Therapist Self-Disclosure as a Natural Language Processing Task

Natalie Shapira*
nd1234@gmail.com

Tal Alfi-Yogev*
talalfi@gmail.com

Abstract

Therapist Self-Disclosure (TSD) within the context of psychotherapy entails the revelation of personal information by the therapist. The ongoing scholarly discourse surrounding the utility of TSD, spanning from the inception of psychotherapy to the present day, has underscored the need for greater specificity in conceptualizing TSD. This inquiry has yielded more refined classifications within the TSD domain, with a consensus emerging on the distinction between immediate and non-immediate TSD, each of which plays a distinct role in the therapeutic process. Despite this progress in the field of psychotherapy, the Natural Language Processing (NLP) domain currently lacks methodological solutions or explorations for such scenarios. This lacuna can be partly due to the difficulty of attaining publicly available clinical data. To address this gap, this paper presents an innovative NLP-based approach that formalizes TSD as an NLP task. The proposed methodology involves the creation of publicly available, expert-annotated test sets designed to simulate therapist utterances, and the employment of NLP techniques for evaluation purposes. By integrating insights from psychotherapy research with NLP methodologies, this study aims to catalyze advancements in both NLP and psychotherapy research.

1 Introduction

Therapist Self-Disclosure (TSD) has various definitions in the literature (e.g., Henretty and Levitt, 2010; Hill, 2009; Knox and Hill, 2003; Vandernoot, 2007; Watkins Jr, 1990), but the one theme that unites these definitions is that TSD involves a therapist's personal self-revelatory statements. In other words, such statements are those that *reveal something personal about the therapist*. This definition refers to verbal disclosures and excludes disclosures that are nonverbal (Hill and Knox, 2001).

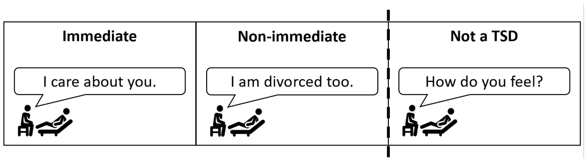


Figure 1: Two types of therapist self-disclosure (TSD).

The attitude toward the use of TSD in psychotherapy has changed over the years. Classical psychoanalytic clinicians tended to emphasize the importance of the therapist's anonymity, equanimity, and abstinence (Freud, 1912; Goldstein, 1997). Many of them viewed TSD as a boundary violation and believed it derailed therapy by removing the focus from the client (Zur, 2004). Over the years, however, therapists and theorists across diverse orientations have increasingly converged around the perspective that TSD can yield a range of positive outcomes when employed purposefully and thoughtfully and that refraining from TSD in every instance may potentially lead to adverse consequences for both the client and the overall therapeutic process (Eagle, 2011; Farber, 2006; Hill and Knox, 2001; McWilliams, 2004; Ziv-Beiman, 2013).

The first to embrace a pro-disclosure approach were the humanistic theorists (Bugental, 1965; Farber, 2006). They have postulated that therapists can demonstrate openness, strength, vulnerability, and the sharing of intense feelings cautiously through TSD. By doing so, they invite the client to follow suit and cultivate an environment of openness, trust, intimacy, gains in self-understanding and change (Henretty et al., 2014; Hill and Knox, 2001; Knox et al., 2001; Kottler, 2003). Cognitive-behavioral therapists describe TSD as a tool that is useful for strengthening the therapeutic bond, normalizing clients' experiences of their difficulties, challenging negative interpretations of emotions and behavior, enhancing positive expectations and motivation for change, and modeling and reinforcing

* Equal contribution.

March 21, 2024 ©2024 Association for Computational Linguistics

	Therapist Self Disclosure (TSD) Therapist reveals something personal about himself		
Category	**Immediate**	**Non-immediate**	**Not a TSD**
Definition	Utterance focuses on articulating the therapist's feelings, thoughts and opinions towards the client, treatment, or therapeutic relationship.	Utterance reveals information about the therapist's personal life outside of therapy, such as beliefs, values, life circumstances and past experiences.	Any comment or other therapeutic intervention (e.g., interpretation, clarification, confrontation, reflection, etc.) that does not include therapist self-disclosure.
Example	*I felt really proud of you when you shared that accomplishment with me.*	*I've used mindfulness exercises in my own life to stay grounded during challenging times.*	*You say you love your family.* (Reflection)

Table 1: Therapist self-disclosure task definition.

desired behaviors (Dryden, 1990; Freeman et al., 1990; Goldfried et al., 2003). Feminist and multicultural approaches also advocate the use of TSD to promote equality, empower the client and reduce clients' feelings of shame, and encourage collaboration in therapy (Brown and Walker, 1990; Mahalik et al., 2000).

In line with the absence of agreement among the mentioned theoretical viewpoints, a body of research presents a multitude of often conflicting or inconclusive findings. These studies delve into diverse facets of TSD, employing different methodologies to assess its influence on clients.

Although there is no consensual conceptualization of the term TSD, as former studies and theoreticians have used a variety of classifications (for a review see Henretty and Levitt, 2010; Ziv-Beiman, 2013), there is growing agreement that one unifying and comprehensive distinction is between *immediate and non-immediate TSD*, which was first put forward by McCarthy and Betz (1978) and later adopted by many psychotherapy researchers (e.g., Alfi-Yogev et al., 2021; Hill et al., 2018; Audet, 2011; McCarthy Veach, 2011; Ziv-Beiman et al., 2017). Whereas ***immediate TSD*** *(also known as self-involving or interpersonal disclosure) focuses on the articulation of the therapist's feelings, thoughts, and opinions toward the client, treatment, or therapeutic relationship*, ***non-immediate TSD*** *(also known as self-revealing or intrapersonal self-disclosure) reveals information about the therapist's personal life outside of therapy, such as beliefs, values, life circumstances, and past experiences.* Immediate TSD and non-immediate TSD are distinctly different utterances. Immediate TSD utterances are primarily “We-focused", whereas non-immediate TSDs are “I-focused". For example, an immediate TSD would be, “I felt proud of you when you shared that accomplishment with me.” Whereas an example of a non-immediate TSD might be, “I've used mindfulness exercises in my own life to stay grounded during challenging times.” Table 1 summarizes all definitions and examples.

