“I think I could probably use Large Language Models to solve my tasks.” Detecting Client Motivational Language in Psychotherapy

Anonymous ACL submission

Abstract

Understand the client’s motivation is crucial for successful therapies. When met with resistance, the therapists are advised to soften it first instead of persisting with goal-related actions and thus risking rapport ruptures. Motivational Interviewing is such an approach: the client’s utterances are coded as they are for or against a certain behaviour change, plus their commitment strength. Yet, there are fewer than 200 samples labelled with strength value. Recently, Large Language Models (LLMs) have shown impressive capabilities in few-shot learning. We compare in-context learning (ICL) and instruction fine-tuning (IFT) with varying training size. Our experiments show that both approaches can learn under low-resourced settings and are sensitive to the instruction formatting. Still, IFT is cheaper, more stable to prompt choice, and yields better performance with more data. However, when the label distribution is heavily imbalanced that the models are unable to learn, ICL is preferred because it can exploit the LLMs more effectively.

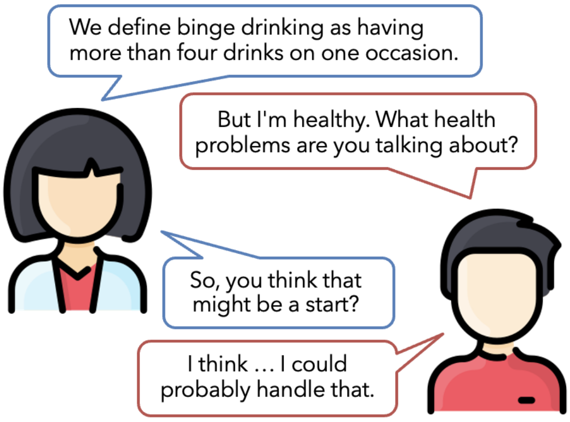

Figure 1: Two sample dialogues from `AnnoMI` (Wu et al., 2023) dataset. The upper one shows a strong resistance from the client (i.e., labelled as “sustain” for `type` and “high” for `strength` in our tasks). In the other dialogue, the client sounds willing to change though still reluctant (i.e., labelled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

1 Introduction

Resistance to social influence is a well-known phenomenon in psychology and social sciences. Cognitive Behavioral Therapy (CBT) is a psychological treatment that helps clients manage their problems by analysing their unhelpful thoughts and behaviours. CBT has been employed widely to treat depression and anxiety. In CBT therapies, resistance proves to be a serious issue, limiting its effectiveness (Westra and Norouzian, 2018). Motivational Interviewing (MI) is an evidence-based client-centred approach to strengthen one’s motivations for behaviour change (Miller and Rollnick, 2023). The core skill of MI is to tailor the therapeutic interventions based on the individuals’ motivational level using the trans-theoretical model of stages of changes (Prochaska and Velicer, 1997).

Understanding client motivational language during therapy helps explain treatment outcomes in psychotherapy up to 35% of variance (Lombardi et al., 2014; Poulin et al., 2019). Observably, in the context of CBT, if the client language shows resistance and 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 that MI is likely to backfire if the client already shows motivation to change early in the session, suggesting personalised interventions at different levels of motivation.

Despite the popularity of self-reported (i.e., questionnaires) measure, observational coding measures is found to correlate better with treatment processes and outcomes in MI (Lombardi et al., 2014; Poulin et al., 2019). And the strength (i.e., the degree of certainty one holds for their utterance), rather than the frequency, of the motivational language is a better predictor (Aharonovich et al., 2008; Campbell et al., 2010; Gaume et al., 2016).

The task of predicting client motivational language can be broken down into two subtasks. The first one, called type task, is to detect the direction of motivation: whether the client is willing to change or not. The other one, called 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? Our experiments utilise AnnoMI (Wu et al., 2023), 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 in total 178 examples with strength annotation, making the second task a low-resourced one.

Recently, Large Language Models (LLMs) have demonstrated their impressive capabilities in few-shot learning (Brown et al., 2020; Chung et al., 2022; Touvron et al., 2023). Ziems et al. (2023) argues that due to reduced costs and increased efficiency in data annotation, LLMs can potentially transform the field of Computational Social Sciences such as psychology and linguistics.

