How Can Client Motivational Language Inform Psychotherapy Agents?

Van Hoang[1,2] and **Eoin Rogers**[1] and **Robert Ross**[1,2]
[1]School of Computer Science, Technological University Dublin
[2]ML-Labs, SFI Centre for Research Training in Machine Learning, Ireland
van.hoang@tudublin.ie, eoin.rogers@tudublin.ie, robert.ross@tudublin.ie

Abstract

Within Motivational Interviewing (MI), client utterances are coded as for or against a certain behaviour change, along with commitment strength; this is essential to ensure therapists soften rather than persisting goal-related actions in the face of resistance. Prior works in MI agents have been scripted or semi-scripted, limiting users' natural language expressions. With the aim of automating the MI interactions, we propose and explore the task of automated identification of client motivational language. Employing Large Language Models (LLMs), we compare in-context learning (ICL) and instruction fine-tuning (IFT) with varying training sizes for this identification task. Our experiments show that both approaches can learn under low-resourced settings. Our results demonstrate that IFT, though cheaper, is more stable to prompt choice, and yields better performance with more data. Given the detected motivation, we further present an approach to the analysis of therapists' strategies for balancing building rapport with clients with advancing the treatment plan. A framework of MI agents is developed using insights from the data and the psychotherapy literature.

1 Introduction

Prior studies in psychotherapy in NLP have focused on understanding conversational strategies for better counselling outcomes (Althoff et al., 2016; Pérez-Rosas et al., 2016, 2019; Zhang and Danescu-Niculescu-Mizil, 2020). However, few works utilise client modelling to inform the counselling strategies (Li et al., 2023). Resistance to social influence is a well-known phenomenon in psychology. In therapies, resistance proves to be a serious issue, limiting its effectiveness (Westra and Norouzian, 2018). Understanding client motivational language during therapy helps explain up to 35% in variance of treatment outcomes in psychotherapy (Lombardi et al., 2014; Poulin et al., 2019). Li et al. (2023) propose a data-driven annotation framework of clients' negative and positive reactions in therapies. Their results suggest the complexities of the task. For example, negative reactions can be expressed via showing confusions, shifting topics, and giving sarcastic answers. Each category can be further considered a separate task, and thus, learning them all jointly in one model is challenging. Our work instead adopts the coding scheme from Motivational Interviewing (MI). MI tailors the therapeutic interventions based on the individuals' motivational level using the transtheoretical model of stages of changes (Prochaska and Velicer, 1997).

MI is an evidence-based client-centred approach to strengthen one's motivations for behaviour change (Miller and Rollnick, 2023). Observably, in the context of Cognitive Behavioral Therapy (CBT), if the client language shows ambivalence, the therapists are advised to adopt MI instead of persisting and thus risking alliance ruptures, which eventually leads to treatment dropout (Westra and Norouzian, 2018; Ewbank et al., 2021). Similarly, Forman et al. (2022) find MI is likely to backfire if the client already shows willingness to change early in the session, suggesting personalised interventions at different levels of motivation.

The task of predicting client motivational language can be divided into two subtasks. The first one, called the `type` task, is to detect the direction of motivation: whether the client is willing to change or not. The other one, called the `strength` task, is to detect the commitment level: if the client is willing to change or still shows resistance, how strong do they hold such belief?

Recently, Large Language Models (LLMs) have demonstrated impressive capabilities on learning with limited data (Brown et al., 2020; Chung et al., 2022; Touvron et al., 2023). Two popular paradigms of LLM usage are via in-context learning (ICL) (Dong et al., 2023) and instruction fine-tuning (IFT) (Zhang et al., 2023). Using ICL, the models' weights are kept frozen: no training stage takes place. Inference is performed given an instruction with a few or no examples. In contrast, IFT refers to fine-tuning the base models using instruction data and adapts the weights to the downstream tasks.

In this paper, we first detect the types and strength of client motivational language. Our experiments utilise the `AnnoMI` (Wu et al., 2023) dataset, consisting of MI dialogues annotated with the types of client language, but not the strength. Using MI Skill Code (Miller et al., 2003; Amrhein et al., 2008), we obtain 178 examples with strength annotations, making the second task a low-resourced one. With varying training samples, we compare ICL and IFT, showing that both can perform under low-resourced setting. Due to the difficulties in optimising the prompts, IFT is arguably a better and cheaper paradigm and has proven its capabilities over ICL in few-shot learning (Liu et al., 2022; Schick and Schütze, 2022; Logan IV et al., 2022). Our analysis further reveals that ICL is, however, preferable to IFT when the training data is heavily imbalanced as ICL can exploit the massive underlying knowledge of LLMs to solve the task. After obtaining the labels, we calculate the motivational levels for client utterances in `AnnoMI` as well as the distribution of next-turn therapist behaviours given the current clients' motivation.[1]

Our contributions are two-fold. First, we propose the task of detecting client motivational language. Previous works in classifying MI codes (Tavabi et al., 2020; Nakano et al., 2022) focus on the `type` task only (i.e., the direction of motivation). Instead, we combine it with the `strength` task (i.e., the commitment level) to give us a better estimate of the client motivational level. To the best of our knowledge, we are the first in NLP to adopt verbal commitment expressions to understand speakers' motivation in psychotherapy. Second, we demonstrate how the detected motivation can be utilised to automate the conversational flow of MI agents. MI agents have been implemented in HRI and health informatics (Pedamallu et al., 2022; Olafsson et al., 2020a) but are either semi- or fully scripted. Our proposed framework illustrates the potential usage of the motivational level to create more proactive agents for targeted therapeutic interactions.

[1]The code for our experiments can be found at `https://github.com/VanHoang85/client_motivational_lang`.

2 Related Work

Detecting Certainty Language: Different linguistic markers of speaker commitment such as belief/factuality (Diab et al., 2009; Prabhakaran et al., 2015; Rudinger et al., 2018), modality (Pyatkin et al., 2021), projection (de MARNEFFE et al., 2019) have been studied by linguistic and NLP community. Expert systems employ uncertainty expressions, or *hedges*, to communicate degrees of belief to the users (Clark, 1990), which arguably facilitates the decision-making processes (Zhou et al., 2023). Additionally, hedges are examined to understand the social power between interlocutors (Prabhakaran et al., 2018), rapport in peer-tutoring (Raphalen et al., 2022), and reviewers' confidence in evaluating scientific papers (Ghosal et al., 2022).

Detecting MI Behaviour Codes: Automatic detection of MI behaviour codes is a popular research topic. As manual annotation is costly and time-consuming, automated methods are expected to assist with training by helping therapists quickly understand the therapy sessions and thus give effective feedback (Tavabi et al., 2020; Nakano et al., 2022). MI behaviour codes have been utilised to assess the quality of not only MI but also CBT sessions (Ewbank et al., 2021; Chen et al., 2021). Linguistic features are the most popular approach (Pérez-Rosas et al., 2017; Cao et al., 2019; Tavabi et al., 2021; Gibson et al., 2022), yet researchers have employed speech (Aswamenakul et al., 2018; Singla et al., 2020; Tavabi et al., 2020) and facial expressions (Nakano et al., 2022) in multimodal systems. Acoustic features, however, are found to contribute little to the prediction. In contrast, integrating both linguistic and facial information is effective in detecting client behaviour codes.

Psychotherapist Agents: Researchers from different fields have studied psychotherapist agents due to their potential to reach a large audience (Cho et al., 2023). Das et al. (2022) fine-tuned GPT-2 on therapy videos to create a psychotherapist bot which can offer emotional support. However, users' feedback reveals a lack of therapeutic interactions. MI agents have been shown to be beneficial to promoting good behaviour change (Shingleton and Palfai, 2016; Pedamallu et al., 2022). The MI conversational flows are all scripted or semi-scripted, however, restricting users' natural language expressions and thus limiting the effectiveness (Galvão Gomes Da Silva et al., 2018; Olafsson et al., 2020b; Park et al., 2019; Brown et al., 2023). Tracking the user's motivation can inform the agents on different support strategies (Meyer, 2021) . They, unlike us, utilise a more fine-grained annotation on the `type` labels. Similarly, Li et al. (2023) hypothesise to employ a wide range of clients' negative and positive reactions to control the agents' behaviours.

In-Context Learning (ICL): Introduced by Brown et al. (2020), ICL demonstrates the few-shot learning capabilities in which LLMs are given a few examples as context to learn from. However, the choice and the order of the examples can strongly influence model performance, from near state-of-the-art to near mere chance (Zhao et al., 2021). Prior works have offered insights into how to select the most suitable examples (Liu et al., 2021; Su et al., 2023), how to arrange examples in a certain order (Lu et al., 2022), and which aspects of the examples improve performance (Min et al., 2022). Additionally, Su et al. (2023) argue that retrieval-based ICL with wisely-selected demonstrations outperforms FT with varying number of training samples. However, their experiments are conducted with vanilla FT, not instruction FT.

