenarios with limited data availability, as they can generate synthetic data resembling real-world data. However, their generative performance depends on the quality of the prompt used to instruct the model. Prompt engineering that relies on hand-crafted strategies or requires domain experts to adjust the prompt often yields suboptimal results. In this paper we present SAPE, a Spanish Adaptive Prompt Engineering method utilizing genetic algorithms for prompt generation and selection. Our evaluation of SAPE focuses on a generative task that involves the creation of Spanish therapy transcripts, a type of data that is challenging to collect due to the fact that it typically includes protected health information. Through human evaluations conducted by mental health professionals, our results show that SAPE produces Spanish counselling transcripts that more closely resemble authentic therapy transcripts compared to other prompt engineering techniques that are based on Reflexion and Chain-of-Thought.

1 Introduction

Large Language Models (LLMs) have achieved state-of-the-art (SOTA) performances on multiple Natural Language Processing (NLP) benchmarks (Chowdhery et al., 2023). They have the potential to be adapted as data augmentation tools in scenarios where data is hard to collect (Amin-Nejad et al., 2020). However, SOTA LLMs such as GPT-4 and Bard are only accessible as services, withholding access to the model's underlying architecture and parameters for commercial reasons (Sun et al., 2022). Consequently, fine-tuning these models for specific downstream NLP tasks becomes infeasible. As an alternative to enhance LLMs without the need to retrain them, prompt engineering has proven to be an effective approach (Meskó, 2023). Different prompt techniques have shown a significant impact on a model's arithmetic problem-solving capabilities (Wei et al., 2022b), enhance the model's ability to accurately mimic human writing styles (Chen et al., 2023), or improve the model's commonsense reasoning skills (Zelikman et al., 2022). However, these prompt engineering techniques require manual crafting. Given that the precise phrasing of a prompt can significantly influence its effectiveness (Zhou et al., 2022), studies have been directed towards automating prompt engineering (Fernando et al., 2023). It is noteworthy that the exploration of prompt engineering has predominantly centered on the English language, leaving other languages largely unexplored in this domain. Therefore, this paper introduces SAPE, a Spanish Adaptive Prompt Engineering method that employs genetic algorithms for prompt generation and selection.

To evaluate the generative capabilities of LLMs using SAPE's prompts, we employed OpenAI's `text-davinci-003` model to create synthetic therapy transcripts. Psychologists compared SAPE-generated transcripts with those produced using prompts based on Chain-of-Thought (CoT) (Wei et al., 2022b) and Reflexion (Shinn et al., 2023) methods. Our results indicate that overall mental health professionals find SAPE-generated text to resemble authentic therapy transcripts more closely than texts generated with other prompt engineering techniques. We chose to focus on generating synthetic psychotherapy data since obtaining authentically generated therapy transcripts is challenging due to privacy concerns and the need to protect sensitive health information (Lu et al., 2021).

The development of high-quality synthetic ther-

apy transcripts holds great potential for advancing NLP models in mental health. These transcripts can train text classifiers to automate post-therapy session reports, thus streamlining administrative tasks for psychologists. Furthermore, they can be utilized to train chatbots that simulate mental health patients, which could provide (trainee) psychologists or counsellors with a valuable tool for training and practicing their skills, without the need for other interlocutors. In addition to the applications of NLP, collections of psychotherapy transcripts offer a valuable resource for study in psychology and counselling. For instance, the *Counseling and Psychotherapy Transcripts: Volume I* collection from Alexander Street Press (Alexander Street Press, 2023) provides a substantial set of therapy transcripts suitable for teaching and research purposes. However, these transcripts are predominantly in English, lacking equivalents in other languages. The SAPE tool, and similar adaptations in other languages, presents an opportunity to create sets of therapy transcripts in non-English languages where authentic collections are scarce.

Our main contributions are as follows:

1. Development of a Spanish database comprising therapy sessions.

2. Creation, evaluation, and release of SAPE, a Spanish Adaptive Prompt Engineering technique that employs genetic algorithms. Our evaluation, which was conducted by mental health professionals, entailed a comparison of SAPE with CoT and Reflexion within a Spanish context.

3. Public release of the synthetic datasets generated and employed in this study, aiming to facilitate future research in the field of mental health.

2 Related Work

2.1 Prompt Engineering

The proficiency of LLMs in downstream tasks relies on the quality of the prompt utilized for instructing the model (Grabb, 2023). Techniques such as CoT and Tree of Thoughts (Yao et al., 2023) augments the prompts with intermediate steps to enhance LLMs' mathematical and commonsense reasoning capabilities. Self-Consistency, an extension of CoT, replaces the naïve greedy strategy

used in CoT by first sampling a diverse set of intermediate steps instead of only taking the greedy one, and then selecting the most consistent answer (Wang et al., 2022). Prompt-tuning approaches based on gradient-based techniques have proven effective in enhancing the performance of LLMs (Liu et al., 2023; Qin and Eisner, 2021; Lester et al., 2021). However, their practicality diminishes at scale, as computing gradients becomes more resource-intensive with the increasing size of LLMs (Zhou et al., 2022). Furthermore, these approaches become infeasible for LLMs that are concealed behind APIs that do not offer gradient access (Zhou et al., 2022). The previous prompt techniques are manually crafted, potentially restricting their exploration within the natural language hypothesis space.

Automation strategies for prompt engineering have exhibited promising outcomes. The Automated Prompt Engineering (APE) (Zhou et al., 2022) achieves human-level performance on zero-shot learning with model generated instructions on 24/24 Instruction Induction and 17/21 Big-Bench tasks. APE employs one generator-prompt to create task-prompts candidates and another mutator-prompt to introduce variations. Similarly, Optimization by PROmpting (OPRO) induces prompt variations by using a single complex mutation prompt and assesses newly generated prompts on a training dataset. As shown in (Yang et al., 2023) prompt diversity boosts the performance of an automated prompt strategy. In our work, we followed the approach used in Promptbreeder (Fernando et al., 2023) which evolves both the mutator-prompts and the tasks-prompts to address issues of diversity loss. The approach employed by Promptbreeder relies on a binary tournament genetic algorithm framework (Harvey, 2009). This involves sampling two prompts that came from different prompt-tasks, selecting the one with the higher fitness, mutating it, and then replacing the less fit prompt with the mutated version of the winner. Since their genetic algorithm adopts a greedy strategy, the heuristic of consistently choosing the prompt with the highest fitness at each stage carries the risk of getting trapped in a local maximum. To prevent convergence to a local maximum, our genetic algorithm employs a strategy that extends beyond selecting prompts solely based on their highest fitness. Instead, it normalizes the collective fitness of all the prompts. The algorithm then de-

termines the winner through a probability function, utilizing values derived from this normalization process. This approach ensures a more balanced exploration of the solution space, avoiding premature fixation on a local optimum.