Theoretically, the two types of TSD serve distinct functions. Immediate TSD may promote dyadic engagement in the therapeutic process, enable clients to recognize their interpersonal impact, foster insight, facilitate the identification, experience, and integration of dissociative components, expand the client's emotional repertoire, and may lead to symptom reduction (Alfi-Yogev et al., 2021, 2024; Hill et al., 2018; Ziv-Beiman et al., 2017). In contrast, non-immediate TSD may enhance client self-acceptance, mitigate feelings of shame and self-criticism, and foster an increased sense of attunement from their therapists, contributing to a greater sense of understanding. It can promote rapport, model new perspectives and behaviors, and help balance the therapeutic relationship (Audet, 2011; Audet and Everall, 2010; Hill et al., 2018).

To investigate the different roles of TSD in treatment, there are a variety of methods (we detailed representative studies in Section 2.1). One of the methods is by using self-report questionnaires. This method has the disadvantages of lack of objectivity and consequently biasing the results of the

Research Work	Literature Domain	Method	Resolution	Clinical Data	Public Testset	Speaker Identity	Subcategories
Valizadeh et al. (2021)	NLP	Experts	Utterance	-	✓	Listener	-
Reuel et al. (2022)	NLP	Analysis	Utterance	-	✓	Listener	-
Ravichander and Black (2018)	NLP	Crowdsourcing	Utterance	-	✓	Chat-bot	-
Welivita and Pu (2022a)	NLP	Crowdsourcing	Utterance	-	✓	Listener	-[1]
Pinto-Coelho et al. (2018a)	Psychotherapy	Experts	Event	✓	-	Therapist	✓
Levitt et al. (2018)	Psychotherapy	Experts	Session	✓	-	Therapist	✓
Alfi-Yogev et al. (2021)	Psychotherapy	Self-report	Session	✓	-	Therapist	✓
Fuertes et al. (2019)	Psychotherapy	Self-report	Session	✓	-	Therapist	✓
Ziv-Beiman et al. (2017)	Psychotherapy	RCT	Treatment	✓	-	Therapist	✓
This paper	Hybrid	Experts	Utterance	$\checkmark^{pseudo}$	✓	Therapist	✓

Table 2: Comparison with related work

research. Another method is by external expert human judges that annotate the session. This method has the disadvantage that it requires time and is also expensive to train expert judges and conduct the annotation process.

Modern technologies, such as automated speech recognition, NLP techniques, and machine learning models, provide the potential to substitute human evaluators, significantly augmenting scale and precision in the study of treatment mechanisms.

These tools can greatly expand the evaluation of TSD and enable the testing of more sophisticated hypotheses about therapeutic change (e.g., determining when to disclose and to whom; Alfi-Yogev et al., 2021). Initial efforts in this direction have been initiated, utilizing NLP to automatically categorize therapist interventions from session transcripts (Cao et al., 2019; Malgaroli et al., 2023). To the best of our knowledge, TSD has not yet been explored using these techniques.

Advancements in the field of Natural Language Processing (NLP) have led to recent developments that offer a variety of advanced methods for automatic detection of self-disclosure within texts (we detailed representative studies in Section 2.2). However, these advancements address *self-disclosure* and not ***therapist*** *self-disclosure* and do not take into account the important subtleties of the various sub-classes within TSD. This lacuna can be partly due to the difficulty of attaining publicly available clinical data due to privacy constraints and the need for collaboration between different disciplines.

In addition, the latest works did not incorporate state-of-the-art tools and methodologies such as using Large Language Models (LLMs; Brown et al., 2020; Bommasani et al., 2021; Zhao et al., 2023).

In this study, we adopt the current clinical definition for immediate and non-immediate TSD to facilitate it as an NLP task. Since clinical data is confidential, we created a first-of-a-kind new artificial open-source expert-based test set for TSD (i.e., utterances that could have been said by therapists during therapy, and ground truth annotations by a TSD expert). This test set emphasizes different linguistic characteristics. In addition, we annotated a sample of utterances from an existing dataset of peer support platforms. We propose a method to solve the task using LLMs and report the results.

The paper continues as follows: In Section 2 we describe related works both from psychotherapy research literature and from NLP and review the previous works. In Section 3 we describe the construction process of the new test set (Expert-TSD) and the annotation process of an existing data set (MI) to create a double-check TSD test set (MI'). In Section 4 we describe the technical details of the usage of LLMs, and in Section 5 we discuss the results of LLMs on the new test sets. Finally in Section 6 we conclude and describe potential future work.

[1] Welivita and Pu (2022a) manually annotated a small amount of the sub-categories of inter- and intra-session disclosure (which corresponds to immediate and non-immediate TSD), though they did not publish the annotation results or statistics and recommended continuing research of the sub-categories for future work.

2 Related Work

In this section, we review the existing works (both from clinical psychology and NLP literature) that refer to the evaluation of self-disclosure. In our review, we refer to the domain of the source (psychotherapy, NLP, or hybrid); the method used to determine self-disclosure (self-report questionnaire, crowdsourcing, experts, analysis or randomized clinical trail); the resolution of the data that was investigated (utterance, event, session or treatment); the type of the data (clinical, non-clinical, or pseudo clinical); whether a public test set has been published; the speaker identity (therapist, listener, or chatbot); and whether referring to the subcategories (immediate and not immediate or only self-disclosure in general).

Table 2 summarizes the related studies according to the categories presented. As can be seen, this work is the first to construct an open expert-based test set for TSD that refers to immediate and non-immediate TSD.

In the next sections, we provide an extensive literature review of both psychotherapy (Section 2.1) and NLP (Section 2.2) approaches for this task.

2.1 Psychotherapy Research Perspective

Immediate and non-immediate TSD have typically been evaluated through judgments of therapist behavior in psychotherapy sessions. One approach involves trained external judges coding TSD interventions as present or absent in sentences or speaking turns in recorded or transcribed sessions (e.g., Hill, 1978; Stiles, 1979). Alternatively, another evaluation method involving trained judges includes listening to entire sessions and estimating the frequency or effectiveness of TSD interventions throughout the session (e.g., Hill et al., 2014; Levitt et al., 2018; Pinto-Coelho et al., 2018a,b).