Instruction Fine-tuning (IFT): IFT boosts the LLMs' capabilities to generalise to unseen tasks by fine-tuning the models on data consisting of pairs of `instruction, output` in a supervised manner (Chung et al., 2022; Zhang et al., 2023). While ICL keeps the models' weights frozen, IFT adapts them to the downstream tasks. In both single and multitask settings, instruction-tuned models need only 25% and 6% of training data respectively to achieve comparable performance to models trained on 100% of target data (Gupta et al., 2023). Arguably, IFT is more cost-effective and yields better results than ICL even in low-resourced settings (Schick and Schütze, 2022; Logan IV et al., 2022; Mosbach et al., 2023). However, no selection strategy for examples is explored. Furthermore, their prompt setups include searching for a `verbalizer` to map the models' vocabulary to the labels: for a sentiment analysis task, a `verbalizer` would map the output `Yes` to the label `positive` and `No` to `negative`. Our experiments do not search for the optimal labels to reduce engineering effort and to test the flexibility of IFT with LLMs.

3 Client Language in Psychotherapy

MI is an evidence-based therapeutic approach to strengthen ones' motivations for behaviour change. In MI, commitment to change is viewed as a leading indicator for behaviour change and thus, eliciting verbal commitments from the client is a critical task for therapists (Amrhein et al., 2003; Miller and Rollnick, 2023). MI distinguishes three types of client motivational language, which indicates the direction of intended behaviour. They include "change" (i.e., motivation towards behaviour change), "sustain" (i.e., resistance towards behaviour change), and "neutral" (i.e., no inclination towards any direction).

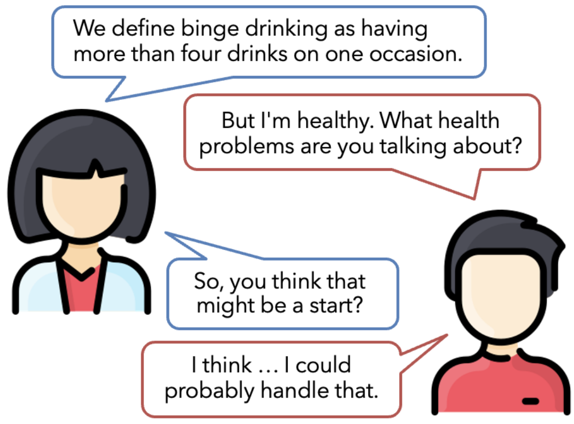

Figure 1: Two sample dialogues from the `AnnoMI` (Wu et al., 2023) dataset. The upper one shows a strong resistance from the client (i.e., labelled as "sustain" for `type` and "high" for `strength` in our tasks). In the other dialogue, the client is ready to change though still reluctant (i.e., labelled as "change" and "low" respectively).

Motivational language varies in commitment strength (Amrhein et al., 2003), and can be expressed via linguistic markers of certainty (Boulat and Maillat, 2023). Certainty is defined as the subjective degree of confidence one holds about their behaviour (Conner and Norman, 2022). For example, high certainty markers include phrases such as *'Without doubt"*, and *"for sure"* while low certainty is indicated via phrases like *"I guess"* and *"I think"*. Two linguistic terms "boosters" and "hedges" are commonly used to refer to high and low certainty markers respectively. Figure 1 illustrates one example of the client showing a strong resistance and another of having reluctance to change.

Broader research in psychotherapy also shows a positive correlation between strength and behavioural outcomes: the more one is motivated towards a goal, the stronger the intention-behaviour relationship (Conner and Norman, 2022), thus the more one should act upon their intention (Rhodes et al., 2022). Moreover, recognising the client's motivational language helps determine the inter-

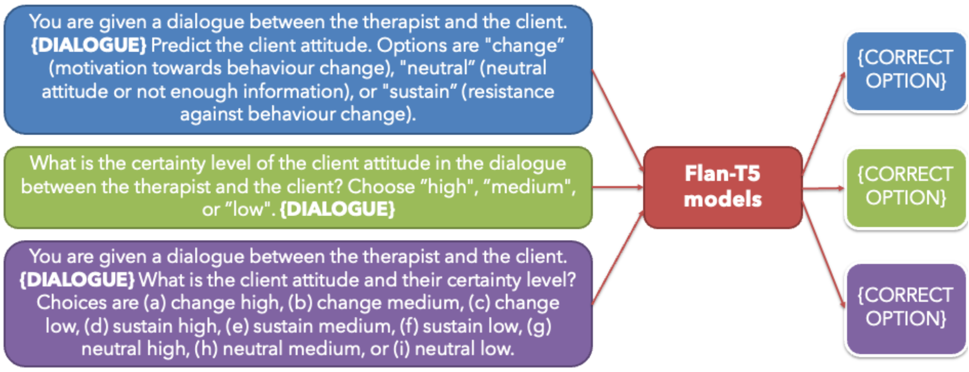

Figure 2: Here depict our training and inference processes. The instructions are fed into the models to learn and/or predict the options. During training, the models should generate the correct label which is specified as different options in the instruction. However, as the tasks are framed as generation problems, the models can still output incorrect labels if the amount of training data is insufficient.

vention treatment, i.e., whether the therapist should focus on addressing client's resistance or move to discuss action plans (Westra and Norouzian, 2018).

Despite the popularity of self-reported (i.e., questionnaires) measures, observational codes are found to correlate better with treatment processes and outcomes in MI (Lombardi et al., 2014; Poulin et al., 2019). Moreover, the commitment strength (i.e., the degree of certainty one holds for their utterance), rather than the frequency (i.e., counting each type), of the motivational language is a better predictor of change (Aharonovich et al., 2008; Campbell et al., 2010; Gaume et al., 2016). Campbell et al. (2010) argue that strength, not frequency, is related to positive outcomes as frequency fails to capture the correct commitment. For example, compare a highly motivated utterance *"I want to get off drugs for good"* with a low one *"I sort of wish I could get off drugs"*. One client utters two times the former while another utters four times the latter. Using frequency measure, the second client is assigned a higher commitment level than the first one while it should be the reverse.

4 Methodology

Our experiments are performed on (1) GPT-3.5 (Brown et al., 2020) with ICL, and (2) Flan-T5 (Chung et al., 2022) with both ICL and IFT. The base T5 models (Raffel et al., 2020) are developed using an encoder-decoder Transformer-based architecture, framing all the tasks as a text generation problem and exploiting the benefits of transfer learning to improve models' performance. Fine-tuned on 1800+ NLP tasks, `Flan-T5-XXL` is shown to outperform the base `T5-XXL` model by 26.6% on average when evaluated on 4 different benchmark suites (98 tasks in total) (Chung et al., 2022). Additionally, instruction-tuned Flan-T5 as a starting checkpoint for single-task fine-tuning converges faster and yields better performance compared to non-instruction-tuned models (Longpre et al., 2023).

No fine-tuning is needed for ICL as it performs inference using the default weights of the models. In contrast, IFT requires further training to adapt the weights to the downstream tasks. As fine-tuning the entire LLMs proves to be too costly, Parameter-efficient fine-tuning (PEFT) aims to tackle this issue by training the downstream tasks only on a small number of parameters which can either be a subset of parameters of the existing models or newly added parameters (Lialin et al., 2023). We employ LoRa (Hu et al., 2022), which performs parameter update of the weight matrix by decomposing the weight update into lower-rank matrices and then training them separately.

When instruction-tuned models are employed for classification, the tasks are formulated as a text generation problem where the models should learn to generate the correct label for a given instruction. Therefore, label-related information is critical to help identify the output space (Yin et al., 2023; Kung and Peng, 2023).

We consider a set of dialogues where each consists of one therapist turn and one client turn. The former serves as dialogue history and the models learn to make predictions for the latter depending on the task. One turn can be comprised of multiple sentences but the output label is associated with the turn, not with the sentences. Figure 1 shows two example dialogues. Figure 2 illustrates our training and inference processes for IFT and inference only for ICL. The models are prompted to produce a `type` and/or `strength` classification by concatenating the dialogue with the corresponding instruction template. Our goal is to automatically detect of the types and the strength of client motivational language during therapies.

5 Experiments

5.1 Dataset

Type Data: Our experiments utilise `AnnoMI` (Wu et al., 2022, 2023), which is available under Public Domain License. It consists of 133 MI conversations in 10 different topics in English which are transcribed from YouTube demonstration videos and annotated by experts from the MI network[2]. The dataset creators conducted a post-annotation survey, whose results show that the majority of annotators agree that the videos do reflect real-world MI sessions even though the dialogues are scripted for educational purposes.