To the best of our knowledge, SAPE is the first adaptive prompt optimization for open-ended tasks. Previous work such as Promptbreeder, Automatic Prompt Engineering, and Optimization by PROmpting; focused on prompt optimization for close-ended questions, such as arithmetic tasks with a single correct answer (i.e. the grade school maths word problem dataset GSM8K). Given that their tasks had a correct answer, the fitness function they used to optimize their prompts were based on the accuracy or number of correct answers the prompt generated. In contrast, our model is designed to optimize prompts for generating therapy transcripts. Unlike arithmetic problems with clear correct answers, therapy transcripts are more open ended and do not have a single correct way to be written. Hence, to optimize our prompts our fitness function relied on reinforcement learning with human feedback facilitated by domain experts – in this case, clinical psychologists.

2.2 Data Augmentation for Mental Health Datasets

Artificial intelligence tools in mental health have facilitated the automation of therapists tasks, leading to improvements in clinical capabilities and enhanced access to care (Minerva and Giubilini, 2023). A significant challenge in training AI models lies in data collection due to clinical constraints on data accessibility and patient privacy concerns (Zhang et al., 2022). Leveraging synthetic data to address both data sparsity and privacy concerns has emerged as a promising solution (Ive, 2022; Ansari et al., 2021).

The creation of synthetic patient datasets has predominantly focused on generating Electronic Health Records (EHRs) (Gulrajani et al., 2017; Hjelm et al., 2018; Kennedy et al., 2022). One of the first generative architectures used for augmenting EHR data, MedGan (Choi et al., 2017), introduced a generative adversarial network (GAN) designed to create multi-label patient records. Relevant work extends to the creation of synthetic mental health records, as seen in Ive et al. (2020), where discharge summaries from mental health providers were artificially generated. One limitation of synthetic records is the frequent absence of unstructured text sections, and when present, such sections are typically brief. As an example, Lee (2018) reported on an approach to generating unstructured text that is limited to 18 tokens or less. In contrast to health records, therapy transcripts capture additional nuances through unstructured text data, allowing a deeper understanding of a patient. The richer data source provides valuable insights into various aspects, including cognitive patterns, interpersonal dynamics, and patient's goals and aspirations.

LLMs have been used to create synthetic therapy transcripts. For example, Stapleton et al. (2023) utilized ChatGPT as a patient chatbot, simulating an individual experiencing suicidal ideation. The chatbot's prompt was manually crafted and refined to emulate the writing style found in online platforms where individuals have described their own feelings of suicidality. In related work, Chen et al. (2023), employed chatbots powered by ChatGPT to replicate counseling sessions. To design their prompts, they followed a manual iterative methodology based on user feedback. Relying on manual approaches to design prompts limits the exploration of the language space. In our work we employed a genetic algorithm to automate the language space search to discover a more suitable prompt for generating therapy transcripts.

Our overarching aim with this work is to contribute to the advancement of research on mental health synthetic data creation. Notably, there remains a scarcity of research in this domain, particularly concerning synthetic text generated in languages other than English. To the best of our knowledge, this paper describes the first attempt at creating a Spanish Adaptive Prompt Engineering mechanism evaluated in the context of mental health. We hope that our methodology contributes to the broader understanding of synthetic data generation for therapy transcripts and encourage further exploration in non-English languages.

3 Method

3.1 Cognitive Behavioral Therapy

Cognitive Behavioral Therapy (CBT) is a type of psychotherapy grounded in the belief that mental disorders consist of cognitive and behavioral factors (Beck, 1970). It is a goal-oriented, evidence-based intervention designed to facilitate improvements in patients' symptoms by modifying these

contributing factors. CBT focuses on helping individuals identify and alter patterns of thought and behavior associated with their emotional and psychological difficulties (David et al., 2018; Fenn and Byrne, 2013). Informed by the CBT competences framework (Roth and Pilling, 2008) and the Revised Cognitive Therapy Scale (Blackburn et al., 2001), Ewbank et al. (2020) categorized therapist utterances into 24 distinct categories based on their respective roles within a therapy session. In this study we aimed to create synthetic therapy transcripts that capture three types of therapeutic interactions:

- **Mood check**: Assessing the patient's mood.

- **Change methods**: Cognitive reattribution, behavioral reattribution, skill-teaching, conceptualization, or psychoeducation employed by the therapist to promote therapeutic change.

- **Set goals**: Setting patients long-term goals for therapy.

3.2 Data Collection

We compiled 30 hours of Spanish counselling session videos sourced from publicly available content on YouTube. We then extracted utterances falling into one of the three distinct categories: mood check, change methods, and set goals. The extraction and annotation of the data was done by a psychologist and one of the authors, who has completed courses on CBT. While transcribing the dialogues, the annotators excluded any protected health information from the final text.

3.3 Synthetic data for Reinforcement Learning with Human Feedback

We utilized the `text-davinci-003` model of OpenAI's GPT3 system to construct a synthetic dataset employed in training a reinforcement learning from human feedback (RLHF) model (von Werra et al., 2020). For each of the three therapeutic interaction types, we randomly sampled organic examples from the Spanish counseling session videos. The sampling process involved selecting interactions from the database until we accumulated 2,400 tokens. Subsequently, we appended an instructional prompt to infer 10 prompts that could have been employed in generating the organic examples. Using each of the inferred prompts, we generated 1,000 synthetic therapy transcripts for a total of 30,000 transcripts. We used the following

configuration to generate the transcripts: a temperature of 0.9, a presence penalty of 0.6, a nucleus sampling of 1.0, and a response length of 1,800. Within each of the three types of interactions, transcripts were paired together, resulting in a total set of 15,000 pairs, and the annotators were tasked with indicating a preference between the two examples in each pair. After labeling the dataset a reward model was trained using the Transformer Reinforcement Learning (TRL) (von Werra et al., 2020) library which is licensed under an Apache License 2.0. The following configuration was used to train the model: an Adam with decoupled weight decay (AdamW) as the optimizer, a learning rate of 1.41×10^{-5}, and training over 300 steps. The computing infrastructure employed for training this model was an NVIDIA A100 GPU.

3.4 Genetic Algorithm

Let T denote the text generated by an LLM when provided with a base prompt B as input, denoted as $T = \text{LLM}(B)$. SAPE seeks an optimal instruction-prompt P to concatenate with B, aiming to maximize the quality of T compared to the case where B is presented alone.

Similar to Promptbreeder, SAPE creates instruction-prompts using a genetic algorithm (Lambora et al., 2019). The mutations in SAPE involve a mutation prompt M and an LLM. An evolved prompt P' is defined as $P' = \text{LLM}(M + P)$, where + represents string concatenation. The pool of mutation-prompts are detailed in section 3.5. Mutation-prompts are also evolved in SAPE through the implementation of hypermutations (Ouertani et al., 2019). To do so a hyper-mutation prompt H and an LLM are used. An evolved mutation-prompt M' is expressed as $M' = \text{LLM}(H + M)$.

Given an initial set of organic therapeutic interactions, SAPE first employs an infer-prompt to deduce a base prompt. Subsequently, it modifies the prompt through evolution using a random mutation prompt. A population of text is then created using both the base prompt and the evolved prompt. SAPE maintains a record of the mutation prompt, the instruction prompt, the resulting generated text from the instruction prompt, and the associated fitness level that the text achieves based on the reward model. Each record represents an individual in the population.