The most popular paradigm to utilise the power of LLMs is via in-context learning (ICL), where the inference is performed given an instruction with a few or no examples. However, ICL is highly sensitive to the prompt format, the choice, and the order of the demonstrated examples (Zhao et al., 2021). Optimising the prompts is, by no means, a trivial task. In contrast, fine-tuning (FT) is arguably a better and cheaper paradigm and instruction FT has proven its capabilities over ICL even in few-shot learning (Liu et al., 2022; Schick and Schütze, 2022; Logan IV et al., 2022).

In this paper, we aim to put the LLMs to the test of detecting the types and strength of client motivational language with the latter task having fewer than 200 gold-labeled samples. Our goal is to explore these following research questions:

RQ1: How does retrieval-based ICL compare with IFT in different training size settings?

With varying training samples for the type and a fixed number for strength tasks, we compare ICL approach by Su et al. (2023) and IFT. The results show that both can perform under low-resourced setting. Yet, IFT yields better performance as the training data increases, whereas that of ICL remains quite stable when the number of in-context examples is low (i.e. fewer than 5).

RQ2: How does IFT with multitask predictions compare with single-task predictions?

During real therapies, the therapists need to perform two tasks simultaneously. Inspired by Varia et al. (2023), we combine two tasks into one instruction and fine-tune the models in a multitasking scenario and compare with single-task instructions. Overall, single-task learning leads to higher scores. Our analysis reveals that ICL is preferable to IFT when the training data is heavily imbalanced as ICL can exploit the massive underlying knowledge of LLMs to solve the task. In contrast, with IFT, the models are unable to learn properly without data.

2 Related Works

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 trainers 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). Even though linguistic features are still the most popular (Pérez-Rosas et al., 2017; Cao et al., 2019; Tavabi et al., 2021; Gibson et al., 2022), researchers have employed speech and facial expressions in a multimodal system. Acoustic features, however, are found to contribute little to the prediction (Aswamenakul et al., 2018; Singla et al., 2020; Tavabi et al., 2020). In contrast, Nakano et al. (2022) show that integrating both linguistic and facial information is effective to detect client behaviour codes.

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 well studied by linguistics 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). Furthermore, researchers examine hedges to understand the social power between interlocutors (Prabhakaran et al., 2018), rapport in peer-tutoring (Raphalen et al., 2022), and reviewers' confidence in their evaluation of scientific papers (Ghosal et al., 2022). Though most works has pursued machine learning solutions, rule-based approach is still a popular choice in detecting certainty and uncertainty cues in texts (Ulinski et al., 2018; Islam et al., 2020; Raphalen et al., 2022). To the best of our knowledge, we are the first in NLP to adopt verbal commitment expressions to understand speakers' motivation in psychotherapy.

In-Context Learning (ICL): ICL is the paradigm introduced by Brown et al. (2020) to demonstrate 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 the 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. Yet, their experiments are conducted with vanilla FT, not instruction FT.

Instruction Fine-tuning (IFT): IFT is the paradigm to boost 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). Additionally, Varia et al. (2023) show that IFT can perform multitask predictions in one prompt: the models are trained with instructions to extract all four elements of the sentiment analysis task. 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% 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, these authors utilise ICL with no selection strategy for examples to use as context despite its importance. Furthermore, their prompt setup includes searching for a `verbalizer` to map the models' vocabulary to the labels. For example, for 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

"Commitment" phenomenon has a long history in linguistics. Markers of commitment have been identified and studied to understand the speakers' attitude towards the truth value conveyed in their utterances (Boulat and Maillat, 2023). 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). 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"*. In this paper, we employ the two linguistic terms **boosters** and **hedges** 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 intervention treatment, e.g., whether the therapist should focus on addressing client's resistance or move to discuss action plans (Westra and Norouzian, 2018).

Compared with the frequency of client language (i.e., counting each type), commitment strength is a better measure of behaviour outcomes(Aharonovich et al., 2008; 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.