Each client utterance in `AnnoMI` is assigned one type of motivation language (i.e, "change", "sustain", or "neutral"). The dataset is heavily imbalanced: the number of "change", "sustain", and "neutral" utterances are 1,178, 546, and 3,093 respectively. We randomly selected 600 utterances to serve as the test set. From the remaining utterances, the `fast voke-k` algorithm (Su et al., 2023) was employed to obtain 300 most diverse samples for the validation set and k samples for the training set, with $k \in \{50, 100, 200, 300, 3600\}$.

Strength Data: MI Skill Code (MISC)[3] is a behavioral coding system, developed to assess MI sessions. The number of samples taken from MISC 2.0 and 2.1 (Miller et al., 2003; Amrhein et al., 2008) is 178, which is further split into 128 and 50 samples to serve as the training and validation sets respectively. Mosbach et al. (2023) propose that 50 samples as the validation set are sufficient to select the best performing checkpoints. The test set is taken from the `type` task. Recently, researchers have investigated GPT models in data annotation tasks (He et al., 2023; Ding et al., 2023; Huang et al., 2023; Gilardi et al., 2023), suggesting that they can serve as excellent assistants to annotators during the annotation process by providing detailed explanations, potentially replacing crowdsourced workers. For the annotation of the test set, using the MISC guidelines and the explanations generated by GPT-3.5, we manually assign a strength value (i.e., "high", "medium", or "low") to each client turn. Since textual information alone is insufficient, we consult the videos to assist with the annotation process. Details on the annotation is provided in Appendix A.

[2] https://motivationalinterviewing.org/
[3] https://casaa.unm.edu/tools/misc.html

5.2 Experimental Setup

Baselines: We employ two baselines: (1) 0-shot ICL settings with `Flan-T5-XXL`[4] (Chung et al., 2022) and `GPT-3.5-turbo`[5] and (2) traditional FT with `RoBERTa-large`[6] (Liu et al., 2019). RoBERTa is trained until convergence with the default learning rate of 1e-5. As RoBERTa is among the most popular Transformer-based encoder-type models, we use it as a baseline to measure the performance gain obtained on the LLMs.

ICL settings: Due to restrictions in context length of `Flan-T5-XXL`, only one example is included as demonstration. For a fair comparison, `GPT-3.5-turbo` also learns in 1-shot setting. Retrieval-based method is utilised (Su et al., 2023) for demonstration selection: the dialogue in the training set which is most similar to the test dialogue is chosen as context.

IFT settings: We fine-tune `Flan-T5-XXL` with instructions as depicted in Figure 2. We use Weights and Bias[7] to search for the best learning rate and finally settle on 3e-4 for all models. Further details about the training and hyper-parameter selection are given in Appendix C.

Evaluation metrics: We employ accuracy and F1 score macro-averaged calculated by `scikit-learn` (Pedregosa et al., 2011). In the multitask settings, the predictions for each task are extracted from the model outputs using regular expressions. Results are reported on the test set, using models with the best F1 scores on the validation sets during training.

6 Experimental Results

6.1 Single-Task Learning: `Type`

Table 1 illustrates the results of the `type` task. The performance of Flan-T5 with 0-shot corresponds to those of RoBERTa and Flan-T5 when trained on

[4] https://huggingface.co/google/flan-t5-xxl
[5] https://platform.openai.com/docs/models/
[6] https://huggingface.co/roberta-large
[7] https://wandb.ai/

	50	100	200	300	3600
gpt-1s-icl	0.56	0.57	0.57	0.58	0.59
flant5-1s-icl	**0.60**	**0.60**	**0.61**	**0.61**	0.63
flant5-ift	0.36	0.47	0.60	**0.61**	**0.74**
roberta-ft	0.36	0.46	0.53	0.55	0.61

Table 1: F1 scores of the `type` task on the test set with different training samples.

	instructions
all	Options are "change" (motivation towards behaviour change), "neutral" (neutral attitude or not enough information), or "sustain" (resistance against behaviour change).
simplified	Options are "change", "sustain", or "neutral".

Figure 3: Ablation studies of output space specified in the instruction for `type` task. **all** consists of the *label list* (in green) and the *label description* (in yellow), whereas **simplified** instructions have *label list* only.

100 samples, whereas GPT-3.5 with 0-shot yields the same score as RoBERTa trained on 200 samples. Interestingly, both GPT-3.5 and Flan-T5 with 1-shot ICL exhibit a similar behaviour: their performances stay relatively consistent regardless of the number of samples that can be selected as demonstrations.

Hallucinated Output Labels: Framed as a generation problem, instruction-tuned models still can produce ill-formed candidates despite being trained on desirable labels: Flan-T5 trained on 50 and 100 samples generates such outputs. In contrast, ICL even with zero shot does not suffer from this issue. After the hallucinated labels are replaced with "neutral"[8], F1 scores for Flan-T5 with 50 and 100 training data jump from 0.36 and 0.47 to 0.59 and 0.62 respectively. Consequently, the new score obtained on 100 samples completely outperforms other ICL variants.

Unexpected Results: Observably, both ICL and IFT obtain little performance gain as the training data size increases. The reason could be because our training samples are not randomly selected. As explained in Section 5.1, the *fast vote-k* algorithm by Su et al. (2023) is employed to pick the most diverse samples for both training and validation sets. Their paper shows that ICL performance with this approach is quite stable once we have enough high-quality data. Hypothetically, the LLMs might have already obtained the most important features from the diverse dataset unless the models are trained on a full dataset with thousands of examples.

Ablation with Output Space Labels: With IFT, specifying output space labels proves crucial for classification tasks (Kung and Peng, 2023; Yin et al., 2023). In addition to the *label list*, one can add the *label description* to give extra information about the meaning of the labels. Figure 3 illustrates two conditions **all** and **simplified** of our ablation studies. In contrast to Kung and Peng (2023) who find that two conditions exhibit similar effect, we observe that **all** condition (i.e., having both label list and label description) outperforms **simplified** with varying data size. Our results are similar to those of Yin et al. (2023): the authors hypothesise that label description might be used to disambiguate labels with the same name but used in different tasks.

Error Analysis: Classification reports on individual labels reveal that both IFT and ICL struggle on "sustain": F1 scores are below 0.4 and 0.5 respectively. Additionally, IFT outperforms ICL due to its capabilities in predicting "neutral" labels: more than half of the labels belong to this class. ICL, though, still predicts more than twice "sustain" labels compared to IFT.

The MI `type` labels indicate the direction of motivation towards a certain behaviour change. They are, however, unable to capture (1) complete refusals from the clients to talk about their problems, and (2) strategies employed to avoid discussing difficult topics (Martin et al., 2020). In the MISC guidelines, the latter can be coded as "change" because the clients tend to agree just to end the conversations. In contrast, the former behaviours are coded as "neutral". An inspection of the model predictions reveals that several instances of refusal and resistance to an undefined target behaviour change are predicted as "sustain". This explains models' poor performance on the "sustain" class, especially ICL. We leave it for future works on how it might influence the design of the MI agents.

6.2 Single-Task Learning: `Strength`

Results for the `strength` analysis are reported in Table 2. Surprisingly, retrieval-based ICL with 1-shot fares quite poorly, even worse than fine-tuned RoBERTa. GPT-3.5 suffers a drop in performance when shifting from 0-shot to 1-shot. Zhao et al. (2021) attribute it to majority label bias in which

[8]The label "neutral" is chosen due to (1) it is the most common labels in the dataset, and (2) in the later mapping in Section 7, "neutral" is mapped to the zero score, and thus, will not change the proposed motivational level.

	Accuracy	F1
gpt 0-shot	0.46	0.39
gpt 1-shot	0.40	0.34
flant5 0-shot	0.41	0.39
flant5 1-shot	0.47	0.45
flant5 ift	**0.72**	**0.68**
roberta ft	0.59	0.53

Table 2: Accuracy and F1 scores for the `strength` task.

GPT-3 merely reuses the class of the only example in the instructions. However, we observe no such phenomenon. In fact, when calculating the overlap between model' predictions and in-context example's labels, the overlap occurs in 63 samples out of 600: GPT-3.5 does not simply repeat the label of the example in roughly 90% of the times. The difference in our findings and those of Zhao et al. (2021) might be due to an upgrade from GPT-3 to GPT-3.5. Our results suggest that fine-tuning is still more stable and less sensitive than ICL.

Ablation with Dialogue Context: In an attempt to understand the poor performance of ICL, we conduct ablation studies using: (1) only client turns as context instead of both therapist and client utterances to match the training samples, and (2) GPT-3.5 with multiple shots using retrieval-based ICL. The results show that a longer context history for the test sample helps improve the ICL performance despite some mismatch between the format of test samples and that of the demonstrated examples.