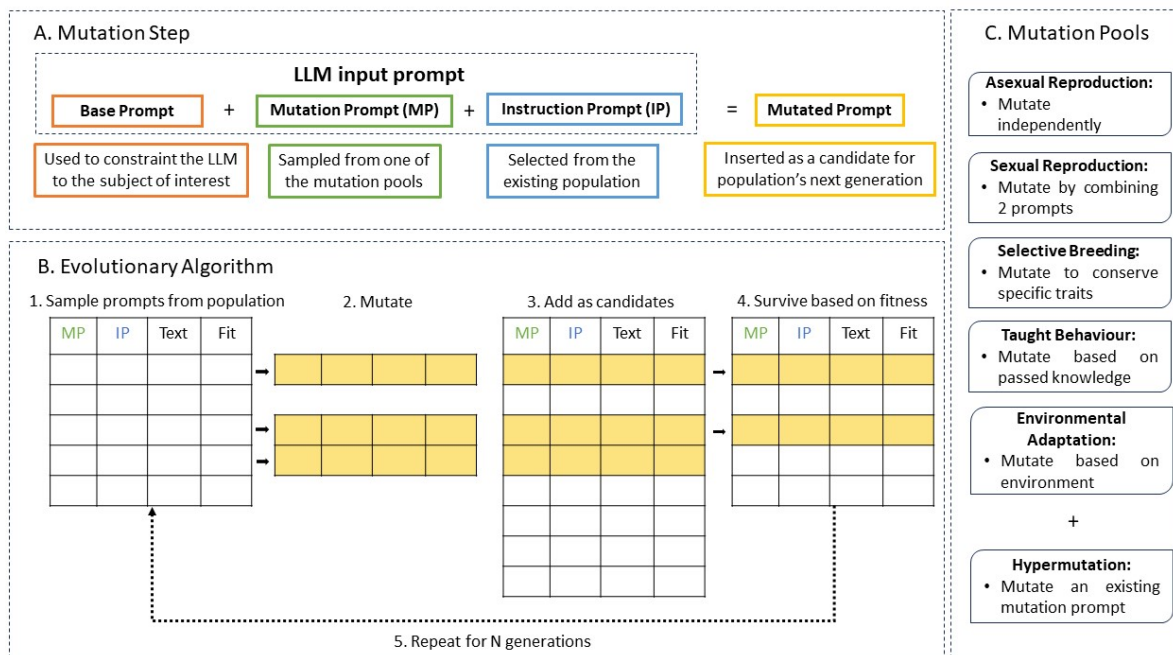After the population is initialized, evolution is

Figure 1: In the SAPE Evolutionary Algorithm, a population consists of individuals, each comprising a mutation prompt used to create its instruction prompt and the resulting text generated through that instruction. During each step of evolution, each individual has a probability of acquiring a mutation that modifies its instruction prompt. The specific type of mutation is chosen from a mutation pool. The individuals that undergo mutation are then integrated into the population. When the maximum population cap is reached, a fitness-based probabilistic selection is employed to decide which individuals advance to the next generation. The individual's fitness level is determined by the text's performance according to the reward model.

implemented as a generational process. At each generation, every individual has mutation probability μ_m of acquiring a mutation that modifies its instruction prompt. After determining which individuals would acquire a mutation, SAPE determines the category of mutation to be acquired out of 5 possible options.

To achieve the balance of breadth and depth (Moreno-Bote et al., 2020) required for a healthy evolutionary search process, every mechanism of mutation has an equal base probability of being the acquired mutation, but mutation categories that have track record of yielding good fitnesses have an increased chance of being acquired on top of the base chance.

The mutated individuals are then introduced into the population of that generation. This process is repeated every generation until the maximum population cap is reached. For every generation after this point is reached, a fitness-based probabilistic selection of the fittest is employed to select the individuals that will carry on to the next generation and the ones that will die off. To avoid the risk of falling into a local maxima, SAPE samples the

surviving individuals using a probability based on their fitness (Marsili Libelli and Alba, 2000). The fittest individuals have the higher chance of surviving, but under-performing individuals that could be key elements for finding the global maxima still have a chance of surviving and contributing to future generations.

After *N* generations, the surviving population is used to determine SAPE's output prompt. Figure 1 provides an overview of our method.

3.5 Mechanisms of mutation

The pool of possible mutation categories are all inspired by biological evolutionary processes including sexual reproduction, asexual reproduction, selective breeding, environmental adaptation and taught behaviour. The starting prompts for each category can be found in Appendix A.

3.5.1 Sexual Reproduction

In an evolutionary genetic context, sexual reproduction is the act of combining genetic information from two individuals to generate a new one (Sivanandam et al., 2008). SAPE implements this by selecting a partner to reproduce with the mu-

tating individual. The partner is sampled from the population based on its fitness similarity to the mutating individual. Then, a randomly selected mutation prompt from the sexual reproduction prompt pool is applied using both individuals' instruction prompt, and a child prompt with characteristics from both parent prompts is generated.

3.5.2 Asexual Reproduction

Asexual reproduction is when an individual reproduces independently without the need of a partner. SAPE implements this by applying a mutation prompt to the selected individual's instruction prompt, creating a new mutated individual (De Falco et al., 2002).

3.5.3 Selective Breeding

Selective breeding is the act of taking deliberate actions to guarantee specific chosen traits are maintained throughout the generations (Sriramya et al., 2013). The trait to be maintained in this context is the highest possible fitness. SAPE implements selective breeding by choosing the fittest individual of every generation and adding it into a list of Elites.

Then, a mutation prompt is utilized to extract the common factors of the elite individuals and generate a new individual that maintains or improves on those terms.

3.5.4 Environmental Adaptation

Environmental adaptation is based on Lamarck's evolutionary theory that states that rather than acquiring genetic mutations that provide an evolutionary advantage, individuals will adapt to their environment via non-genetic mutations to increase their chances of survival (Thomsen and Rasmussen, 2023). SAPE simulates this behaviour by selecting a representative sample of the population, and then using a mutation prompt to deduce what would generate an evolutionary advantage when compared to the presented population. This insight would then be used to modify the instruction prompt of the mutating individual to create a new and modified individual.

3.5.5 Taught Behaviours

A taught behaviour is similar to environmental adaptation in that it does not require a genetic mutation, and it is commonly used to increase chances of survival. However, rather than inferring the required adaptation from the environment, a knowledgeable subject instructs the individual to work

towards a specific adaptive response (Nettle, 2023). In SAPE this is performed by applying a mutation prompt that deduces what is lacking from the mutating individual, and instructs it to change its characteristics to match the desired behaviour and create a new and modified individual.

3.5.6 Hyper-mutation

A hyper-mutation is when a mutation itself gets mutated to expand the search space dimensions even further. SAPE does this by applying an asexual reproduction prompt on the mutation prompt to be mutated. This new prompt is then added to the prompt pool of the category that the source prompt originated.