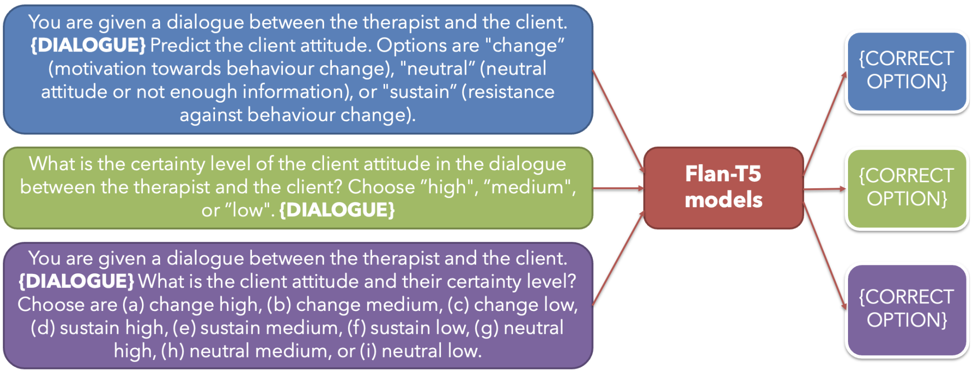

Figure 2: Considered as a generation problem, the models should generate the correct label which is specified as different options in the instruction.

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.

Our paper employs the strength rating approach similar to that of Gaume et al. (2016)[1]: 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. In this paper, we use the word lists of **boosters** and **hedges** by Hyland (2005); Islam et al. (2020); Zhou et al. (2023).

4 Methodology

We consider a set of dialogues where each consists of one therapist turn and one client turn. The former serves as dialogue history while the model should 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 a sentence. If the client starts the conversations, not the therapist, the dialogue consists of one client turn only.

Our experiments utilise Flan-T5 models which are fine-tuned on 1k8+ NLP tasks and shown to outperform other models with the same size up to 26% (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). 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 small number of parameters which can either be a subset of parameters of the existing models or a newly added parameters (Lialin et al., 2023). We employ LoRa (Hu et al., 2022), which performs parameter update of the weight matrix by decomposing it into lower-rank matrices and then train 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). Figure 2 illustrates our instruction fine-tuning (IFT) process. An example dialogue is *"Therapist: Yeah. Hmm, that might be a start. Client: I think I could- I think I could probably handle that."*. The correct options for three instruction are "change", "low", and "change low" respectively. The model is prompted to produce a `type` and/or `strength` classification by concatenating the dialogue with the corresponding instruction template depicted in Figure 2. Our goal is to automatically detect of both 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 conversations in English annotated by MI experts. Each client utterance is assigned one type of motivation language (i.e, "change", "sustain", or "neutral"). The dataset

[1]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.

is heavily imbalanced: the number of "change", "sustain", and "neutral" utterances are 1178, 546, 3093 respectively. We randomly select 600 utterances to serve as test set. From the remaining utterances, `fast voke-k` algorithm (Su et al., 2023) is employed to obtain 300 most diverse samples for the validation set and k samples for training set, with $k \in \{50, 100, 200, 300, 3k6\}$.

`Strength` Data: MI Skill Code (MISC) is a behavioral coding system, developed to assess MI session. It is open-source and available to download from CASAA's website[2]. The number of samples 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 training and validation sets respectively. Mosbach et al. (2023) propose that 50 samples as validation set are sufficient to select the best performing checkpoints. Using the MISC 2.0 (Miller et al., 2003) guideline and the list of certainty markers from Section 3, the first author of this paper, who has both bachelor and master degrees in Computational Linguistics, manually assigns a strength value (i.e., "high", "medium", or "low") for each client turn in the test set from the previous task. When textual information alone is insufficient, we consult the videos to assist with annotation process.

Mixed Data: In the mixed multitask settings, we mix a maximum number of k {instructions, outputs} pairs of each prompt formula, with $k \in \{100, 200, 300\}$. As the number of gold-labelled samples with strength value is limited, *mixed-200* and *mixed-300* datasets contain more samples with the `type` prompt than the other two. The `strength` and multitask instructions use the same dialogues but with different labels: only 3 labels for `strength` samples but 9 for multitask data.

5.2 Experimental Setup

Baselines: Two baselines are employed: (1) zero-shot ICL settings with `Flan-T5-XXL`[3] (Chung et al., 2022) and `GPT-3.5-turbo`[4] and (2) traditional FT with `RoBERTa-large`[5] (Liu et al., 2019).