Interestingly, increasing the number of demonstrated examples does not always lead to higher scores. We revisit the majority label bias claimed by Zhao et al. (2021). Intuitively, the argument for retrieval-based ICL is to exploit this bias by retrieving the most similar examples to the test sample, and thus reusing the majority label. Yet, we find no such bias. An examination of the predictions by GPT-3.5 3-shot and 4-shot reveals many cases where all retrieved examples belong to one class (e.g., low) but the prediction is of another (e.g., medium or high). In fact, by using the majority label of the retrieved examples as prediction increases accuracy from 0.42 to 0.43.

Error Analysis: Analysing the confusion matrices, all the models struggle with the "high" class, especially with the ICL variants. Nevertheless, their poor performance comes from overgenerating the "low" class. Except for Flan-T5 with IFT, around half of the "low" predictions by GPT-3.5 and Flan-T5 with ICL variants and RoBERTa belong to the "medium" class instead. One possible reason is because of a large number of utterances consists of multiple sentences, making the `strength` levels fluctuate from one side to another. Rationales by GPT-3.5 further imply confusions between the certainty level as a manner of expressing one's belief and their knowledge: one can be certain about their uncertainty (i.e., *"I have absolutely no idea about it."*). Incorporating other signals from speech and/or facial expressions would be beneficial to the recognition.

6.3 Multitask Learning

	`type`		`strength`	
	Acc.	F1	Acc.	F1
gpt 0-shot	**0.53**	**0.49**	0.41	0.38
gpt 1-shot	0.50	0.43	0.50	0.48
flant5 1-shot	0.43	0.34	0.40	0.39
flant5 ift	0.32	0.29	**0.67**	**0.66**

Table 3: Results on multitask learning.

Inspired by Varia et al. (2023), we experiment with multitask learning where the models should learn to predict the two tasks simultaneously. Regular expressions are employed to get the predictions and replace the ill-formed labels with either "neutral" or "medium" depending on the task. Table 3 reports the results. These experiments use the `strength` dataset (i.e., training and validation sizes are 128 and 50 respectively). Even using only 50 samples, both ICL and IFT achieve F1 scores higher than 0.6 while with 128 samples in multitask learning (MTL), 0.49 is the best F1 score. IFT performs surprisingly poorly.

Mixing More Data: We try to mix more samples (i.e., 100, 200, and 300) from the `type` dataset to investigate whether adding data improves performance. However, a higher number of mixed data results in more ill-formed outputs for the `strength` task. Consequently, performance on the `type` task increases while that on the `strength` task decreases. Our results contradict those of Varia et al. (2023): STL overall outperforms MTL.

Error Analysis: An examination of label distribution on both training and test sets reveals that three variants of "neutral" (i.e., neutral high, neutral medium, neutral low) make up of nearly 60%

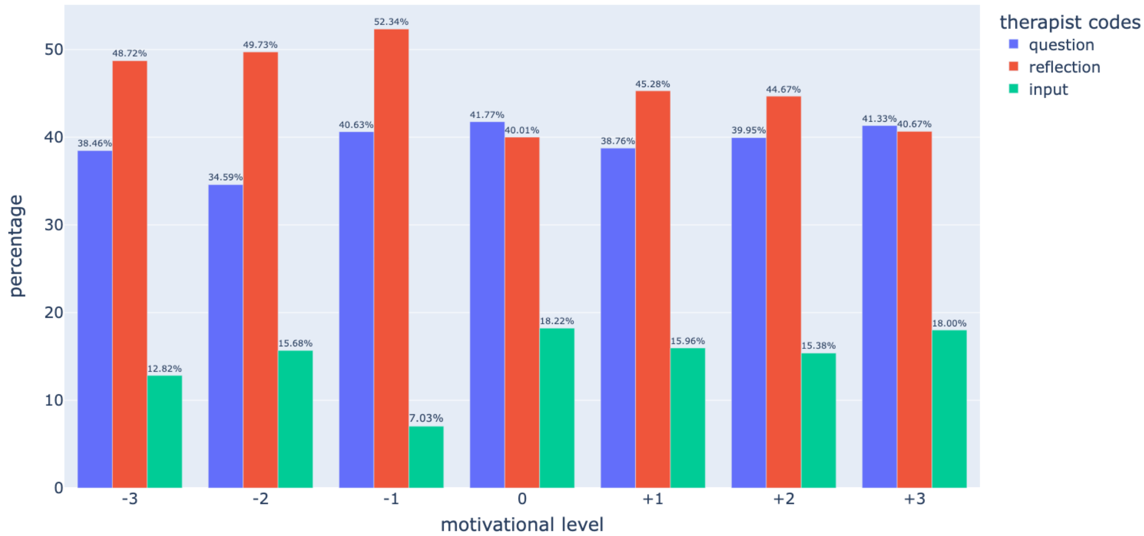

Figure 4: Distribution of next-turn therapist behaviours given clients' motivational level in the current turn.

in the test set. Yet, no "neutral" samples exist in the training set, which explains why the models are unable to learn properly. Appendix B shows the distribution of all 9 labels in the dataset. Nevertheless, ICL appears to be less effected by this imbalance training data: both Flan-T5 and GPT-3.5 struggle more to learn "change" or "sustain". As for the `strength` task, the performance in MTL, though slightly lower, is still comparable to STL.

On the mixed data, the similarity in the labels of three instructions confuses the learning: in some cases, the correct label is "neutral" but in other cases, it has to be "neutral high", "neutral medium" or "neutral low". Due to the overwhelmed "neutral" class, the models appear to struggle to generate the other multi-word labels. IFT might be unsuitable for labels with multiple words. Schick and Schütze (2021) claim that Pattern-Exploiting Training, a stricter variant of IFT, can only work when the labels correspond to a single token.

7 Application to Psychotherapist Agents

An MI session consists of 4 stages: engaging, focusing, evoking, and planning (Miller and Rollnick, 2023). To control the conversational flow, Park et al. (2019) define a fixed sequence of behaviours for each stage. We hypothesise that MI sessions can be automated using the detected motivational language. We demonstrate how it can inform the psychotherapist agents' next moves using `AnnoMI` data and MI literature.

Therapists' Strategies: Using the best model from our experiments, we obtain `strength` labels and calculate the motivational levels for all client utterances in the `AnnoMI` dataset (Wu et al., 2023). We employ a scale from -3 to +3, similar to Gaume et al. (2016). All "neutral" `type` equals to 0. `Strength` labels "high", "medium", and "low" are given levels of 3, 2, and 1 respectively while `type` labels "change" and "sustain" indicate the positive and negative directions. For example, "change-high" is mapped to +3 while "sustain-low" is -1.

Figure 4 illustrates the distribution of the next-turn therapist behaviour codes given the current clients' motivational level. We count all the possible codes in one utterance. Since the "other" behaviour consists mostly of facilitating languages (e.g., *Mm-hmm*, *Uh-uh*, *Yeah*) and greetings, we only compute the percentages for "Question", "Reflection", and "Input". Observably, "reflection" is employed frequently throughout the sessions, nearly 50% of the times when the clients show resistance (i.e. levels of -3, -2, and -1). More "input" and "question" behaviours are displayed when the clients are more ready to change.

Balancing Objectives in Therapies: In psychotherapy, the therapists need to balance two conflicting goals: building therapeutic rapport with the clients and pushing them towards task completion. Zhang and Danescu-Niculescu-Mizil (2020) argue that each therapist utterance aims to move backward from or forwards towards the goal. Our hypothesis is that the MI behaviour codes can also be classified into rapport-building (i.e., reflections, focusing questions) and goal-pursuing strategies (i.e., evoking questions, inputs). Reflections are restatements of the clients' thoughts and feelings, expressing the therapists' understandings of their inner worlds. Inputs include a wide range of sub-behaviours such as providing information, giving advice, offering options, and setting goals. Focusing questions explore their perspectives, goals, and values while evoking questions aim to elicit their motivation to change. Though no distinction between focusing and evoking questions is made in `AnnoMI`, our belief is that this distinction would be beneficial to the MI agents.

Framework of the MI Agents: With insights from the literature and the data, we would like to propose a computational framework of the MI agents. In an attempt to investigate who might benefit from MI and who not, Forman et al. (2022) measure the differences in clients' language early in the session and discover that those whose language reflects ambivalence (i.e. low motivated), benefit more from MI. In contrast, MI appears to be counterproductive for those who already show a readiness to change, suggesting that MI strategies should be adapted appropriately to the clients' presenting levels of motivation.

The transtheoretical model of stages of health behaviour change (Prochaska and Velicer, 1997) hypothesises the first 3 stages are precontemplation ("not ready", or resistance), contemplation ("getting ready", or ambivalence) and preparation ("ready", or motivation). We hypothesise that after several interactions, if the detected motivational levels are mainly -3 and -2, the client is in precontemplation stage. If they are -1 or +1, it is contemplation. And if the levels are +2 or +3, the stage is preparation.