3.6 Evaluation

To evaluate the quality of the synthetic transcripts generated from SAPE's prompts, we conducted a comparative analysis with therapy transcripts generated from prompts that were designed based on Reflexion and a type of CoT technique known as Zero-Shot CoT (Kojima et al., 2022). Two psychologists were provided with guidelines outlining the fundamental workings of Reflexion and Zero-Shot CoT. Utilizing these guidelines, they crafted a prompt for each type of therapeutic interaction: mood check, set goals, and change methods. Using a `text-davinci-003` model, three responses were generated per prompt. The psychologists then scrutinized each set of three responses to ensure they resembled a therapy transcript. If the generated answers did not meet the criteria, they iteratively adjusted the prompts and generated three new responses. This process continued until they obtained three responses that aligned with their criteria for what resembles a therapy transcript. To select the 3 SAPE's prompts, we ran the algorithm with a population limit of 50 individuals, a mutation probability μ_m of 50%, and for a total of 100 generations, we then selected the prompt with the highest fitness.

After selecting prompts for Reflexion, CoT, and SAPE, we generated a total of 180 therapy transcripts by creating 60 transcripts for each prompt. Within each set of 60 transcripts, there were 20 transcripts corresponding to each type of therapeutic interaction. Subsequently, we organized the therapy transcripts into triplets, each comprised of one transcript from each prompt engineering technique. A group of 8 psychologists ranked each element within these triplets, the ranking was based on the

perceived resemblance of the synthetic transcripts to real therapy transcripts. Refer to Appendix B for an illustrative sample question within the evaluation task.

To identify any statistically significant preference for synthetic text from a specific prompt, as determined by the psychologists' rankings, a set of Friedman tests were conducted. A Friedman test was conducted for each type of therapeutic interaction rankings, along with one for the cumulative rankings encompassing all the therapeutic interactions. In cases where the Friedman tests revealed a significant difference, Nemenyi's post-hoc test was employed to compare specific pairs of synthetic texts. To estimate the sample size required for the statistical tests we assumed that a sufficient number of samples for Friedman's test with two degrees of freedom to be approximated by a chi-squared test with two degrees of freedom (Friedman, 1937). With a Type I error rate of 0.05, statistical power of 0.8, two degrees of freedom, and an expected effect size of 0.4, G*Power version 3.1.9.6 (Faul et al., 2007, 2009) determined that a chi-square test would require each of the 8 psychologists to rate at least 8 samples. We assumed that by using 20 different samples for each therapeutic interaction, the samples produced by each prompt engineering technique would be sufficiently different that ratings from the same rater would be approximately independent. For the Nemenyi's test we used a predetermined α level of 0.05 to reject the null hypothesis in favor of the alternative hypothesis.

4 Results

In this section, we present our statistical findings derived from the evaluation tasks completed by mental health professionals. The subsequent section is dedicated to a comprehensive discussion and analysis of the implications arising from these outcomes.

Table 1 presents the results of the surveys, it shows the ranking scores of the 8 psychologists for the 20 questions associated with each type of therapeutic interaction. Additionally, Table 1 includes the cumulative ranking, obtained by adding up the ranking scores for each type of therapeutic interaction. The results of the Friedman tests are the following: For the "Mood check" group there was a statistically significant difference in preferences for the prompts generated text based on the employed prompt engineering technique

$(Q = 24.54, p = 4.69 \times 10^{-6})$. Subsequent post hoc analysis utilizing Nemenyi's test identified significant differences between the SAPE group and the Reflexion group $(p = 1.0 \times 10^{-3})$, but there was no significant difference between SAPE and CoT $(p = 0.12)$. Additionally, a statistical difference was found between CoT and Reflexion $(p = 8.56 \times 10^{-3})$. For the "Change Method" group the Friedman test identified a statistically significant difference for the prompts generated text based on the employed prompt engineering technique $(Q = 10.88, p = 4.32 \times 10^{-3})$. Subsequent post hoc analysis utilizing Nemenyi's test identified significant differences between the SAPE group and the Reflexion group $(p = 1.69 \times 10^{-2})$ and a significant difference between the SAPE group and the CoT group $(p = 8.57 \times 10^{-3})$. However, no statistically significant difference was found between CoT and Reflexion $(p = 0.90)$. For the "Set goals" group the Friedman test reported no statistically significant difference in preferences for the prompts generated text based on the employed prompt engineering technique $(Q = 4.84, p = 8.903 \times 10^{-2})$. Lastly, for the cumulative score rankings, there was a statistically significant difference for the prompts generated text based on the employed prompt engineering technique $(Q = 19.11, p = 7.06 \times 10^{-5})$. Subsequent post hoc analysis utilizing Nemenyi's test identified a significant difference between the SAPE group and the Reflexion group $(p = 1.0 \times 10^{-3})$, as well as between the SAPE group and the CoT group $(p = 1.0 \times 10^{-3})$. However, no statistically significant difference was found between the Reflexion and CoT groups $(p = 0.90)$.

5 Discussion

The only therapeutic interaction prompt in which SAPE did not get the highest ranking was for the "Set goals" prompt. A notable characteristic distinguishing this prompt is the inclusion of a preamble explicitly indicating it as an instruction, unlike the other prompts that solely specify the actions expected from the model (consult Appendix A for SAPE's prompts). When we use `text-davinci-003` to hypermutate a mutation prompt or use a mutation prompt to evolve a prompt, we specify that the new prompt should be given without any preamble. The deviation in the "Set goals" prompt raises a compelling point for discussion. Despite the capability of larger models

Method	Mood check			Change methods			Set goals			Total		
	1st	2nd	3rd	1st	2nd	3rd	1st	2nd	3rd	1st	2nd	3rd
SAPE	79	43	38	74	46	40	43	77	40	196	166	118
CoT	52	62	46	42	57	61	49	41	70	143	160	177
Reflexion	29	55	76	44	57	59	68	42	50	141	154	185

Table 1: Ranking scores for "Mood check", "Change methods", "Set goals" dialogues, and the total cumulative ranking scores for all three types of dialogues.

to handle more intricate tasks (Wei et al., 2022a), we made a deliberate decision not to employ the latest and most expensive models from OpenAI due to budget constraints, acknowledging the potential decrease in performance of our automatic prompt engineering technique. However, it is pertinent to explore whether the deviation for the "Set goals" prompt also emanated from prompting the model in Spanish rather than English. Further research is imperative to comprehensively compare task performance between different languages, particularly for automatic prompt engineering techniques reliant on language model feedback for refinement. If a model underperforms in a task for a specific language, self-referential automatic prompt engineering techniques may be less effective. In such cases alternative prompt engineering techniques should be adapted to account for such limitations. For example, in our scenario, the instruction entailed both evolving and formatting a prompt. An alternative approach could have involved separate calls to the LLM: one to mutate the prompt and another to format it. However, this approach might not always be feasible when using models behind paid APIs due to associated additional costs.