Furthermore, the assessment of immediate and non-immediate TSD has also been conducted through self-report questionnaires provided to clients, therapists, or both. Participants receive definitions of immediate and non-immediate TSD and then retrospectively report the use of these interventions within sessions (e.g., Ain, 2008, 2011; Alfi-Yogev et al., 2021, 2023, 2024; Fuertes et al., 2019).

An additional assessment method involves training therapists to either employ immediate TSD, non-immediate TSD, or refrain from using TSD with their clients. In this randomized clinical trial (RCT) method, clients are categorized into three conditions based on the type of self-disclosure employed by their therapists (e.g., Ziv-Beiman et al., 2017).

Several disadvantages are associated with these methods. First, in self-report measurement, there is a potential for bias in retrospective recall, as feelings and reactions may evolve over time, leading to changes in how participants interpret their experience. Second, in self-report measurement, there is difficulty in identifying the session's specific location when recalled immediate/non-immediate TSD occurred, posing challenges in assessing the interventions' context, manner of delivery, and associated subsequent processes. Third, in evaluation through external judgments, achieving agreement among judges is sometimes marginal due to the intricate task of distinguishing verbal response modes that predominantly focus on grammatical form, while overlooking intent, quality, or manner of delivery. This limitation results in diminished clinical relevance. Fourth, the reliance on training for external judges or therapists is highly time-consuming, introducing inefficiencies to the assessment procedure. Lastly, using an RCT may not always mimic real-life treatment situations.

2.2 Self Disclosure Within NLP Litrature

Valizadeh et al. (2021) created a 6,639-instance dataset comprised of public online social posts covering a wide range of mental and physical health issues, categorized into three groups (no self-disclosure, possible self-disclosure, and clear self-disclosure) with high inter-annotator agreement (= 0.88). They demonstrated that a large percentage of instances from the possible self-disclosure class were misclassified than were instances from the other two classes, suggesting room for future work that disentangles the nuances of ambiguous cases.

Reuel et al. (2022) Analysed several existing self-disclosure related datasets (Wang et al., 2015; Jaidka et al., 2020; Pei and Jurgens, 2020; Omitaomu et al., 2022; Valizadeh et al., 2021) with variety of techniques (e.g., RoBERTa-, LIWC-, LDA-, and EmoLex-based models). All datasets are based on publicly available conversations (forums, Reddit, online platforms, and more) with crowdsourcing annotations for self-disclosure and related tasks (e.g., intimacy, empathy, emotional disclosure, and more). They showed that it is hard for models to generalize between datasets. They found that self-disclosure linguistic correlates with the expression of negative emotions and the use of first-person personal pronouns like “I”. They provide a multi-task model across all available data sets to assess self-disclosure. However, they noted that the data sets they took into account were not annotated based on validated definitions of self-disclosure in psychological literature, but rather had differing labeling instructions, which might lead to inaccuracies when predicting self-disclosure. They recommended that in future work, data that is labeled for a validated self-disclosure definition should be collected and analyzed.

Ravichander and Black (2018) built an open-domain chatbot that engages in social conversation with hundreds of Amazon Alexa users and ran a large-scale quantitative analysis on the effect of self-disclosure by analyzing these interactions. In their work, their definition of self-disclosure was binary. They noted that a more nuanced version that considers both the magnitude and valence of self-disclosure would open up several further research directions, such as analyzing reciprocity matching in the depth of disclosure and analyzing user behavior based on the valence of disclosure.

Welivita and Pu (2022a) created large-scale publicly available datasets (17k) from peer support platforms, annotated by trained crowdsourcing counselors. They labeled TSD, as well as other interventions (e.g., clarification). In their paper, the authors recommend that future work consider the distinction between intra- and extra-session disclosure (equivalent to immediate and non-immediate disclosure).

3 Data

In this section, we describe the creation of two test sets: Expert-TSD and MI’. The first was developed from scratch by an expert, and the second was created by expertly annotating an existing dataset. Both test sets are in English.[2]

The purpose of the first test set is to provide an adequate test for TSD (precision). The purpose of the second test set is to strengthen the findings and to enable an assessment of real data distribution. Real data contains surprising behaviors such as syntax and grammar errors, informal or non-verbal utterances, and more phenomena. It is important to examine behavior in a wide variety of situations (recall) to strengthen our conclusions.

The subsequent paragraphs provide the construction process for each test set.

Expert-TSD. The initial phase of the test set creation process involved a collaborative effort between the authors (an NLP researcher and a clinical psychologist specializing in TSD research). In a comprehensive brainstorming session, the authors discussed the precise definition of TSD and its subtypes as described in psychotherapy literature (see Section 1 and Table 1), as well as potential solutions for recognizing TSD types using shallow heuristics and machine learning.

Next, utterances were generated by the clinical psychologist along with their respective type label. The NLP researcher reviewed the proposed samples marking potential shallow heuristics, such as syntactic features, that a machine learning model might exploit to predict the correct label for the incorrect reasons (see shallow heuristics: Hendrycks et al., 2021; Wu et al., 2021; Kaushik et al., 2019; Geirhos et al., 2020; Glockner et al., 2018). This writing and reviewing procedure was conducted throughout five iterations, with new samples proposed and previous ones fixed.

To mitigate the effect of shallow heuristics, we made sure to diversify utterances over the following properties: (1) the balance of positive and negative examples (i.e., including “Not a TSD” utterances) (2) the length of the utterance (i.e., short sentence below 10 words vs. numerous or long sentences above 20 words), (3) the presence or absence of first-person pronouns words (e.g., I, me, our), (4) the existence of positive or negative sentiment, and (5) the incorporation of questions.

The test set generation rounds were stopped once we surpassed 100 instances (108), which is a sufficient quantity for testing significance.

MI’. We first sampled 650 examples from the MI dataset (Welivita and Pu, 2022b, summarized in Section 2.2), ensuring diversity by extracting 25 instances from each category and 300 from self-disclosure. For each utterance, the TSD expert annotated the TSD type based on the task definition outlined in Section 1 and Table 1.