ICL setting: 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 one-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 setting: We fine-tune `Flan-T5-XXL` with instructions as specified in Section 4. In single-task settings, each model is fed with either `type` or `strength` instructions only. Our multitask settings employ the multitask one while the mixed setup uses all three instructions. Figure 2 depicts the instructions used in our experiments.

Number of parameters: We use LoRa implemented in `peft` library[6] 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.

Hyper-parameters selection: RoBERTa is trained until convergence with the learning rate of 1e-5. As for Flan-T5, we use Weights and Bias[7] to search for the best learning rate and finally 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. Further details about our training is in Appendix B.

Evaluation metrics: We employ `accuracy` and `f1 score` macro-averaged calculated by `scikit-learn` (version 1.3) (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 best f1 scores on the validation sets during training.

6 Results

6.1 Single-Task Learning: `Type`

Figure 3 shows the results of the `type` task (i.e., predicting whether the client has "change", "neutral", or "sustain" attitude to behaviour change) on the test set. Flan-T5 and GPT-3.5 with zero-shot obtain f1 scores of 0.45 and 0.53 respectively. The performance of Flan-T5 with zero-shot corresponds to those of RoBERTa and Flan-T5 when trained on 100 samples, whereas GPT-3.5 with zero-shot yields the same score as RoBERTa trained on 200 samples. Interestingly, both GPT-3.5 and Flan-T5

[2] https://casaa.unm.edu/tools/misc.html
[3] https://huggingface.co/google/flan-t5-xxl
[4] https://platform.openai.com/docs/models/gpt-3-5
[5] https://huggingface.co/roberta-large
[6] https://huggingface.co/docs/peft/index
[7] https://wandb.ai/

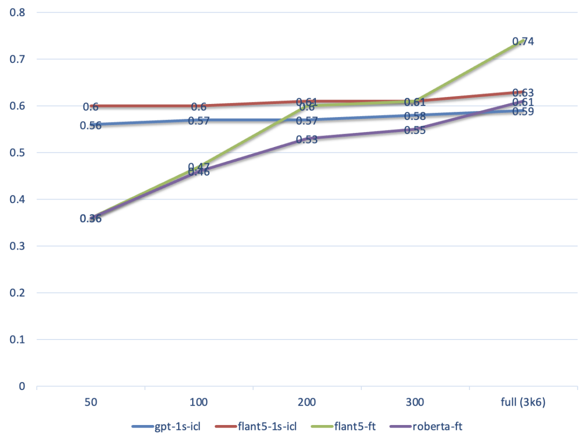

Figure 3: F1 scores on `type` task with different training samples shown on the horizontal axis.

with one-shot ICL exhibit similar behaviour: their performances stay relatively consistent regardless of the number of samples that can be selected as demonstrations. In contrast, for fine-tuning, normally the model performance is positively correlated with the data size. Additionally, Flan-T5 with IFT converges with 200 samples, similar to the findings of Gupta et al. (2023).

Hallucinated Output Label: Framed as a generation problem, instruction-tuned models can produce ill-formed outputs. When analysing the results, we discover that Flan-T5 trained on 50 and 100 samples generates such outputs: 2 for each condition. In contrast, ICL with either zero- or multiple shots does not cause the same issue. After 2 hallucinated labels are replaced with "neutral", F1 scores for Flan-T5 models with 50 and 100 training data size jump from 0.36 and 0.47 to 0.59 and 0.62 respectively. As a result, the new score obtained on 100 samples completely outperforms two one-shot ICL variants while the one on 50 samples is analogous to one-shot Flan-T5. Observably, under this condition, IFT with varying training data from 50 to 300 leads to comparable results unless trained on full dataset with thousands of examples.

6.1.1 Ablation with Output Space Label

	50	100	200	300	full
all	0.59	0.62	0.60	0.61	0.74
simplified	0.56	0.58	0.59	0.59	0.71

Table 1: F1 scores in our ablation studies using **all** and **simplified** instructions with different data size.

With IFT, specifying output space label proves

	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 4: 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.

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 4 illustrates two conditions **all** and **simplified** for our ablation studies. Table 1 reports results on f1 scores across different training data size. All hallucinated outputs are converted to "neutral" label. 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. These 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.