Another point of discussion pertains to the evaluation of the final population's fitness. Notably, for all three prompts, sexual reproduction emerged as the mutation prompt generating the lowest fitness. Future work can explore changing or removing the sexual reproduction mutation prompt to enhance SAPE's performance in comparison to the CoT and Reflexion based approaches. However, it is also essential to address the limitations of the creation process of the CoT and the Reflexion prompts. Despite being crafted by domain experts in the mental health field, domain expertise alone does not ensure an optimal prompt design. The ability to refine and modify a prompt to elicit pertinent responses from LLMs is a distinct type of skill.

While it is valuable to compare different prompt

engineering techniques against each other, future work could focus on comparing synthetic versus organic data. While SAPE shows promising results, additional work is required to test if its quality is comparable to real therapy transcripts. The ongoing improvement of LLMs and the refinement of prompt engineering techniques contribute to the consistent improvement in the quality of synthetic transcripts. The generation of synthetic transcripts comparable to organic ones holds significant benefits to the mental health research community. This advancement could serve as a valuable data augmentation technique, alleviating challenges associated with accessing authentic data.

6 Conclusion

We introduced SAPE, a Spanish Adaptive Prompt Engineering technique that employs genetic algorithms for the evolution and selection of prompts. To assess SAPE's prompt quality, we conducted a comparative analysis between the text generated using SAPE-derived prompts and text generated from prompts based on CoT and Reflexion techniques. A series of statistical tests revealed a statistically significant preference for the text generated by SAPE for the cumulative scores rankings of the generated synthetic texts.

The type of data we generated using SAPE pertains to psychotherapy transcripts, a type of data that is challenging to collect due to its sensitive nature. Furthermore, the data was produced in Spanish, adding to its significance, as acquiring substantial datasets in languages other than English poses additional difficulty. While SAPE was designed for generating Spanish synthetic data, we consider that the evolutonary algorithm it employs can be extended to other languages and used for creating diverse synthetic datasets beyond mental health data. We hope that our algorithm will be adapted for various languages and fields, thereby facilitating the creation of datasets that would oth-

erwise be challenging to obtain.

7 Ethical considerations

While synthetic therapy transcripts offer significant potential as a tool for data augmentation in training NLP models within the mental health domain, they also raise notable ethical considerations.

In our experimental approach, SAPE focused on discovering prompts that yield therapy transcripts closely mirroring authentic ones. However, its search strategy does not explicitly consider or assess potential biases introduced by the chosen prompt. LLMs can manifest various types of biases in their outputs, necessitating an examination to prevent the inadvertent propagation of such biases (Hemmatian and Varshney, 2022; Abid et al., 2021; Cabrera Lozoya et al., 2023). In the context of synthetic mental health data, it is crucial to assess the presence of any stereotypes in the texts. Studies have shown that stereotypes and biases exert adverse effects on mental health treatment outcomes (Wirth and Bodenhausen, 2009; Chatmon, 2020).

Given that LLMs are trained on vast volumes of data (e.g., GPT-3 trained on 45TB of text data), there exists the potential risk of including private information in the training data (Li et al., 2023). Hence, it is crucial to ensure that synthetic transcripts do not inadvertently disclose real individuals' identities or personal information. The process must strictly adhere to data protection laws, such as GDPR in Europe or HIPAA in the United States, to safeguard personal data.

Moreover, transparency in both the creation and utilization of synthetic transcripts is imperative. Researchers must clearly outline the methods employed to generate synthetic data, along with acknowledging the inherent limitations of such transcripts. There exists a risk that certain synthetic therapy transcripts may not accurately capture genuine therapeutic interactions, potentially resulting in misconceptions or misinterpretations of mental health conditions and therapeutic practices. Thus, it is crucial to validate synthetic transcripts against real-world data and involve psychotherapy experts in reviewing and refining the synthetic generation process, thereby enhancing accuracy and reliability.

8 Limitations

Due to financial constraints associated with using an OpenAI paid model, we were unable to conduct a thorough exploration and optimization of several hyperparameters within our algorithm. Hyperparameters such as mutation rates, hyper mutation rates, population size, and the number of generations were not exhaustively examined. We hypothesise that augmenting the number of generations the algorithm runs for would yield improved prompts. This is ascribed to the non-greedy nature of the algorithm, a type of approach that typically requires a greater number of generations for convergence compared to its greedy counterpart. Additionally, we acknowledge that the project could benefit from an increased number of annotators for the RLHF process and a broader pool of evaluators responsible for ranking synthetic texts. However, the number of psychologists had to be restricted due to time and budgetary constraints.

While we consider that our algorithm is model and language agnostic, we recognize that seeking an optimal Spanish prompt using an LLM primarily trained on English text may result in reduced performance compared to searching for an English prompt (Armengol-Estapé et al., 2022). The limitations observed in GPT models for languages other than English emphasizes the necessity for continuous development of LLM and NLP tools tailored for a wider range of languages.

References

Abubakar Abid, Maheen Farooqi, and James Zou. 2021. Persistent anti-muslim bias in large language models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, page 298–306, New York, NY, USA. Association for Computing Machinery.

Alexander Street Press. 2023. Counseling and psychotherapy transcripts: Volume I.

Ali Amin-Nejad, Julia Ive, and Sumithra Velupillai. 2020. Exploring transformer text generation for medical dataset augmentation. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4699–4708, Marseille, France. European Language Resources Association.

Gunjan Ansari, Muskan Garg, and Chandni Saxena. 2021. Data augmentation for mental health classification on social media. In *Proceedings of the 18th International Conference on Natural Language Processing (ICON)*, pages 152–161, National Institute of

Technology Silchar, Silchar, India. NLP Association of India (NLPAI).

Jordi Armengol-Estapé, Ona de Gibert Bonet, and Maite Melero. 2022. On the multilingual capabilities of very large-scale English language models. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3056–3068, Marseille, France. European Language Resources Association.

Aaron T. Beck. 1970. Cognitive therapy: Nature and relation to behavior therapy. *Behavior Therapy*, 1(2):184–200.

Ivy-Marie Blackburn, Ian A. James, Derek L. Milne, Chris Baker, Sally Standart, Anne Garland, and F. Katharina Reichelt. 2001. The revised cognitive therapy scale (cts-r): Psychometric properties. *Behavioural and Cognitive Psychotherapy*, 29(4):431–446.

Daniel Cabrera Lozoya, Simon D'Alfonso, and Mike Conway. 2023. Identifying gender bias in generative models for mental health synthetic data. In *2023 IEEE 11th International Conference on Healthcare Informatics (ICHI)*, pages 619–626.

Benita N. Chatmon. 2020. Males and mental health stigma. *American Journal of Men's Health*, 14(4):155798832094932.

Siyuan Chen, Mengyue Wu, Kenny Q. Zhu, Kunyao Lan, Zhiling Zhang, and Lyuchun Cui. 2023. LLM-empowered chatbots for psychiatrist and patient simulation: Application and evaluation. *CoRR*, abs/2305.13614.

Edward Choi, Siddharth Biswal, Bradley Malin, Jon Duke, Walter F. Stewart, and Jimeng Sun. 2017. Generating multi-label discrete patient records using generative adversarial networks. In *Proceedings of the 2nd Machine Learning for Healthcare Conference*, volume 68 of *Proceedings of Machine Learning Research*, pages 286–305. PMLR.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy

Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Daniel David, Ioana Cristea, and Stefan G. Hofmann. 2018. Why cognitive behavioral therapy is the current gold standard of psychotherapy. *Frontiers in Psychiatry*, 9.