A total of 277 items were tagged. An effort was made to equally represent each class (“Immediate TSD”, “Non-immediate TSD” and “Not a TSD”). Except for one instance, all of our utterance labels agreed with the MI labels for the binary self-disclosure classification.

[2] The data is available at: https://github.com/NatalieShapira/TherapistSelfDisclosure/

Therapist Self-Disclosure Instructions Prompt

TEST:

Below are definitions of two subcategories of self-disclosure and not self-disclosure:

Non-immediate TSD: Self-disclosure of information about the therapist.
* Relates to disclosing, during a treatment session, facts about the therapists' life outside of the treatment and personal insights they gained, the way they reached these insights, effective / in-effective ways of coping based on their experience and the way they formulated them, emotions that they experience in different situations in their life, etc...
* Example:
Speech turn: I remember going through a career change a few years ago, and it was a challenging time for me. It's normal to feel uncertain during transitions, but it's also a chance to explore new possibilities.
Answer: Non-immediate TSD

Immediate TSD: Self-disclosure that relates to the "here and now".
* Relates to sharing therapists' feelings, associations, and thoughts relating to the client and the issues and topics raised during the session and of their emotions, feelings, and thoughts on the therapy process which they are both part of, etc...
* Example:
Speech turn: I was genuinely excited to hear about the progress you've made.
Answer: Immediate TSD

Not a TSD: Not a Self-disclosure
* Any comment or other therapeutic intervention (e.g., interpretation, clarification, confrontation, reflection, etc.) that does not include therapist self-disclosure.
* Example:
Speech turn: You say you love your family
Answer: Not a TSD (clarification)

For the next speech turn, determine whether it is non-immediate TSD or immediate TSD according to the above definitions.
Speech turn: ***If what you are experiencing seems fine and normal to you, it may be nothing to worry about.***
Answer:

Table 3: Therapist self-disclosure instructions prompt. The bold-italics text is a variable utterance we want to automatically tag with a label (Immediate, Non-immediate, or Not a TSD), all the rest is a constant template.

4 Method

In line with the latest works that examine automated detection of psychology-related tasks by LLMs in-context learning or zero-shot setup (e.g., Murthy et al., 2023; Shapira et al., 2023a,c,b), we investigate the TSD automatic detection abilities of LLMs. We evaluated the two test sets mentioned in Section 3 in-context learning setup.

LLMs and Decoding Parameters. We used two different LLMs: Flan-T5 (Chung et al., 2022)[3] of different sizes `flan-t5-{small, base, large, xl}` and GPT-4 (Brown et al., 2020; Ouyang et al., 2022; Achiam et al., 2023).[4] A single sample (the first) was selected from each model for the analysis of the tagging evaluation. We chose hyper-parameters that minimize randomness, predict the most probable answer (i.e., low temperature, sampling method), and allow for a sufficient number of tokens.

Prompt. As input to the LLMs, we used a prompt that contained the definition of TSD with examples concatenated to the utterance that we wished to automatically tag. The full exact prompt is detailed in Table 3.

5 Results and Discussions

	Immediate	**Non-immediate**	**Not a TSD**	**Total**
Model	**29**	**28**	**51**	**108 (100%)**
Flan-T5-small	0	28	0	28 (23%)
Flan-T5-base	3	28	5	36 (33%)
Flan-t5-large	3	28	0	31 (29%)
Flan-t5-xl	9	28	0	37 (34%)
GPT-4	26	28	43	97 (90%)

Table 4: Evaluation on the expert test set for therapist self-disclosure task (Expert-TSD). The first row represents the number of samples for each category. The rest, each cell represents the number of correct responses for each model.

Expert-TSD Results. The results of the Expert-TSD test set appear in Table 4.

As evident, Flan-T5 exhibits a bias toward the "Non-immediate" class. The results of GPT-4 were surprisingly good (accuracy of 90% on the task; above expected human annotation agreement; and higher than previous self-disclosure literature as reported by Reuel et al., 2022). Note that this method is proposed for practice and as a proof-of-concept and not for real use, see more discussions in the Limitation Section and Ethical Statement.

For the GPT-4, 10% utterances where discrepancies emerged between the labels assigned by the human annotator and those generated by GPT-4, we conducted a manual error analysis and consulted with three additional psychotherapists. Notably, there was no consensus among the therapists regarding whether these utterances constituted TSD.

Upon examining the inconsistencies in labeling between the human annotator and GPT-4, it became apparent that the discrepancies pertained solely to immediate TSD. Specifically, two types of differences were identified: First, instances where the human annotator identified "Immediate TSD" while GPT-4 identified "Not a TSD"; and second, cases where GPT-4 detected "Immediate TSD", but the human annotator detected "Not a TSD".

Determining the frequency of immediate TSD in real therapy sessions poses a considerable challenge. Therapists and clients typically perceive these interventions as integral to the therapeutic dialogue, leading to their routine exclusion from TSD reports. Nevertheless, it is assumed that such disclosures transpire more frequently in therapeutic dialogues than what has been officially reported (Farber, 2006; Ziv-Beiman, 2013).

Moreover, as for instances where GPT-4 identified immediate TSD, but the human annotator did not, it appears that some of the utterances were characterized as *immediacy*.

The term ***immediacy*** was defined by Hill et al. (2014) as *"discussion of the therapeutic relationship by both the therapist and client in the here-and-now, involving more than social chitchat"*. While earlier literature used *immediacy* to refer to immediate TSD utterances, researchers have evolved from defining immediacy exclusively as immediate TSD and now use the term to refer to a more complex phenomenon (McCarthy Veach, 2011). Immediacy extends to therapist responses and behaviors such as feedback, inquiries to gather more information about the client's here-and-now reactions, and primary and advanced empathy to reflect the client's momentary experiences. At times, immediacy utterances are more client-focused, than

[3] **Python packages:** *transformers* (*AutoModelForSeq2SeqLM*, *AutoTokenizer*) and *torch*; **Generation function:** *generate*; **Hyper-parameters:** do_sample=True, max_length=50, from_pretrained:{google/flan-t5-small, google/flan-t5-base, google/flan-t5-large, google/flan-t5-xl}, temperature=0.0001

[4] **Python package:** *openai*; **Generation function:** *ChatCompletion.create*; **Hyper-parameters:** model=gpt-4-0314 temperature=0

therapist-focused. An illustrative example from our data involves the utterance: *“I’ve noticed you seem unhappy when we talk about the disagreement we had last time. Do you think there might be some anger or resentment towards me?”* The human annotator labeled it as “Not a TSD,” while GPT-4 tagged it as “Immediate TSD,” when in fact it represents *immediacy*. This clarification aims to shed light on some of the observed gaps in labeling.