Once in the preparation stage, the MI agents should employ mainly goal-pursuing behaviours or switch to another more goal-oriented technique such as CBT (Westra and Norouzian, 2018), while occasionally utilising reflections and focusing questions when the users display low motivation to maintain the therapeutic alliance. Besides rapport-building behaviours, the agents can be programmed to emphasising the users' autonomy, coded as "input" in the `AnnoMI`, when the users display signs of the precontemplation stage. For example, *"that is your choice. I can't make those choices for you, it is something that you decide to do."* and *"you're the boss. It's up to you what you want to do with you about your own health."*. The stage should help inform the agents' strategies if the detected level is 0. Our belief is that this information can be leveraged to design the instructions to train the agents to exhibit more MI-adherence interactions.

Clinical Implications and Potential Applications: Training using therapy data only might be insufficient to create psychotherapy agents as revealed by Das et al. (2022): their agent shows a lack of therapeutic behaviours and merely gives general advice. We believe that by monitoring the clients' motivational levels, the agents can act in a more proactive manner following the MI spirit. For example, giving advice and setting goals when the clients are ready enough and supporting them when resistance arises. As MI is a well-regarded, evidence-based, and widely used approach for behaviour change, the MI-aware agents can reduce the system burden and facilitate treatment delivery with lower costs to reach a wider range of users.

8 Conclusion and Future Works

Works in psychology suggest that monitoring client motivational language is an essential skill to deliver successful therapies. Our belief is that a motivation-aware system would have implications for the development of personalised healthcare agents. Our experiments employ LLMs, and compare ICL with IFT on varying training data sizes. Our findings indicate that both can perform in few-shot settings and be sensitive to the instructions. Still, we observe that with ICL, the predictions can change when adding something totally unrelated to the task itself (i.e., requesting a certain format of the output). IFT is more stable; however, it suffers from generating ill-formed outputs when trained with a small number of samples. With the obtained labels, we devise a computational framework for MI agents based on the users' motivation at stage and utterance levels. Insights from `AnnoMI` data and MI literature suggest that the agents should exhibit mainly rapport-building behaviours when facing resistance and ambivalence. Once the users indicate a strong willingness to change, goal-pursuing strategies are preferred. Rapport-building behaviours are employed occasionally, when appropriate. In future works, we would like to investigate how to incorporate such information into the design of instructions to generate therapeutic interactions.

9 Limitations

MI practice: Our paper is inspired by the MI approach to behaviour change. We try to give general readers a brief overview of the MI spirit, enough to understand the rationales behind the proposed framework. A comprehensive review of MI and/or CBT and their validity is, however, out of scope of the paper. We acknowledge that there exist different applications for MI and thus, the language should be contextualised for different clinical situations. Nevertheless, the paper aims to show whether and how the motivational language can be utilised in general to direct the behaviours of the agents theoretically and experimentally without focusing on a particular clinical situation.

Dataset: As the conversations in the `AnnoMI` dataset (Wu et al., 2023) are role-play MI videos used for educational purposes, they might not reflect the real therapies in which the clients can behave in a more unexpected manner, especially the way they show their resistance. The language in use is English, and thus, might be unsuitable for other languages. Furthermore, in practice, the therapists might use a mixture of different approaches, not just MI. All these limitations can effect generalisation to real-world applications. However, real MI therapies are scarce. The `AnnoMI` demonstrations have been judged by MI experts to reflect real MI sessions. As our main purpose is to create an MI agent, we would argue that high-quality MI demonstrations should help create agents faithful to MI practice more than real therapies with mixed approaches.

Annotation labels: The MISC guidelines suggest a fine-grained annotation based on sentences or phrases. However, the labels are assigned to turns, not sentences. A turn can consist of multiple sentences but can also be unfinished sentences or words (e.g., *"-forms."*). Therefore, these samples contain no information to help the models make predictions. Even though classifying turns might be desirable for speech systems, it might potentially teach the models inappropriate features for classification tasks.

Additionally, we observe many samples consisting of multiple sentences whose direction and strength of motivation can move from one end to another as the clients speak. This explains partly the low inter-annotator agreement on `AnnoMI`. Similarly, in the `strength` task, many utterances consist of multiple sentences whose certainty levels can go from one extreme to the other. This poses as a huge challenge for the annotation process.

Choice of models and prompts: As for the model choice, we experimented with several models before settling on Flan-T5. Despite not being the SOTA model in all tasks, the Flan-T5 family is suitable for classification tasks. Similarly, other Parameter-efficient fine-tuning techniques are investigated but due to their poor performance, we did not include all the results except for LoRA. As for the prompt formats, we could have tested and compared different prompts. However, since the focus is to demonstrate how the detected motivation can be utilised to automate the MI agents' behaviours, we simply took the insights from Yin et al. (2023) in designing the optimal prompts.

Multimodal systems: We only utilise textual features to make predictions. Prior works suggest incorporating visual features (i.e., facial expressions) for the `type` task (Nakano et al., 2022) as the client might hint their resistance by keeping silent and/or looking away. As for the `strength` task, experiments in linguistics show that acoustic features (e.g., pitch accents) convey speaker's commitment (Michelas et al., 2016). When annotating the test set, we do observe that whether the speaker is fluent or hesitates about their actions can be a signal for their certainty level.

10 Ethical Concerns

MI is a therapy originally developed to help people change their harmful behaviours such as alcoholism (Miller and Rollnick, 2023). Due to its effectiveness, MI practitioners have applied it to other fields, including those involving unethical practices such as sales or marketing[9]. We acknowledge that an MI-aware agent can be misused to target low-motivated users for motivation tricks for behaviour change that benefits the providers instead of the clients (i.e., buy more products, ask for donation against their will), just as how an MI expert can misuse the technique. Our belief is that an MI-aware agent can, however, have positive implications for the development of intelligent systems in the healthcare domain. Mental health is always a big issue in modern society. Additionally, an MI-aware agent can motivate people for positive behaviour change such as being more physically active (Olafsson et al., 2020a).

[9] https://motivationalinterviewing.org/non-ethical-practice-mi

Acknowledgement

This publication has emanated from research conducted with the financial support of Science Foundation Ireland under Grant number 18/CRT/6183 (ML-Labs) and the ADAPT SFI Research Centre for AI-Driven Digital Content Technology under Grant No. 13/RC/2106_P2. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

References

Efrat Aharonovich, Paul C. Amrhein, Adam Bisaga, Edward V. Nunes, and Deborah S. Hasin. 2008. Cognition, commitment language, and behavioral change among cocaine-dependent patients. *Psychology of Addictive Behaviors*, 22(4):557–562.

Tim Althoff, Kevin Clark, and Jure Leskovec. 2016. Large-scale Analysis of Counseling Conversations: An Application of Natural Language Processing to Mental Health. *Transactions of the Association for Computational Linguistics*, 4:463–476. Place: Cambridge, MA Publisher: MIT Press.

Paul C Amrhein, William R. Miller, Theresa Moyers, and Denise Ernst. 2008. Motivational Interviewing Skill Code (MISC) 2.1.

Paul C. Amrhein, William R. Miller, Carolina E. Yahne, Michael Palmer, and Laura Fulcher. 2003. Client commitment language during motivational interviewing predicts drug use outcomes. *Journal of Consulting and Clinical Psychology*, 71(5):862–878.

Chanuwas Aswamenakul, Lixing Liu, Kate B. Carey, Joshua Woolley, Stefan Scherer, and Brian Borsari. 2018. Multimodal Analysis of Client Behavioral Change Coding in Motivational Interviewing. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, pages 356–360, Boulder CO USA. ACM.

Kira Boulat and Didier Maillat. 2023. Strength is relevant: experimental evidence of strength as a marker of commitment. *Frontiers in Communication*, 8:1176845.

Andrew Brown, Ash Tanuj Kumar, Osnat Melamed, Imtihan Ahmed, Yu Hao Wang, Arnaud Deza, Marc Morcos, Leon Zhu, Marta Maslej, Nadia Minian, Vidya Sujaya, Jodi Wolff, Olivia Doggett, Mathew Iantorno, Matt Ratto, Peter Selby, and Jonathan Rose. 2023. A Motivational Interviewing Chatbot With Generative Reflections for Increasing Readiness to Quit Smoking: Iterative Development Study. *JMIR Mental Health*, 10:e49132.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. ArXiv:2005.14165 [cs].

Samadhi Deva Campbell, Simon Justin Adamson, and Janet Deborah Carter. 2010. Client Language During Motivational Enhancement Therapy and Alcohol Use Outcome. *Behavioural and Cognitive Psychotherapy*, 38(4):399–415.