6.2 Single-Task Learning: `Strength`

	accuracy	f1
gpt 0-shot	0.43	0.35
gpt 1-shot	0.39	0.30
flant5 0-shot	0.30	0.29
flant5 1-shot	0.38	0.38
flant5 ift	**0.67**	**0.61**
roberta ft	0.52	0.48

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

This task utilises the `strength` data as specified in Section 5.1, consisting of 50 "high", 35 "medium", and 43 "low" labels in the training set. Results on the test set of 600 samples are reported in Table 2. Surprisingly, retrieval-based ICL with 1-shot fares quite poorly, even worse than fine-tuned RoBERTa. Analysing the confusion matrices, Flan-T5 and GPT-3.5 appear to struggle with "medium" and "high" labels respectively with both recall scores are below 0.1.

GPT-3.5 suffers a drop in performance when shifting from zero-shot to one-shot. Previous works attribute it to majority label bias in which GPT-3 merely reuses the class of the only example in the instructions (Zhao et al., 2021). However, we observe no such phenomenon in this task. 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.

6.2.1 Ablation with Dialogue Context

	accuracy	f1
gpt 1-shot w-th	0.39	0.30
gpt 2-shot w-th	0.43	0.34
gpt 3-shot w-th	0.42	0.33
gpt 4-shot w-th	**0.43**	**0.35**
gpt 1-shot wo-th	0.39	0.34
gpt 2-shot wo-th	0.38	0.33
gpt 3-shot wo-th	0.40	0.35
gpt 4-shot wo-th	0.37	0.33

Table 3: Results for GPT with and without the previous therapist utterance in the demonstrations, shortened as *w-th* and *wo-th* respectively.

One hypothesis about the poor performance of ICL is due to the mismatch between the dialogue served as context and the test dialogue. As indicated in Section 5.1, the test set is taken from `AnnoMI` dataset (Wu et al., 2023): each dialogue consists of one therapist turn and one client turn. However, the examples from MISC guidelines have only one client turn. Therefore, we conduct an ablation studies to understand the effect of this mismatch: in the original experiments, called *w-th*, the test dialogue have both therapist and client turns while in the *wo-th* condition, the test dialogue contains only the client turn. Additionally, we use GPT-3.5 with multiple shots using retrieval-based ICL (Su et al., 2023).

Table 3 reports the results of our ablation. The overall trend suggests that having longer context history for the test sample helps improve the ICL performance despite some mismatch between the format of test sample and that of the demonstrated example. 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-shot w-th` 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. We leave the investigation of the sensitivity of in-context examples to future works.

6.3 Multitask Learning

	`type`		`strength`	
	acc.	f1	acc.	f1
gpt 0-shot	**0.53**	**0.49**	0.45	0.39
gpt 1-shot	0.50	0.43	0.48	0.47
flant5 1-shot	0.43	0.34	0.34	0.34
flant5 ift	0.32	0.29	**0.61**	**0.58**

Table 4: 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 by using the third instruction shown in Figure 2. Because of hallucination issue, we use regular expressions to get the predictions and replace the ill-formed labels with either "neutral" or "medium" depending on the task. Table 4 reports the results. These experiments use the `strength` dataset (Section 5.1) because the samples from MISC guidelines have both `type` and `strength` labels.

The first observation is that overall, single-task learning (STL) still yields better performance on a large margin, especially for `type` task. 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. 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% in the test set. Yet, no "neutral" samples exist in the training set, which explains why the models are unable to learn properly. Appendix A 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.

6.3.1 Multitask Learning with Mixed Data

	type		strength	
	acc.	f1	acc.	f1
flant5 ift mix100	0.36	0.36	0.68	**0.59**
flant5 ift mix200	0.34	0.36	0.69	0.56
flant5 ift mix300	**0.44**	**0.43**	**0.71**	0.58

Table 5: Results on multitask learning using mixed data.