Model	Immediate 6	Non-immediate 135	Not a SD 136	Total 277 (100%)
Flan-t5-small	0	135	0	135 (49%)
Flan-t5-xl	0	133	30	163 (59%)
GPT-4	6	111	134	251 (91%)

Table 5: Evaluation on our annotated sample (MI’) from the MI dataset (Welivita and Pu, 2022b). The first row represents the number of samples for each category. The rest, each cell represents the number of correct responses for each model.

MI’ Results. The results of the MI’ test set appear in Table 5.

MI’ test set, unlike the Expert-TSD test set, contains quotes from peer support platforms and thus does not necessarily represent therapist utterances, nevertheless, we classify the utterances as if they were of a therapist.

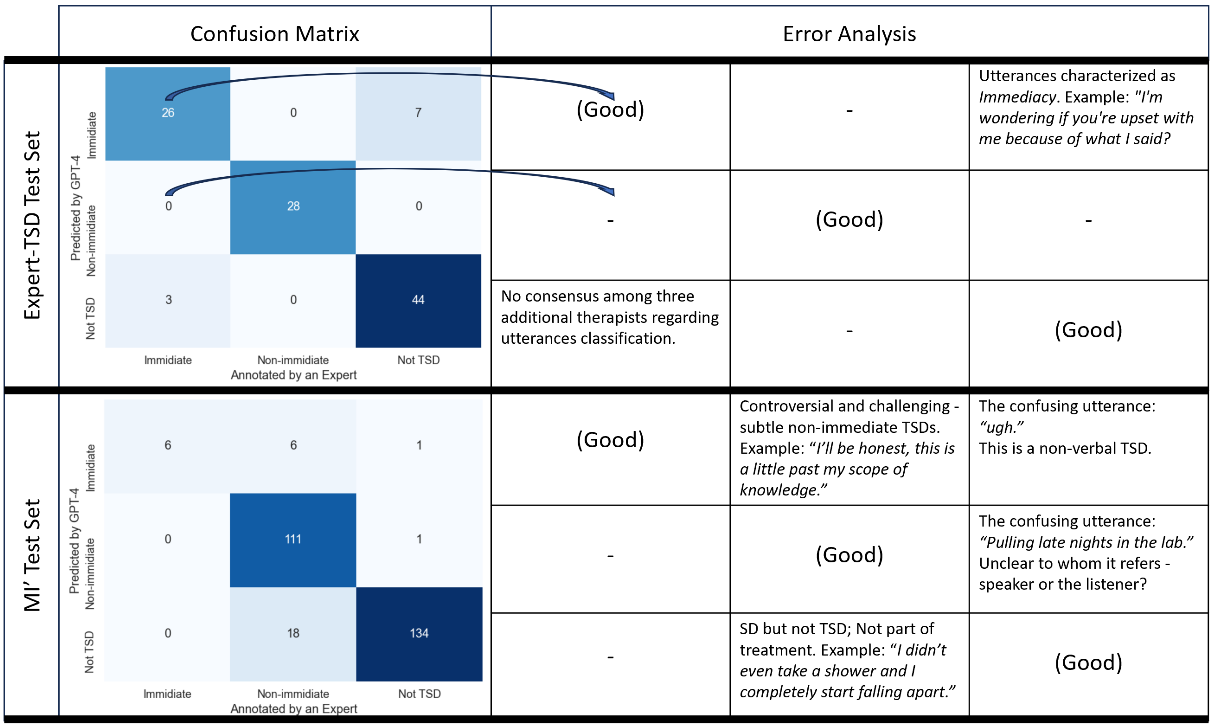
We analyzed utterances in which discrepancies between our human expert annotator and GPT-4 were observed regarding TSD.

Four types of differences were identified: First, instances where the human expert annotator identified “Non-immediate TSD” while GPT-4 identified “Immediate TSD.” Second, instances where the human expert annotator identified “Non-immediate TSD” while GPT-4 identified “Not a TSD.” Third, instances where the human expert annotator identified “Not a TSD” while GPT-4 identified “Non-immediate TSD.” Fourth, instances where the human annotator identified “Not a TSD” while GPT-4 identified “Immediate TSD.” The distinction between the first and second types appears to lie in the level of controversy associated with the TSD. Non-immediate TSD is considered a controversial technique and is seen as challenging fundamental therapeutic principles (Ziv-Beiman, 2013). It appears that GPT-4 labeled more subtle Non-immediate TSDs as “Immediate TSD” (e.g., *“I’ll be honest, this is a little past my scope of knowledge”*), whereas less subtle non-immediate TSDs, to the extent that they may not theoretically be considered part of treatment (e.g., *“I didn’t even take a shower and I completely start falling apart”* note that this example is not only untypical therapist discourse but also grammatically incorrect), were identified by GPT-4 as “Not a TSD.” The third and fourth type included only one utterance. *“ugh.”* was labeled as “Immidiate TSD” by GPT-4 but is a non-verbal disclosure while the formal TSD definition includes only verbal disclosures. *“Pulling late nights in the lab.”* was labeled as “Non-immediate TSD” while it is unclear to whom it refers - (speaker or the listener).