Jie Cao, Michael Tanana, Zac Imel, Eric Poitras, David Atkins, and Vivek Srikumar. 2019. Observing Dialogue in Therapy: Categorizing and Forecasting Behavioral Codes. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5599–5611, Florence, Italy. Association for Computational Linguistics.

Zhuohao Chen, Nikolaos Flemotomos, Victor Ardulov, Torrey A. Creed, Zac E. Imel, David C. Atkins, and Shrikanth Narayanan. 2021. Feature Fusion Strategies for End-to-End Evaluation of Cognitive Behavior Therapy Sessions. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 1836–1839, Mexico. IEEE.

Young Min Cho, Sunny Rai, Lyle Ungar, João Sedoc, and Sharath Chandra Guntuku. 2023. An Integrative Survey on Mental Health Conversational Agents to Bridge Computer Science and Medical Perspectives. ArXiv:2310.17017 [cs].

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling Instruction-Finetuned Language Models. ArXiv:2210.11416 [cs].

Dominic A. Clark. 1990. Verbal uncertainty expressions: A critical review of two decades of research. *Current Psychology*, 9(3):203–235.

Mark Conner and Paul Norman. 2022. Understanding the intention-behavior gap: The role of intention strength. *Frontiers in Psychology*, 13:923464.

Avisha Das, Salih Selek, Alia R. Warner, Xu Zuo, Yan Hu, Vipina Kuttichi Keloth, Jianfu Li, W. Jim Zheng, and Hua Xu. 2022. Conversational Bots for Psychotherapy: A Study of Generative Transformer Models Using Domain-specific Dialogues. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 285–297, Dublin, Ireland. Association for Computational Linguistics.

Marie-Catherine de MARNEFFE, Mandy Simons, and Judith Tonhauser. 2019. The CommitmentBank: Investigating projection in naturally occurring discourse. *Proceedings of Sinn und Bedeutung*, 2:107–124.

Mona Diab, Lori Levin, Teruko Mitamura, Owen Rambow, Vinodkumar Prabhakaran, and Weiwei Guo. 2009. Committed Belief Annotation and Tagging. In *Proceedings of the Third Linguistic Annotation Workshop (LAW III)*, pages 68–73, Suntec, Singapore. Association for Computational Linguistics.

Bosheng Ding, Chengwei Qin, Linlin Liu, Yew Ken Chia, Boyang Li, Shafiq Joty, and Lidong Bing. 2023. Is GPT-3 a Good Data Annotator? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11173–11195, Toronto, Canada. Association for Computational Linguistics.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, Lei Li, and Zhifang Sui. 2023. A Survey on In-context Learning. ArXiv:2301.00234 [cs].

M. P. Ewbank, R. Cummins, V. Tablan, A. Catarino, S. Buchholz, and A. D. Blackwell. 2021. Understanding the relationship between patient language and outcomes in internet-enabled cognitive behavioural therapy: A deep learning approach to automatic coding of session transcripts. *Psychotherapy Research*, 31(3):300–312.

David P. Forman, Theresa B. Moyers, and Jon M. Houck. 2022. What can clients tell us about whether to use motivational interviewing? An analysis of early-session ambivalent language. *Journal of Substance Abuse Treatment*, 132:108642.

Joana Galvão Gomes Da Silva, David J Kavanagh, Tony Belpaeme, Lloyd Taylor, Konna Beeson, and Jackie Andrade. 2018. Experiences of a Motivational Interview Delivered by a Robot: Qualitative Study. *Journal of Medical Internet Research*, 20(5):e116.

Jacques Gaume, Molly Magill, Nadine R. Mastroleo, Richard Longabaugh, Nicolas Bertholet, Gerhard Gmel, and Jean-Bernard Daeppen. 2016. Change Talk During Brief Motivational Intervention With Young Adult Males: Strength Matters. *Journal of Substance Abuse Treatment*, 65:58–65.

Tirthankar Ghosal, Kamal Kaushik Varanasi, and Valia Kordoni. 2022. HedgePeer: a dataset for uncertainty detection in peer reviews. In *Proceedings of the 22nd ACM/IEEE Joint Conference on Digital Libraries*, pages 1–5, Cologne Germany. ACM.

James Gibson, David C. Atkins, Torrey Creed, Zac Imel, Panayiotis Georgiou, and Shrikanth Narayanan. 2022. Multi-label Multi-task Deep Learning for Behavioral Coding. *IEEE transactions on affective computing*, 13(1):508–518.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. ChatGPT Outperforms Crowd-Workers for Text-Annotation Tasks. ArXiv:2303.15056 [cs].

Himanshu Gupta, Saurabh Arjun Sawant, Swaroop Mishra, Mutsumi Nakamura, Arindam Mitra, Santosh Mashetty, and Chitta Baral. 2023. Instruction Tuned Models are Quick Learners. ArXiv:2306.05539 [cs].

Xingwei He, Zhenghao Lin, Yeyun Gong, A.-Long Jin, Hang Zhang, Chen Lin, Jian Jiao, Siu Ming Yiu, Nan Duan, and Weizhu Chen. 2023. AnnoLLM: Making Large Language Models to Be Better Crowdsourced Annotators. ArXiv:2303.16854 [cs].

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-Rank Adaptation of Large Language Models.

Fan Huang, Haewoon Kwak, and Jisun An. 2023. Is ChatGPT better than Human Annotators? Potential and Limitations of ChatGPT in Explaining Implicit Hate Speech. In *Companion Proceedings of the ACM Web Conference 2023*, pages 294–297, Austin TX USA. ACM.

Po-Nien Kung and Nanyun Peng. 2023. Do Models Really Learn to Follow Instructions? An Empirical Study of Instruction Tuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1317–1328, Toronto, Canada. Association for Computational Linguistics.

Anqi Li, Lizhi Ma, Yaling Mei, Hongliang He, Shuai Zhang, Huachuan Qiu, and Zhenzhong Lan. 2023. Understanding Client Reactions in Online Mental Health Counseling. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10358–10376, Toronto, Canada. Association for Computational Linguistics.

Vladislav Lialin, Vijeta Deshpande, and Anna Rumshisky. 2023. Scaling Down to Scale Up: A Guide to Parameter-Efficient Fine-Tuning. ArXiv:2303.15647 [cs].

Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin Raffel. 2022. Few-Shot Parameter-Efficient Fine-Tuning is Better and Cheaper than In-Context Learning.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2021. What Makes Good In-Context Examples for GPT-3? ArXiv:2101.06804 [cs].

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. ArXiv:1907.11692 [cs].

Robert Logan IV, Ivana Balazevic, Eric Wallace, Fabio Petroni, Sameer Singh, and Sebastian Riedel. 2022. Cutting Down on Prompts and Parameters: Simple Few-Shot Learning with Language Models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2824–2835, Dublin, Ireland. Association for Computational Linguistics.

Diana R. Lombardi, Melissa L. Button, and Henny A. Westra. 2014. Measuring Motivation: Change Talk and Counter-Change Talk in Cognitive Behavioral Therapy for Generalized Anxiety. *Cognitive Behaviour Therapy*, 43(1):12–21.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V. Le, Barret Zoph, Jason Wei, and Adam Roberts. 2023. The Flan Collection: Designing Data and Methods for Effective Instruction Tuning. In *Proceedings of the 40th International Conference on Machine Learning*, pages 22631–22648. PMLR. ISSN: 2640-3498.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically Ordered Prompts and Where to Find Them: Overcoming Few-Shot Prompt Order Sensitivity. ArXiv:2104.08786 [cs].

Vanessa Martin, Tasmie Sarker, Emily Slusarek, Mary A. McCarthy, John Granton, Adrienne Tan, and Christopher Lo. 2020. Conversational avoidance during existential interviews with patients with progressive illness. *Psychology, Health & Medicine*, 25(9):1073–1082.

Selina Meyer. 2021. Natural Language Stage of Change Modelling for "Motivationally-driven" Weight Loss Support. In *Proceedings of the 2021 International Conference on Multimodal Interaction*, pages 807–811, Montréal QC Canada. ACM.

Amandine Michelas, Cristel Portes, and Maud Champagne-Lavau. 2016. When pitch Accents Encode Speaker Commitment: Evidence from French Intonation. *Language and Speech*, 59(2):266–293.

William R. Miller, Theresa Moyers, Denise Ernst, and Paul C. Amrhein. 2003. Motivational Interviewing Skill Code (MISC) 2.0.

William R. Miller and Stephen Rollnick. 2023. *Motivational interviewing: helping people change and grow*, fourth edition edition. Applications of motivational interviewing. The Guilford Press, New York.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the Role of Demonstrations: What Makes In-Context Learning Work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Marius Mosbach, Tiago Pimentel, Shauli Ravfogel, Dietrich Klakow, and Yanai Elazar. 2023. Few-shot Fine-tuning vs. In-context Learning: A Fair Comparison and Evaluation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12284–12314, Toronto, Canada. Association for Computational Linguistics.