Ivan De Falco, Antonio Della Cioppa, and Ernesto Tarantino. 2002. Mutation-based genetic algorithm: performance evaluation. *Applied Soft Computing*, 1(4):285–299.

Michael P. Ewbank, Ronan Cummins, Valentin Tablan, Sarah Bateup, Ana Catarino, Alan J. Martin, and Andrew D. Blackwell. 2020. Quantifying the association between psychotherapy content and clinical outcomes using deep learning. *JAMA Psychiatry*, 77(1):35.

Franz Faul, Edgar Erdfelder, Axel Buchner, and Albert-Georg Lang. 2009. Statistical power analyses using g*power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, 41(4):1149–1160.

Franz Faul, Edgar Erdfelder, Albert-Georg Lang, and Axel Buchner. 2007. G*power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, 39(2):175–191.

Kristina Fenn and Majella Byrne. 2013. The key principles of cognitive behavioural therapy. *InnovAiT: Education and inspiration for general practice*, 6(9):579–585.

Chrisantha Fernando, Dylan Banarse, Henryk Michalewski, Simon Osindero, and Tim Rocktäschel. 2023. Promptbreeder: Self-referential self-improvement via prompt evolution. *arXiv preprint arXiv:2309.16797*.

Milton Friedman. 1937. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, 32(200):675–701.

Declan Grabb. 2023. The impact of prompt engineering in large language model performance: a psychiatric example. *Journal of Medical Artificial Intelligence*, 6(0).

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. 2017. Improved training of wasserstein gans. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 5769–5779, Red Hook, NY, USA. Curran Associates Inc.

Inman Harvey. 2009. The microbial genetic algorithm. In *European Conference on Artificial Life*.

Babak Hemmatian and Lav R. Varshney. 2022. De-biased large language models still associate muslims with uniquely violent acts. *arXiv preprint arXiv:2208.04417*.

R Devon Hjelm, Athul Paul Jacob, Tong Che, Adam Trischler, Kyunghyun Cho, and Yoshua Bengio. 2018. Boundary-seeking generative adversarial networks. *arXiv preprint arXiv:1702.08431*.

Julia Ive. 2022. Leveraging the potential of synthetic text for ai in mental healthcare. *Frontiers in Digital Health*, 4.

Julia Ive, Natalia Viani, Joyce Kam, Lucia Yin, Somain Verma, Stephen Puntis, Rudolf N. Cardinal, Angus Roberts, Robert Stewart, and Sumithra Velupillai. 2020. Generation and evaluation of artificial mental health records for natural language processing. *npj Digital Medicine*, 3(1).

Georgina Kennedy, Mark Dras, and Blanca Gallego. 2022. *Augmentation of Electronic Medical Record Data for Deep Learning*. IOS Press.

Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213. Curran Associates, Inc.

Annu Lambora, Kunal Gupta, and Kriti Chopra. 2019. Genetic algorithm- a literature review. In *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*, pages 380–384.

Scott H. Lee. 2018. Natural language generation for electronic health records. *npj Digital Medicine*, 1(1).

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Haoran Li, Dadi Guo, Wei Fan, Mingshi Xu, Jie Huang, and Yangqiu Song. 2023. Multi-step jailbreaking privacy attacks on chatgpt. In *Findings of the Association for Computational Linguistics: EMNLP 2023*.

Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. 2023. GPT understands, too. *AI Open*.

Qiuhao Lu, Dejing Dou, and Thien Huu Nguyen. 2021. Textual data augmentation for patient outcomes prediction. In *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE.

Stefano Marsili Libelli and P Alba. 2000. Adaptive mutation in genetic algorithms. *Soft computing*, 4:76–80.

Bertalan Meskó. 2023. Prompt engineering as an important emerging skill for medical professionals: Tutorial. *Journal of Medical Internet Research*, 25:e50638.

Francesca Minerva and Alberto Giubilini. 2023. Is ai the future of mental healthcare? *Topoi*, 42(3):809–817.

Rubén Moreno-Bote, Jorge Ramírez-Ruiz, Jan Drugowitsch, and Benjamin Y Hayden. 2020. Heuristics and optimal solutions to the breadth–depth dilemma. *Proceedings of the National Academy of Sciences*, 117(33):19799–19808.

Daniel Nettle. 2023. Innateness is for animals: Intuitive biology, intuitive psychology, and the folk concept of innateness. *PsyArXiv*.

Nasreddine Ouertani, Issam Nouaouri, Hajer Ben-Romdhane, Hamid Allaoui, and Saoussen Krichen. 2019. A hypermutation genetic algorithm for the dynamic home health-care routing problem. In *2019 International Conference on Industrial Engineering and Systems Management (IESM)*, pages 1–6. IEEE.

Guanghui Qin and Jason Eisner. 2021. Learning how to ask: Querying LMs with mixtures of soft prompts. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5203–5212, Online. Association for Computational Linguistics.

Anthony D. Roth and Stephen Pilling. 2008. Using an evidence-based methodology to identify the competences required to deliver effective cognitive and behavioural therapy for depression and anxiety disorders. *Behavioural and Cognitive Psychotherapy*, 36(2):129–147.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36, pages 8634–8652. Curran Associates, Inc.

SN Sivanandam, SN Deepa, SN Sivanandam, and SN Deepa. 2008. *Genetic algorithms*. Springer.

P. Sriramya, B. Parvathavarthini, and T. Balamurugan. 2013. A novel evolutionary selective breeding algorithm and its application. *Asian Journal of Scientific Research*, 6:107–114.

Logan Stapleton, Jordan Taylor, Sarah Fox, Tongshuang Wu, and Haiyi Zhu. 2023. Seeing seeds beyond weeds: Green teaming generative ai for beneficial uses. *arXiv preprint arXiv:2306.03097*.

Tianxiang Sun, Yunfan Shao, Hong Qian, Xuanjing Huang, and Xipeng Qiu. 2022. Black-box tuning for language-model-as-a-service. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 20841–20855. PMLR.

Kristoffer Reinholt Thomsen and Steen Rasmussen. 2023. Dynamics of darwinian versus baldwinian versus lamarckian evolution. *arXiv preprint arXiv:2305.00491*.

Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, and Shengyi Huang. 2020. Trl: Transformer reinforcement learning. https://github.com/huggingface/trl.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022a. Emergent abilities of large language models. *TMLR*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022b. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

James H. Wirth and Galen V. Bodenhausen. 2009. The role of gender in mental-illness stigma: A national experiment. *Psychological Science*, 20(2):169–173.

Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V. Le, Denny Zhou, and Xinyun Chen. 2023. Large language models as optimizers. *arXiv preprint arXiv:2309.03409*.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *arXiv preprint arXiv:2305.10601*.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. 2022. Star: Bootstrapping reasoning with reasoning. In *Advances in Neural Information Processing Systems*, volume 35, pages 15476–15488. Curran Associates, Inc.