In this setup, we experiment with mixing a maximum number of samples from type and strength tasks with multitask samples (See Section 5.1). In other words, the models are fine-tuned with three instructions all together as depicted in Figure 2. This setup is similar to that of Varia et al. (2023) but we frame it as a cloze-quiz problem, not a generation one. Our aim is to investigate whether adding data from other tasks can improve performance on a downstream task. More importantly, type data is expected to help the models learn to predict “neutral” class. Results are reported in Table 5. Though the models still struggle to learn “neutral” class, the more type samples are in the training set, the higher the recall scores are. However, the higher the number of mixed data is, the more ill-formed outputs are generated for the strength task. As a result, performance on type increase while that on strength task decreases. The reasonable strength scores are due to a high amount of “medium” predictions by the models where the test set is imbalanced with nearly 60% samples belonging to this class. Overall, our results contradict those of Varia et al. (2023): STL outperforms MTL in our setup.

Our hypothesis is that the similarity in the labels of three instructions confuse the learning (e.g., in some cases, the correct label is “neutral” but in other cases, it has to be “neutral high”, “neutral medium” or “neutral low”). Additionally, as the likelihood that the correct label starting with type class is twice higher than with strength class, the models are unable to learn it properly. Indeed, when employed the models trained on mixed dataset to make predictions on single tasks, the outputs for *strength* task are overwhelmed with type labels. It is unclear whether the issue is due to similarity in label space or IFT is 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. In future works, we would like investigate this problem further with varying data size.

7 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 multimodal system would have implications for the development of personalised healthcare agents. In this paper, we break it down into two sub-tasks: predicting the direction of their motivation (i.e., type task), and the verbal commitment strength (i.e., strength task). Our experiments employ GPT-3.5 and Flan-T5, and compare retrieval-based ICL with IFT on varying training data size. Regarding **RQ1**, our findings indicate that both can perform under few-shot settings. Both appear to be sensitive to the instructions: removing label descriptions for IFT or context history for ICL hurts the performance. 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). In contrast, IFT is more stable: adding more data generally leads to better performance, while it has no such effect for ICL. However, IFT suffers from generating ill-formed outputs when trained with a small number of samples. As for **RQ2**, when framing the multitask instructions as a single task of choosing the correct option, ICL outperforms IFT when the label distribution is heavily imbalanced, e.g. some labels might not exist in the training data. In this case, exploiting the massive knowledge of the LLMs to solve the tasks is preferable. Mixing data from different tasks appears to confuse the models by the similarity and/or the multiple-word format of the output labels. In the future works, we would like to investigate this issue on varying training data and model size.

8 Limitations

Annotation of AnnoMI dataset: As the conversations in AnnoMI (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. Furthermore, the labels are assigned to turns, not sentences. Therefore, many samples contain no information to help the models make predictions (e.g., *"-forms."*). The MISC guidelines, however, suggest a fine-grained annotation based on sentences or phrases. Additionally, we observe many samples consisting of multiple sentences whose direction and strength of motivation can move from one end to another as the client speak. This explains partly the low inter-annotator agreement on `AnnoMI`.

Annotation of certainty level: As explained in Section 3, we use lists of linguistic certainty markers to manually annotate the strength value of an utterance. Yet, our observation is that some markers' class can depend on context. For example, *"I think"* is often classified as "low" strength because it shows the lack of confidence of the speaker. However, when watching the videos, we sometimes do not detect such low confidence. In fact, "I think" as a **hedge** word might probably imply politeness or reflect social and power relations between the interlocutors (Prabhakaran et al., 2018). Additionally, the motivation for this paper is to have a sanity check on whether LLMs can be employed for low-resourced tasks in psychotherapy and if yes, how we can best leverage them. Therefore, we only have one annotator for the test set in the `strength` task. In future works, we would approach the annotation process in a more controlled manner.

Multimodal system: 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.

9 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[8]. 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 have implications for the development of intelligent systems in 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., 2020).

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

Acknowledgement

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.

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.

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.

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.

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.

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.

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.

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].

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. ArXiv:2106.09685 [cs].

Ken Hyland. 2005. Stance and engagement: a model of interaction in academic discourse. *Discourse Studies*, 7(2):173–192.

Jumayel Islam, Lu Xiao, and Robert E. Mercer. 2020. A Lexicon-Based Approach for Detecting Hedges in Informal Text. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3109–3113, Marseille, France. European Language Resources Association.

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.

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. ArXiv:2205.05638 [cs].

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. ArXiv:2301.13688 [cs].

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].