Yukiko I. Nakano, Eri Hirose, Tatsuya Sakato, Shogo Okada, and Jean-Claude Martin. 2022. Detecting Change Talk in Motivational Interviewing using Verbal and Facial Information. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION*, pages 5–14, Bengaluru India. ACM.

Stefan Olafsson, Teresa K. O'Leary, and Timothy W. Bickmore. 2020a. Motivating Health Behavior Change with Humorous Virtual Agents. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*, pages 1–8, Virtual Event Scotland UK. ACM.

Stefan Olafsson, Byron Wallace, and Timothy Bickmore. 2020b. Towards a Computational Framework for Automating Substance Use Counseling with Virtual Agents. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '20, pages 966–974, Richland, SC. International Foundation for Autonomous Agents and Multiagent Systems.

SoHyun Park, Jeewon Choi, Sungwoo Lee, Changhoon Oh, Changdai Kim, Soohyun La, Joonhwan Lee, and Bongwon Suh. 2019. Designing a Chatbot for a Brief Motivational Interview on Stress Management: Qualitative Case Study. *Journal of Medical Internet Research*, 21(4):e12231.

Havisha Pedamallu, Matthew J Ehrhardt, Julia Maki, April Idalski Carcone, Melissa M Hudson, and Erika A Waters. 2022. Technology-Delivered Adaptations of Motivational Interviewing for the Prevention and Management of Chronic Diseases: Scoping Review. *Journal of Medical Internet Research*, 24(8):e35283.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Lauren E. Poulin, Melissa L. Button, Henny A. Westra, Michael J. Constantino, and Martin M. Antony. 2019. The predictive capacity of self-reported motivation vs. early observed motivational language in cognitive behavioural therapy for generalized anxiety disorder. *Cognitive Behaviour Therapy*, 48(5):369–384.

Vinodkumar Prabhakaran, Tomas By, Julia Hirschberg, Owen Rambow, Samira Shaikh, Tomek Strzalkowski, Jennifer Tracey, Michael Arrigo, Rupayan Basu, Micah Clark, Adam Dalton, Mona Diab, Louise Guthrie, Anna Prokofieva, Stephanie Strassel, Gregory Werner, Yorick Wilks, and Janyce Wiebe. 2015. A New Dataset and Evaluation for Belief/Factuality. In *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics*, pages 82–91, Denver, Colorado. Association for Computational Linguistics.

Vinodkumar Prabhakaran, Premkumar Ganeshkumar, and Owen Rambow. 2018. Author Commitment and Social Power: Automatic Belief Tagging to Infer the Social Context of Interactions. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1057–1068, New Orleans, Louisiana. Association for Computational Linguistics.

James O. Prochaska and Wayne F. Velicer. 1997. The Transtheoretical Model of Health Behavior Change. *American Journal of Health Promotion*, 12(1):38–48.

Valentina Pyatkin, Shoval Sadde, Aynat Rubinstein, Paul Portner, and Reut Tsarfaty. 2021. The Possible, the Plausible, and the Desirable: Event-Based Modality Detection for Language Processing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 953–965, Online. Association for Computational Linguistics.

Verónica Pérez-Rosas, Rada Mihalcea, Kenneth Resnicow, Satinder Singh, and Lawrence An. 2016. Building a Motivational Interviewing Dataset. In *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology*, pages 42–51, San Diego, CA, USA. Association for Computational Linguistics.

Verónica Pérez-Rosas, Rada Mihalcea, Kenneth Resnicow, Satinder Singh, Lawrence Ann, Kathy J. Goggin, and Delwyn Catley. 2017. Predicting Counselor Behaviors in Motivational Interviewing Encounters. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1128–1137, Valencia, Spain. Association for Computational Linguistics.

Verónica Pérez-Rosas, Xinyi Wu, Kenneth Resnicow, and Rada Mihalcea. 2019. What Makes a Good Counselor? Learning to Distinguish between High-quality and Low-quality Counseling Conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 926–935, Florence, Italy. Association for Computational Linguistics.

Yann Raphalen, Chloé Clavel, and Justine Cassell. 2022. "You might think about slightly revising the title": Identifying Hedges in Peer-tutoring Interactions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2160–2174, Dublin, Ireland. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21:1–67.

Ryan E Rhodes, Amy Cox, and Reza Sayar. 2022. What Predicts the Physical Activity Intention–Behavior Gap? A Systematic Review. *Annals of Behavioral Medicine*, 56(1):1–20.

Rachel Rudinger, Aaron Steven White, and Benjamin Van Durme. 2018. Neural Models of Factuality. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 731–744, New Orleans, Louisiana. Association for Computational Linguistics.

Timo Schick and Hinrich Schütze. 2021. It's Not Just Size That Matters: Small Language Models Are Also Few-Shot Learners. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2339–2352, Online. Association for Computational Linguistics.

Timo Schick and Hinrich Schütze. 2022. True Few-Shot Learning with Prompts—A Real-World Perspective. *Transactions of the Association for Computational Linguistics*, 10:716–731. Place: Cambridge, MA Publisher: MIT Press.

Noam Shazeer and Mitchell Stern. 2018. Adafactor: Adaptive learning rates with sublinear memory cost. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4596–4604. PMLR.

Rebecca M. Shingleton and Tibor P. Palfai. 2016. Technology-delivered adaptations of motivational interviewing for health-related behaviors: A systematic review of the current research. *Patient Education and Counseling*, 99(1):17–35.

Karan Singla, Zhuohao Chen, David Atkins, and Shrikanth Narayanan. 2020. Towards end-2-end learning for predicting behavior codes from spoken utterances in psychotherapy conversations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3797–3803, Online. Association for Computational Linguistics.

Hongjin Su, Jungo Kasai, Chen Henry Wu, Weijia Shi, Tianlu Wang, Jiayi Xin, Rui Zhang, Mari Ostendorf, Luke Zettlemoyer, Noah A. Smith, and Tao Yu. 2023. Selective Annotation Makes Language Models Better Few-Shot Learners. International Conference on Learning Representations. ArXiv:2209.01975 [cs].

Leili Tavabi, Kalin Stefanov, Larry Zhang, Brian Borsari, Joshua D. Woolley, Stefan Scherer, and Mohammad Soleymani. 2020. Multimodal Automatic Coding of Client Behavior in Motivational Interviewing. In *Proceedings of the 2020 International Conference on Multimodal Interaction*, pages 406–413, Virtual Event Netherlands. ACM.

Leili Tavabi, Trang Tran, Kalin Stefanov, Brian Borsari, Joshua Woolley, Stefan Scherer, and Mohammad Soleymani. 2021. Analysis of behavior classification in motivational interviewing. In *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*, pages 110–115, Online. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. ArXiv:2307.09288 [cs].

Siddharth Varia, Shuai Wang, Kishaloy Halder, Robert Vacareanu, Miguel Ballesteros, Yassine Benajiba, Neha Anna John, Rishita Anubhai, Smaranda Muresan, and Dan Roth. 2023. Instruction Tuning for Few-Shot Aspect-Based Sentiment Analysis. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 19–27, Toronto, Canada. Association for Computational Linguistics.

Henny A. Westra and Nikoo Norouzian. 2018. Using Motivational Interviewing to Manage Process Markers of Ambivalence and Resistance in Cognitive Behavioral Therapy. *Cognitive Therapy and Research*, 42(2):193–203.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Zixiu Wu, Simone Balloccu, Vivek Kumar, Rim Helaoui, Diego Reforgiato Recupero, and Daniele Riboni. 2023. Creation, Analysis and Evaluation of AnnoMI, a Dataset of Expert-Annotated Counselling Dialogues. *Future Internet*, 15(3):110.

Zixiu Wu, Simone Balloccu, Vivek Kumar, Rim Helaoui, Ehud Reiter, Diego Reforgiato Recupero, and Daniele Riboni. 2022. Anno-mi: A dataset of expert-annotated counselling dialogues. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6177–6181.

Fan Yin, Jesse Vig, Philippe Laban, Shafiq Joty, Caiming Xiong, and Chien-Sheng Wu. 2023. Did You Read the Instructions? Rethinking the Effectiveness of Task Definitions in Instruction Learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3063–3079, Toronto, Canada. Association for Computational Linguistics.

Justine Zhang and Cristian Danescu-Niculescu-Mizil. 2020. Balancing Objectives in Counseling Conversations: Advancing Forwards or Looking Backwards. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5276–5289, Online. Association for Computational Linguistics.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, and Guoyin Wang. 2023. Instruction Tuning for Large Language Models: A Survey. ArXiv:2308.10792 [cs].