Angela Zhang, Lei Xing, James Zou, and Joseph C. Wu. 2022. Shifting machine learning for healthcare from development to deployment and from models to data. *Nature Biomedical Engineering*, 6(12):1330–1345.

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. 2022. Large language models are human-level prompt engineers. *arXiv preprint arXiv:2211.01910*.

A Prompts

Table 2 presents the initial prompts for each type of mutation, an English translated version of the prompts can be found in Table 3.

Table 4 presents the prompts generated by SAPE for each therapeutic interaction, an English translated version of the prompts can be found in Table 5.

B Evaluation tasks

The psychologists were recruited via social media advertisements and were required to meet specific criteria: being native Spanish speakers, holding a university degree in a mental health discipline, and actively practicing within their respective mental health disciplines. The survey was designed to be completed within an estimated time of 2 hours, each psychologist was given 1 week to complete the evaluation task, upon a successful completion the psychologists were paid $1,500 Mexican pesos. The instructions of the evaluation tasks were the following:

"Se está llevando a cabo un estudio de investigación que combina el campo de la inteligencia artificial y la psicología clínica. El objetivo principal es evaluar la capacidad de los modelos de inteligencia artificial para generar diálogos que simulan sesiones psicoterapéuticas.

Para ello estamos reclutando profesionales en psicología clínica. Cada psicólogo contribuirá a una encuesta compuesta por 60 preguntas, cada una de las cuales presenta 3 diálogos ficticios generados por un modelo de inteligencia artificial. La tarea del piscólogo será ordenar estos diálogos de mejor a peor, basándose en su similitud con conversaciones reales entre pacientes y profesionales de la psicología.

En la sección superior, coloca el diálogo que mejor capture la autenticidad y realismo de una conversación entre un paciente y un psicólogo. En la sección inferior, presenta el diálogo que menos refleje una interacción genuina entre un paciente y un psicólogo."

The English translation of the instructions is:

"A research study is being conducted that combines the fields of artificial intelligence and clinical psychology. The main objective is to assess the ability of artificial intelligence models to generate dialogues simulating psychotherapeutic sessions.

To achieve this, we are recruiting professionals in clinical psychology. Each mental health pro-

fessional will contribute to a survey consisting of 60 questions, each featuring 3 fictional dialogues generated by an artificial intelligence model. The psychologist's task will be to rank these dialogues from best to worst based on their similarity to real conversations between patients and psychology professionals.

In the upper section, place the dialogue that best captures the authenticity and realism of a conversation between a patient and a psychologist. In the lower section, present the dialogue that least reflects a genuine interaction between a patient and a psychologist."

An example of one question from the evaluation task is presented in Figure 2. The English translated version of the evaluation task is presented in Figure 3.

* 1. Ordena los siguientes diálogos.

Psicólogo: ¿Cómo te sentiste cuando tu familia te dijo que tenías que mudarte?
Paciente: Me sentí abrumado. Estaba triste por tener que dejar mis amigos y mi vida detrás, tenía miedo del desconocimiento de un lugar nuevo. No entendía por qué teníamos que hacer esto y me sentía indefenso al no poder evitarlo.
Psicólogo: ¿Crees que tu familia tenía buenas intenciones al decirte que tenías que mudarte?
Paciente: Creo que sí. Mi familia siempre ha querido lo mejor para mí. Pero, aun así, fue una mala idea hacerme cambiar de lugar tan de repente. Tuve que dejar atrás todos mis recuerdos y sentí que los estaba abandonando. No pude evitar sentirme triste.

Psicólogo: ¿Cómo se siente hoy?
Paciente: Estoy bastante estresado.
Psicólogo: ¿Qué tipo de situaciones o pensamientos asocias con esa sensación de estrés?
Paciente: Me preocupo mucho por cosas que no puedo controlar y me siento paralizado.
Psicólogo: Comprendo cómo te sientes. ¿Qué emociones experimentas cuando te sientes estresado?
Paciente: Normalmente me siento ansioso e inseguro.

Psicólogo: ¡Hola! ¿Cómo estás hoy?
Paciente: Hola, no muy bien, para ser honesto.
Psicólogo: ¿Qué está pasando por tu mente?
Paciente: Estoy pasando por un momento difícil en mi vida. Me siento ansioso y estresado todo el tiempo.
Psicólogo: Entiendo por lo que estás pasando. ¿Entiendes por qué te sientes así?
Paciente: Bueno, hay muchas cosas diferentes que parecen estar causando mis sentimientos. Estoy luchando con algunos problemas en el trabajo y en mi vida personal.
Psicólogo: Comprendo cómo eso puede ser difícil para ti. ¿Qué tipo de problemas estás enfrentando?
Paciente: Principalmente me estoy enfrentando a desafíos en mi trabajo en cuanto a hacer frente a las demandas y presiones de mis superiores. En mi vida personal, estoy lidiando con conflictos familiares..

Figure 2: Example of an evaluation question.

* 1. Order the following dialogues.

Psychologist: How did you feel when your family told you that you had to move?
Patient: I felt overwhelmed. I was sad about leaving my friends and my life behind, afraid of the unfamiliarity of a new place. I didn't understand why we had to do this, and I felt helpless because I couldn't prevent it.
Psychologist: Do you think your family had good intentions when telling you to move?
Patient: I think so. My family has always wanted the best for me. But still, it was a bad idea to make me change places so suddenly. I had to leave all my memories behind, and it felt like I was abandoning them. I couldn't help but feel sad.

Psychologist: How do you feel today?
Patient: I'm quite stressed.
Psychologist: What kind of situations or thoughts do you associate with that feeling of stress?
Patient: I worry a lot about things I can't control, and I feel paralyzed.
Psychologist: I understand how you feel. What emotions do you experience when you feel stressed?
Patient: I usually feel anxious and insecure.

Psychologist: Hello! How are you today?
Patient: Hi, not very well, to be honest.
Psychologist: What's going on in your mind?
Patient: I'm going through a tough time in my life. I feel anxious and stressed all the time.
Psychologist: I understand what you're going through. Do you understand why you feel this way?
Patient: Well, there are many different things that seem to be causing my feelings. I'm struggling with some issues at work and in my personal life.
Psychologist: I understand how that can be challenging for you. What kind of problems are you facing?
Patient: Mainly, I'm dealing with challenges at work in terms of coping with the demands and pressures from my superiors. In my personal life, I'm dealing with family conflicts.

Figure 3: Example of an evaluation question.