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? ArXiv:2202.12837 [cs].

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. 2020. 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.

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, 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.

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.

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.

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].

Morgan Ulinski, Seth Benjamin, and Julia Hirschberg. 2018. Using Hedge Detection to Improve Committed Belief Tagging. In *Proceedings of the Workshop on Computational Semantics beyond Events and Roles*, pages 1–5, New Orleans, Louisiana. Association for Computational Linguistics.

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.

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].

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2023. Can Large Language Models Transform Computational Social Science? ArXiv:2305.03514 [cs].

A Label Distribution

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

Table 6: Label distribution for `type` task.

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

Table 7: 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 8: Label distribution for multitask learning.

Table 6 and Table 7 show the label distribution for `type` and `strength` tasks respectively.

Table 8 shows the number of labels and Figure 5 depicts the percentage of each label for multitask learning in Section 6.3. In the mixed datasets, we add the data with **type** and **strength** labels but the amount of multitask data remains unchanged.

B Training Details

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. To search for the best learning rate with Flan-T5, we use Weights and Bias[9] to randomly sample from the range of 5e-3

[9] https://wandb.ai/

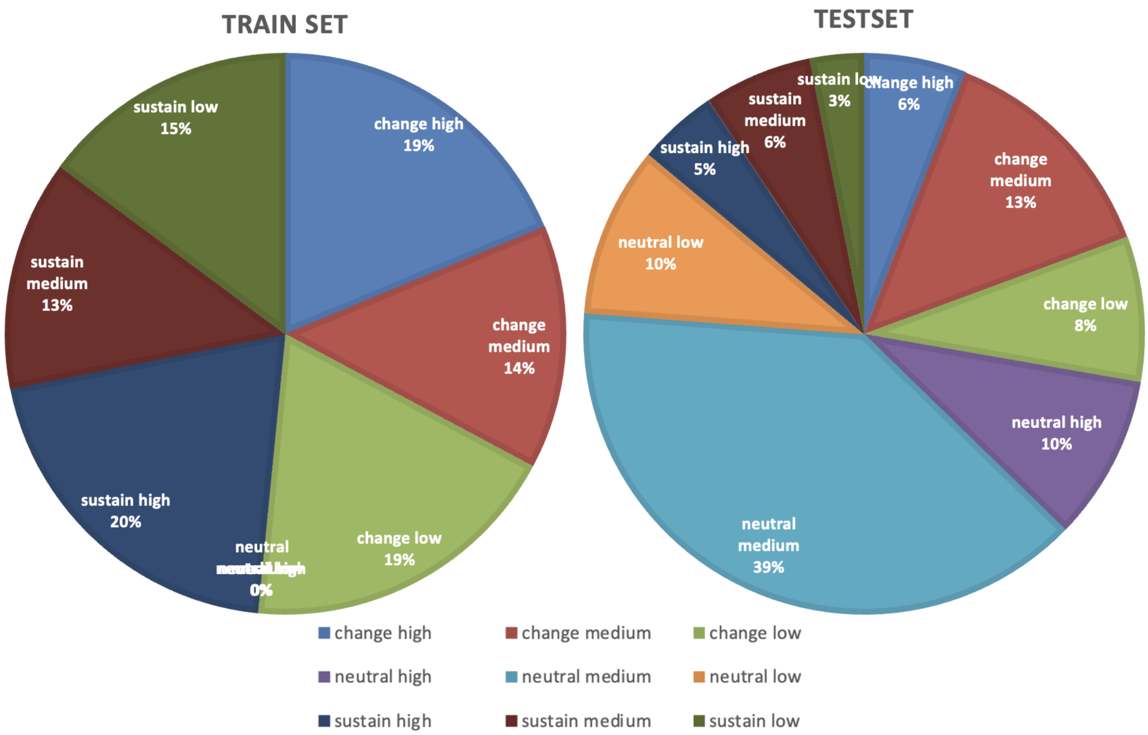

Figure 5: Label distribution for multitask learning (Section 6.3). The training set contains no samples of any "neutral" variants even though they make up for nearly 60% of the test set.

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 use a fixed seed for reproducibility purposes.

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. However, but the instruction-tuned models with LoRa adapters take more than twice the latency even after the adapters have been merged with the original models.