Note that this test set contained only a few examples (6) of immediate TSD. This is due to the nature of the data on which it was based. It is crucial to emphasize that the MI dataset was extracted from online peer support forums, as opposed to therapeutic interactions between a therapist and a client. Therefore, the TSD utterances identified in the study’s data do not portray instances of TSD. The distinction between the MI data in the Welivita and Pu (2022a) study and data derived from therapeutic interactions is also evident in the prevalence of immediate and non-immediate TSDs. Notably, therapeutic sessions tend to feature a higher frequency of immediate TSDs than non-immediate TSDs (e.g., Levitt et al., 2018). Conversely, the MI’ sample from MI indicates a greater prevalence of non-immediate TSD. In peer support conversations, participants predominantly engage in sharing their lived experiences (which is parallel to using non-immediate TSDs- often characterized by an emphasis on individual perspectives; “I-focused”). Given the potentially less committed therapeutic relationships or absence of genuine connections, peers may be less inclined to disclose their immediate feelings in response to the other’s experiences or emotions (referred to as immediate TSDs- where the focus is on shared experiences; “We-focused”).

While analyzing the differences between the two datasets, we observed that in the Expert-TSD dataset, the disparities between labels assigned by the human annotator and those generated by GPT-4 were exclusively related to immediate TSD. Conversely, in the MI dataset, the discrepancies between labels assigned by the human annotator and GPT-4 were particularly associated with non-immediate self-disclosure. This discrepancy may be attributed, in part, to the higher frequency of non-immediate self-disclosure utterances in the MI dataset.

Overall, the results of GPT-4 in MI’ dataset are



Expert-TSD Test Set	(Good)	-	Utterances characterized as *Immediacy*. Example: *"I'm wondering if you're upset with me because of what I said?*
	-	(Good)	-
	No consensus among three additional therapists regarding utterances classification.	-	(Good)
MI' Test Set	(Good)	Controversial and challenging - subtle non-immediate TSDs. Example: *"I'll be honest, this is a little past my scope of knowledge."*	The confusing utterance: *"ugh."* This is a non-verbal TSD.
	-	(Good)	The confusing utterance: *"Pulling late nights in the lab."* Unclear to whom it refers - speaker or the listener?
	-	SD but not TSD; Not part of treatment. Example: *"I didn't even take a shower and I completely start falling apart."*	(Good)

Figure 2: Confusion matrix (left) and error analysis (right) between GPT-4 predictions and the gold standard annotated by expert in the Expert-TSD test set (above) and MI' test set (below). Each cell on the right represents an explanation for a significant part of the examples of the corresponding cell.

similar to the results in Expert-TSD dataset (i.e., high accuracy classification). This is despite the complexity of real data, which does not always allow a clear decision regarding whether or not there was self-disclosure (e.g., mixes other interventions that make it difficult to decide which of them is more significant). Error analysis shows that the error type differs between the test sets. While the errors in Expert-TSD were mostly controversial among experts, here, there were clear errors in places labeled "Not a TSD" by GPT-4. At the same time, the utterance contained a clear "Non-immediate TSD" (e.g., *I've always thought suicide was something I would never do, but lately I'm getting scared that I'm gonna reach a point where I simply can't handle any more of this.*). Note that all these places (18) were utterances that therapists would not say during therapy. This raises the suspicion that the model was pretrained on a task related to self-disclosure in a clinical-related domain rather than a general domain. Analyzing the different behaviors in different data distributions can give a glimpse into the findings of Reuel et al. (2022) that showed that models that involve self-disclosure exhibit limited generalization capabilities when applied to different datasets.

Figure 2 summarises GPT-4 confusions and error analysis in both test sets.

6 Conclusion

In this study, we have formalized Therapist Self-Disclosure (TSD) as a Natural Language Processing (NLP) task by introducing expert-annotated test sets to simulate therapist utterances and utilizing Large Language Models (LLMs) for in-context learning as a solution. This work demonstrates how psychotherapy literature can help capture language nuances. In addition, this work shows the potential of NLP tools to enhance theoretical understanding of existing issues in psychotherapy.

The contribution to the NLP domain lies in the task's potential to serve as a challenging benchmark for optimizing results of accuracy or efficiency while the proposed method serves as a baseline. In addition, The expert-annotated utterance set can function as a test set for model evaluation (as in this study) or as valuable training examples for few-shot learning or other methods.

In the field of Psychotherapy Research, our study offers carefully documented guidelines and a testing ground for human annotators aiming to engage in manual annotations of TSD. Our proposed method lays a promising foundation, however, it necessitates ongoing exploration and refinement before implementation. Future research will have to examine its readiness and effectiveness for automated TSD tagging in real-world data contexts.

Limitations

Data. We annotated at the utterance level only, without considering a broader context. For instance, utterances where the therapist responds to a personal question without initiating the disclosure are also considered disclosures, such as:
Client: *Do you care about me?*
Therapist: *Of course.*
The last example represents a TSD, though in our test set, there is no option to represent such a scenario. Another example that requires a broader context:
Client: *I want to tell him "it's especially for you"*
Therapist: *that I care about you*
The last example does not represent a TSD but rather constitutes a reflection in which the therapist employs the first person. When taken as an isolated utterance without context, the therapist's response may be perceived as TSD.

Furthermore, given that the utterances were both generated and annotated by a single expert, there is a potential for unconscious bias in the data, and the utterances may not be as representative as those found in actual treatment data. Different annotators can have different labels for the same utterance.

Method. The method we suggested uses a long and expensive prompt. We did not try to optimize the length of the prompt. Moreover, there might be more efficient and accurate methods available.

Results and Conclusions. The notably favorable outcomes observed with GPT-4 on the test sets may indicate a seemingly straightforward task that GPT adeptly handles. Conversely, these results could stem from the limited diversity and insufficient representation of real data within the examples we generated. In practical scenarios, real data often diverge from artificial test sets. Therapists' utterances commonly extend beyond 1-2 sentences, incorporating a combination of interventions, thereby complicating the task's definition. This reality highlights the challenge of accurately capturing the complexity and diversity inherent in therapist communications. Thus, while our proposed method presents a promising foundation, it requires further exploration and refinement before implementation. Continued research is essential to enhance its readiness and effectiveness in the context of automated TSD tagging in real data.

Ethical Statement

Data. The new test set used in this study is publicly available. The authors evaluated the utterances to ensure that they did not contain offensive content. None of the samples found in the test set were taken from a real therapy.

Models. LLMs may generate offensive content if prompted with certain inputs. However, we used them for evaluation only, with non-offensive inputs, and we did not encounter any problematic responses.