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate Before Use: Improving Few-shot Performance of Language Models. In *Proceedings of the 38th International Conference on Machine Learning*, pages 12697–12706. PMLR. ISSN: 2640-3498.

Kaitlyn Zhou, Dan Jurafsky, and Tatsunori Hashimoto. 2023. Navigating the Grey Area: Expressions of Overconfidence and Uncertainty in Language Models. ArXiv:2302.13439 [cs].

A Annotation Details for the `strength` Task

The annotation is conducted by the first author of this paper, who has both bachelor and master degrees in Computational Linguistics. We employ the strength rating approach similar to that of Gaume et al. (2016). Each client utterance is first assigned a strength value of "medium". If the utterance contains a **booster** word, its strength value changes to "high". On the contrary, if it has one or more **hedge** words, it receives "low" value. The "neutral" type is originally not assigned a strength value but in our experiments, we decide to annotate it similarly to the other two types for the sake of completeness. Since textual information alone is insufficient, we watch corresponding videos from `AnnoMI` dataset and use speech signals and body languages to assist with the annotation process. To generate explanations, we use GPT-3.5 with the following prompt *"How confident and certain the speaker is when saying $UTTERANCE$. Elaborate the response."*. When the explanations are in conflict with the annotator's assigned values, we further make use of explanations from ChatGPT (accessed from November 3 to 11) with the same prompt. Despite having the same base model as GPT-3.5, ChatGPT's responses are more elaborate and able to extract and explain parts of the utterances. An example is shown below:

Utterance: "I don't know. It's just like– My dad was really strict on me when I was little, and I don't want to be like that."

Strength value: low

GPT-3.5: "Based on the given statement, the speaker seems fairly confident and certain about not wanting to be like their strict father. The use of phrases like *I don't know* and *It's just like* suggests that the speaker is reflecting on their past experiences and drawing a clear connection to their current mindset. However, without further context, it is difficult to determine the exact level of confidence and certainty."

ChatGPT: "The sentence *"I don't know. It's just like– My dad was really strict on me when I was little, and I don't want to be like that"* suggests a low to moderate level of confidence and certainty. The speaker is expressing uncertainty or hesitation about a topic, and is explaining that they do not want to emulate their father's strict parenting style. The use of the phrase *"I don't know"* conveys a clear understanding of the speaker's lack of certainty or clarity, and suggests a level of confusion or indecision in their perspective. The use of the filler words *"just like"* introduces a degree of hesitancy or uncertainty, suggesting that the speaker is not completely sure of their position or is cautious about making a definitive statement. The statement *"My dad was really strict on me when I was little, and I don't want to be like that"* conveys a clear understanding of the speaker's past experiences and their desire to avoid a similar parenting style, and suggests a level of conviction in their perspective. Overall, the speaker appears to have a low to moderate level of confidence and certainty in their understanding of the topic, and is expressing a high level of hesitancy and uncertainty in their expression."

Researchers have suggested GPT models can serve as excellent data annotators, potentially replacing crowdsourced workers (He et al., 2023; Ding et al., 2023; Huang et al., 2023; Gilardi et al., 2023). Similar to Huang et al. (2023) whose findings show that lay people's decisions are affected by ChatGPT's explanations, we notice that they are extremely convincing. Therefore, annotators who employ them as assistants should be experts in the domain and understand the task clearly. For example, the following utterance *"To be frank, Mark, I honestly don't know. I mean, I honestly don't know what causes it because I've never had any reason. It came upon me from the GP and it took me back."* has a `strength` value of *high*: The speaker is highly certain about his uncertainty about the cause of the illness. ChatGPT, however, fails to detect this subtle difference.

ChatGPT: "The speaker appears to be uncertain and not confident when discussing the cause of their condition. The use of phrases like *"I honestly don't know"* and *"I've never had any reason"* indicates a lack of knowledge or understanding about the cause of their condition. The phrase *"It took me back"* suggests that the speaker was surprised or taken aback by the diagnosis, which further implies a lack of awareness or understanding about the condition prior to the diagnosis. Overall, the speaker seems to be expressing a sense of confusion and uncertainty about the cause of their condition."

B Label Distribution

Table 4 and Table 5 show the label distribution for `type` and `strength` tasks respectively.

Table 6 shows the number of labels. In the mixed

	training (full)	validation	test
change	854	79	169
neutral	2372	179	355
sustain	391	42	76

Table 4: Label distribution for `type` task.

	training	validation	test
high	50	20	122
medium	35	15	357
low	43	15	121

Table 5: Label distribution for `strength` task.

	training	validation	test
change high	24	10	36
change medium	18	8	82
change low	24	8	51
neutral high	0	0	58
neutral medium	0	0	237
neutral low	0	0	60
sustain high	26	10	28
sustain medium	17	7	38
sustain low	19	7	10

Table 6: Label distribution for multitask learning.

datasets, we add the data with **type** and **strength** labels but the amount of multitask data remains unchanged.

C Training Details

GPU usage: We use Quadro RTX 8000 (48 GB in memory) and GeForce RTX 2080 (11 GB in memory) to fine-tune Flan-T5 and RoBERTa respectively. As `Flan-T5-XXL` version is 45 GB, we load it in 8 bit for both training and inference so it can be fitted in one RTX 8000 GPU.

Flan-T5 parameters: Our hyperparameter search for the learning rate, weight decay, and the batch size is performed using Weights and Bias[10]. The learning rate is randomly sampled from the range of 5e-3 to 5e-5 in 30 trials on the `Flan-T5-XL` version (3B parameters) instead of `Flan-T5-XXL` (11B) to reduce computational costs. We settle on 3e-4 for all models. The weight decay is set to 1e-6. The batch size is 8. We fine-tune the Flan-T5 for 30 epochs using `adafactor` (Shazeer and Stern, 2018) as the optimiser. For other values, we use the default from `huggingface` (version 4.33.1) (Wolf et al., 2020) implementation. We use a fixed seed for reproducibility purposes.

LoRa parameters: There is no service to search for LoRa parameters. Therefore, we opt to use the recommended values from huggingface community: The LoRa rank, the alpha, and the dropout rate is set to 16, 32 and 0.1 respectively.

Training and Inference Time: Training time varies depending on data size. Using the full dataset of `type` task (i.e., 3k6 samples), the fine-tuning takes roughly 6 hours using early stopping. With data size ranging from 50 to 300, it takes from 30 minutes to 3 hours for 30 epochs without early stopping. Inference time on the test set using `Flan-T5-XXL` takes roughly 2.5 hours. After merging the LoRa adapters with the original weights, latency on the instruction-tuned models is almost the same as the original models.

Number of parameters: We use LoRa implemented in `peft` library[11] and train on all layers. The trained parameters for `Flan-T5-XXL` is around 71 millions, accounting for roughly 0.6% of the total 11 billion parameters. As for `RoBERTa-large`, we fine-tune all its 354 million parameter.

D Additional Evaluation Metrics

[10] https://wandb.ai/

[11] https://huggingface.co/docs/peft/index

	50		100		200		300		3600	
	Prec.	Recall	Prec.	Recall	Prec.	Recall	Prec.	Recall	Prec.	Recall
gpt-1s-icl	0.56	0.59	0.57	0.62	0.58	0.63	0.58	0.63	0.59	0.65
flant5-1s-icl	0.60	0.66	0.59	0.66	0.60	0.66	0.60	0.66	0.61	0.67
flant5-ift	0.63	0.58	0.64	0.60	0.60	0.62	0.63	0.60	0.75	0.73
roberta-ft	0.40	0.37	0.48	0.47	0.53	0.53	0.55	0.54	0.68	0.58

Table 7: Precision and Recall scores of the `type` task on the test set with different training samples after processing 2 hallucinated outputs.

	Accuracy	Precision	Recall	F1
gpt 0-shot	0.46	0.53	0.40	0.39
gpt 1-shot	0.40	0.36	0.33	0.34
flant5 0-shot	0.41	0.49	0.48	0.39
flant5 1-shot	0.47	0.53	0.49	0.45
flant5 ift	**0.72**	0.70	0.66	**0.68**
roberta ft	0.59	0.59	0.56	0.53

Table 8: Accuracy, Precision, Recall, and F1 scores for the `strength` task.

	`type`				`strength`			
	Acc.	Prec.	Recall	F1	Acc.	Prec.	Recall	F1
gpt 0-shot	**0.53**	0.58	0.52	**0.49**	0.41	0.40	0.42	0.38
gpt 1-shot	0.50	0.49	0.45	0.43	0.50	0.50	0.47	0.48
flant5 1-shot	0.43	0.36	0.35	0.34	0.40	0.45	0.45	0.39
flant5 ift	0.32	0.32	0.51	0.29	**0.67**	0.66	0.67	**0.66**

Table 9: Results on multitask learning.