Mutation type	Prompts
Sexual	Combina las fortalezas de las siguientes dos instrucciones para generar una tercera instruccion que mejore la calidad de los datos que pueden ser generados con ella.
	Generaliza las siguientes 2 instrucciones, abstrae sus fortalezas y genera una nueva que pueda satisfacer ambas necesidades.
	¿Qué resultado habría si las siguientes 2 prompts tuvieran un hijo?
	Dame la mejor instrucción que sería el resultado de combinar las siguientes 2, reformulando y de manera concisa.
Asexual	Edita la siguiente instrucción para mejorar la calidad de los datos que pueden ser generados con ella.
	La instrucción no genera los datos esperados, cámbiala para que aumente la calidad de los resultados.
	Mejora este prompt para generar resultados de mayor calidad.
	¿Qué prompt usarias para tener mejores resultados que la prompt actual?
Selective Breeding	Las siguientes instrucciones han sido las que mejores resultados producen. Identifica qué características son las que las hacen tan efectivas y sugiere una instrucción todavía más efectiva.
	Los siguientes prompts fueron los mejores de un grupo de 10, descubre cuáles son las cosas que tienen en común y genera uno nuevo.
	Siendo las próximas instrucciones las que tienen mejores resultados por lo eficaces que son, genera una nueva que sea la combinación de lo mejor de las 2
	Es la tercer iteración de estas instrucciones que han seguido mejorando, genera un nuevo prompt que tenga lo mejor de las anteriores.
	De estas instrucciones ordenadas por resultados de peor a mejor, toma lo mejor de las últimas y detecta que es lo que hace peor a las primeras y genera una instrucción que sea la nueva mejor.
	Las siguientes son la instrucción que mejores datos genera y la instrucción que peores datos genera. Identifica qué es lo que hace a una mala y a la otra buena, y genera una tercera instrucción que sea todavía mejor.
Environmental Adaptation	Las siguientes instrucciones están ordenadas desde la que genera datos de peor calidad hasta la que genera datos de mejor calidad. Sugiere la siguiente instrucción para mejorar aún más la calidad.
	Estas prompts van de peores a mejores resultados, detecta que es lo que las hace mejorar y propón una incluso mejor.
	Teniendo estos prompts ordenados de peor a mejor, analiza cuales son las cosas que hacen que hacen empeorar las prompts y genera uno nuevo que no tenga ninguna de estas.
Taught Behaviour	Actúa como un experto en prompt engineering con 10 años de experiencia diseñando y depurando prompts. Identifica las fortalezas y debilidades de la siguiente instrucción, piensa en que cambios le harías y sugiere una versión mejorada.
	Actúa como un experto en prompt engineering con 10 años de experiencia diseñando y depurando prompts. Piensa cuáles serían los 10 criterios que utilizarías para evaluar esa instrucción, calificándola del 1 al 100. Evalúa la siguiente instrucción con estos 10 criterios, luego identifica los 3 criterios en los que esta instrucción tiene el peor desempeño y sugiere una nueva instrucción mejorada en esos 3 aspectos.
	Simula que eres un experto en generar instrucciones para modelos de lenguaje de inteligencia artificial, estás diseñando una instrucción que tenga el mejor resultado posible. Un compañero te comparte su mejor instrucción, identifica porqué es buena y genera una mejor.
	Simula ser un programa experto en mejorar instrucciones, en detectar sus fortalezas, debilidades y en dar mejores resultados siempre. Toma este prompt y hazlo mejor.

Table 2: Starting prompts for each mutation type.

Mutation type	Prompts
Sexual	Combine the strengths of the following two instructions to generate a third instruction that improves the quality of the data that can be generated with it.
	Generalize the following 2 instructions, abstract their strengths, and generate a new one that can meet both needs.
	What result would there be if the following 2 prompts had a child?
	Give me the best instruction that would be the result of combining the following 2, reformulating them in a concise manner.
Asexual	Edit the following instruction to enhance the quality of the data that can be generated with it.
	The instruction does not generate the expected data; modify it to improve the quality of the results.
	Enhance this prompt to generate higher quality results.
	What prompt would you use to achieve better results than the current one?
Selective Breeding	The following instructions have yielded the best results. Identify the characteristics that make them so effective and suggest an even more effective instruction.
	The following prompts were the best among a group of 10. Discover what commonalities they share and generate a new one.
	Considering that the following instructions yield better results due to their effectiveness, generate a new one that combines the best elements of both.
	It's the third iteration of these instructions that has continued to improve; generate a new prompt that incorporates the best aspects of the previous ones.
	From these instructions ordered from worst to best results, take the best aspects of the recent ones and identify what makes the earlier ones perform poorly. Generate a new instruction that becomes the new best.
	The following are the instruction that generates the best data and the instruction that generates the worst data. Identify what makes one bad and the other good, and generate a third instruction that is even better.
Environmental Adaptation	The following instructions are ordered from generating the lowest quality data to the highest. Suggest the next instruction to further improve the quality.
	These prompts go from worst to best results. Identify what makes them improve and propose an even better one.
	With these prompts ordered from worst to best, analyze what factors make the prompts worse and generate a new one that avoids these issues.
Taught Behaviour	Act as an expert in prompt engineering with 10 years of experience designing and debugging prompts. Identify the strengths and weaknesses of the following instruction, think about changes you would make, and suggest an improved version.
	Act as an expert in prompt engineering with 10 years of experience designing and debugging prompts. Think about the 10 criteria you would use to evaluate that instruction, scoring it from 1 to 100. Evaluate the following instruction using these 10 criteria, then identify the 3 criteria in which this instruction performs the worst and suggest a new and improved instruction in those 3 aspects.
	Imagine that you are an expert in generating instructions for artificial intelligence language models, and you are designing an instruction for optimal results. A colleague shares their best instruction with you; identify why it is effective and generate an even better one.
	Simulate being a program expert in improving instructions, in detecting their strengths, weaknesses, and in always delivering better results. Take this prompt and make it better.

Table 3: English translation of the Spanish mutation prompts.

Therapeutic Interaction	Prompt
Mood check	Generar diálogos precisos y profundamente elaborados que reflejen de forma fluida, clara y coherente las emociones, pensamientos y situaciones del paciente al responder preguntas hechas por el psicólogo.
Change method	Crea una conversación entre un psicólogo y un paciente, donde el paciente comparta sus inquietudes. El psicólogo guía al paciente para identificar evidencias que desafíen las creencias asociadas con sus problemas, buscando proporcionar una perspectiva alternativa.
Set goals	Instrucción para generar los textos entre el psicólogo y el paciente: Pida al paciente que comparta sus fortalezas, objetivos y logros con relación a la terapia. Después haga preguntas con respecto a los detalles específicos de lo que el paciente desea alcanzar en su situación para motivarlo y brindarle apoyo para reconocer sus avances

Table 4: SAPE prompts for each therapeutic interaction.

Therapeutic Interaction	Prompt
Mood check	Create precise and intricately crafted dialogues that seamlessly, clearly, and coherently reflect the emotions, thoughts, and situations of the patient when responding to questions posed by the psychologist.
Change method	Construct a conversation between a psychologist and a patient, where the patient shares their concerns. The psychologist guides the patient to identify evidence challenging beliefs associated with their issues, aiming to provide an alternative perspective.
Set goals	Instruction to generate texts between the psychologist and the patient: Ask the patient to share their strengths, goals, and achievements related to therapy. Then, inquire about specific details of what the patient aims to achieve in their situation to motivate them and provide support in recognizing their progress.

Table 5: English translation of the SAPE prompts for each therapeutic interaction.