Privacy. In our experiments we did not use confidential data. Therefor we had no problem using the GPT-4 model that processes the data through OpenAI's servers. **Please note that if confidential data is used, a thorough check must be performed regarding models and data leakage from the local computer to the outside.**

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Stacie Ain. 2008. *Chipping away at the blank screen: Therapist self-disclosure and the real relationship*. University of Maryland, College Park.

Stacie Claire Ain. 2011. *The real relationship, therapist self-disclosure, and treatment progress: A study of psychotherapy dyads*. University of Maryland, College Park.

Tal Alfi-Yogev, Ilanit Hasson-Ohayon, Gal Lazarus, Sharon Ziv-Beiman, and Dana Atzil-Slonim. 2021. When to disclose and to whom? examining within- and between-client moderators of therapist self disclosure-outcome associations in psychodynamic psychotherapy. *Psychotherapy Research*, 31(7):921–931.

Tal Alfi-Yogev, Yogev Kivity, Dana Atzil-Slonim, Adar Paz, Libby Igra, Adi Lavi-Rotenberg, and Ilanit Hasson-Ohayon. 2024. Transdiagnostic effects of therapist self-disclosure on diverse emotional experiences of clients with emotional disorders and schizophrenia. *Journal of Clinical Psychology*.

Tal Alfi-Yogev, Yogev Kivity, Ilanit Hasson-Ohayon, Sharon Ziv-Beiman, Ido Yehezkel, and Dana Atzil-Slonim. 2023. Client-therapist temporal congruence in perceiving immediate therapist self-disclosure and its association with treatment outcome. *Psychotherapy Research*, 33(6):704–718.

Cristelle T Audet. 2011. Client perspectives of therapist self-disclosure: Violating boundaries or removing barriers? *Counselling Psychology Quarterly*, 24(2):85–100.

Cristelle T Audet and Robin D Everall. 2010. Therapist self-disclosure and the therapeutic relationship: A phenomenological study from the client perspective. *British Journal of Guidance & Counselling*, 38(3):327–342.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Laura S Brown and Lenore EA Walker. 1990. Feminist therapy perspectives on self-disclosure. In *Self-disclosure in the therapeutic relationship*, pages 135–154. Springer.

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

James FT Bugental. 1965. The search for authenticity: An existential-analytic approach to psychotherapy. *(No Title)*.

Jie Cao, Michael Tanana, Zac E Imel, Eric Poitras, David C Atkins, and Vivek Srikumar. 2019. Observing dialogue in therapy: Categorizing and forecasting behavioral codes. *arXiv preprint arXiv:1907.00326*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Windy Dryden. 1990. Self-disclosure in rational-emotive therapy. In *Self-disclosure in the therapeutic relationship*, pages 61–74. Springer.

Morris N Eagle. 2011. *From classical to contemporary psychoanalysis: A critique and integration*, volume 70. Taylor & Francis.

Barry Alan Farber. 2006. *Self-disclosure in psychotherapy*. Guilford Press.

Arthur Freeman, James Pretzer, Barbara Fleming, and Karen M Simon. 1990. *Clinical applications of cognitive therapy*. Springer.

S Freud. 1912. Recommendation to physicians practicing psycho-analysis. *Standard Edition*, 12.

Jairo N Fuertes, Michael Moore, and Jennifer Ganley. 2019. Therapists' and clients' ratings of real relationship, attachment, therapist self-disclosure, and treatment progress. *Psychotherapy Research*, 29(5):594–606.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673.

Max Glockner, Vered Shwartz, and Yoav Goldberg. 2018. Breaking nli systems with sentences that require simple lexical inferences. *arXiv preprint arXiv:1805.02266*.

Marvin R Goldfried, Lisa A Burckell, and Catherine Eubanks-Carter. 2003. Therapist self-disclosure in cognitive-behavior therapy. *Journal of clinical psychology*, 59(5):555–568.

Eda G Goldstein. 1997. To tell or not to tell: The disclosure of events in the therapist's life to the patient. *Clinical Social Work Journal*, 25(1):41–58.

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. 2021. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15262–15271.

Jennifer R Henretty, Joseph M Currier, Jeffrey S Berman, and Heidi M Levitt. 2014. The impact of counselor self-disclosure on clients: A meta-analytic review of experimental and quasi-experimental research. *Journal of Counseling Psychology*, 61(2):191.

Jennifer R Henretty and Heidi M Levitt. 2010. The role of therapist self-disclosure in psychotherapy: A qualitative review. *Clinical psychology review*, 30(1):63–77.

Clara E Hill. 1978. Development of a counselor verbal response category. *Journal of Counseling Psychology*, 25(5):461.

Clara E Hill. 2009. *Helping skills: Facilitating, exploration, insight, and action*. American Psychological Association.

Clara E Hill, Charles J Gelso, Harold Chui, Patricia T Spangler, Ann Hummel, Teresa Huang, John Jackson, Russell A Jones, Beatriz Palma, Avantika Bhatia, et al. 2014. To be or not to be immediate with clients: The use and perceived effects of immediacy in psychodynamic/interpersonal psychotherapy. *Psychotherapy Research*, 24(3):299–315.

Clara E Hill and Sarah Knox. 2001. Self-disclosure. *Psychotherapy: Theory, Research, Practice, Training*, 38(4):413.

Clara E Hill, Sarah Knox, and Kristen G Pinto-Coelho. 2018. Therapist self-disclosure and immediacy: A qualitative meta-analysis. *Psychotherapy*, 55(4):445.

Kokil Jaidka, Iknoor Singh, Jiahui Liu, Niyati Chhaya, and Lyle Ungar. 2020. A report of the cl-aff offmychest shared task: Modeling supportiveness and disclosure. In *AffCon@ AAAI*, pages 118–129.

Divyansh Kaushik, Eduard Hovy, and Zachary C Lipton. 2019. Learning the difference that makes a difference with counterfactually-augmented data. *arXiv preprint arXiv:1909.12434*.

Sarah Knox, Shirley A Hess, David A Petersen, and Clara E Hill. 2001. A qualitative analysis of client perceptions of the effects of helpful therapist self-disclosure in long-term therapy. In *Annual Meeting of the Society for Psychotherapy., Jun, 1996, Amelia Island, FL, US; A version of this article was mentioned in the aforementioned conference.* American Psychological Association.

Sarah Knox and Clara E Hill. 2003. Therapist self-disclosure: Research-based suggestions for practitioners. *Journal of clinical psychology*, 59(5):529–539.

JA Kottler. 2003. On being a therapist . hoboken.

Heidi M Levitt, Takuya Minami, Scott B Greenspan, Jae A Puckett, Jennifer R Henretty, Catherine M Reich, and Jeffery S Berman. 2018. How therapist self-disclosure relates to alliance and outcomes: A naturalistic study. In *Disclosure and Concealment in Psychotherapy*, pages 7–28. Routledge.

James R Mahalik, E Alice Van Ormer, and Nicole L Simi. 2000. Ethical issues in using self-disclosure in feminist therapy.

Matteo Malgaroli, Thomas D Hull, James M Zech, and Tim Althoff. 2023. Natural language processing for mental health interventions: a systematic review and research framework. *Translational Psychiatry*, 13(1):309.

Patricia R McCarthy and Nancy E Betz. 1978. Differential effects of self-disclosing versus self-involving counselor statements. *Journal of Counseling Psychology*, 25(4):251.

Patricia McCarthy Veach. 2011. Reflections on the meaning of clinician self-reference: Are we speaking the same language? *Psychotherapy*, 48(4):349.

Nancy McWilliams. 2004. *Psychoanalytic psychotherapy: A practitioner's guide*. Guilford Press.

Sonia Murthy, Kiera Parece, Sophie Bridgers, Peng Qian, and Tomer Ullman. 2023. Comparing the evaluation and production of loophole behavior in humans and large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4010–4025.

Damilola Omitaomu, Shabnam Tafreshi, Tingting Liu, Sven Buechel, Chris Callison-Burch, Johannes Eichstaedt, Lyle Ungar, and João Sedoc. 2022. Empathic conversations: A multi-level dataset of contextualized conversations. *arXiv preprint arXiv:2205.12698*.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *NeurIPS*.

Jiaxin Pei and David Jurgens. 2020. Quantifying intimacy in language. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5307–5326, Online. Association for Computational Linguistics.

Kristen G Pinto-Coelho, Clara E Hill, Monica S Kearney, Elissa L Sarno, Elizabeth S Sauber, Sydney M Baker, Jennifer Brady, Glenn W Ireland, Mary Ann Hoffman, Patricia T Spangler, et al. 2018a. When in doubt, sit quietly: A qualitative investigation of experienced therapists' perceptions of self-disclosure. *Journal of Counseling Psychology*, 65(4):440.

Kristen G Pinto-Coelho, Clara E Hill, and Dennis M Kivlighan. 2018b. Therapist self-disclosure in psychodynamic psychotherapy: A mixed methods investigation. In *Disclosure and Concealment in Psychotherapy*, pages 29–52. Routledge.

Abhilasha Ravichander and Alan W Black. 2018. An empirical study of self-disclosure in spoken dialogue systems. In *Proceedings of the 19th annual SIGdial meeting on discourse and dialogue*, pages 253–263.

Ann-Katrin Reuel, Sebastian Peralta, João Sedoc, Garrick Sherman, and Lyle Ungar. 2022. Measuring the language of self-disclosure across corpora. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1035–1047.

Natalie Shapira, Oren Kalinsky, Alex Libov, Chen Shani, and Sofia Tolmach. 2023a. Evaluating humorous response generation to playful shopping requests. In *European Conference on Information Retrieval*, pages 617–626. Springer.

Natalie Shapira, Mosh Levy, Seyed Hossein Alavi, Xuhui Zhou, Yejin Choi, Yoav Goldberg, Maarten Sap, and Vered Shwartz. 2023b. Clever hans or neural theory of mind? stress testing social reasoning in large language models. *arXiv preprint arXiv:2305.14763*.

Natalie Shapira, Guy Zwirn, and Yoav Goldberg. 2023c. How well do large language models perform on faux pas tests? In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10438–10451, Toronto, Canada. Association for Computational Linguistics.

William B Stiles. 1979. Verbal response modes and psychotherapeutic technique. *Psychiatry*, 42(1):49–62.

Mina Valizadeh, Pardis Ranjbar-Noiey, Cornelia Caragea, and Natalie Parde. 2021. Identifying medical self-disclosure in online communities.

Andrew J Vandernoot. 2007. *The relationship between the attachment-style of therapists and their utilization of self-disclosure within the therapeutic relationship*. Ph.D. thesis, Alliant International University, Los Angeles.

Yi-Chia Wang, Robert E Kraut, and John M Levine. 2015. Eliciting and receiving online support: using computer-aided content analysis to examine the dynamics of online social support. *Journal of medical Internet research*, 17(4):e99.

C Edward Watkins Jr. 1990. The effects of counselor self-disclosure: A research review. *The Counseling Psychologist*, 18(3):477–500.

Anuradha Welivita and Pearl Pu. 2022a. Curating a large-scale motivational interviewing dataset using peer support forums. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3315–3330.

Anuradha Welivita and Pearl Pu. 2022b. Curating a large-scale motivational interviewing dataset using peer support forums. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3315–3330, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel S Weld. 2021. Polyjuice: Generating counterfactuals for explaining, evaluating, and improving models. *arXiv preprint arXiv:2101.00288*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A survey of large language models.

Sharon Ziv-Beiman. 2013. Therapist self-disclosure as an integrative intervention. *Journal of Psychotherapy Integration*, 23(1):59.

Sharon Ziv-Beiman, Giora Keinan, Elad Livneh, Patrick S Malone, and Golan Shahar. 2017. Immediate therapist self-disclosure bolsters the effect of brief integrative psychotherapy on psychiatric symptoms and the perceptions of therapists: A randomized clinical trial. *Psychotherapy Research*, 27(5):558–570.

Ofer Zur. 2004. To cross or not to cross: Do boundaries in therapy protect or harm. *Psychotherapy bulletin*, 39(3):27–32.