Language-Based Detection of Adherence to Evidence-Based Psychotherapy Scripts

Samuel T. Campione[1], Elizabeth C. Stade[1], Stefanie T. LoSavio[2], Shreya Singhvi[1], William Xuan[1], Phi Long Bui[3], Maria Martin Lopez[1], Shashanka Subrahmanya [1], Bailee B. Schuhmann[2], Courtney B. Worley[4], Shannon Wiltsey Stirman[1,5], Johannes C. Eichstaedt[1,6], H. Andrew Schwartz[3],

[1]Stanford University, [2]University of Texas Health Sciences Center at San Antonio, [3]Vanderbilt University, [4]U.S. Department of Veterans Affairs, [5]National Center for PTSD, VA Palo Alto Health Care System, [6]INSEAD

Correspondence: campione@stanford.edu

Abstract

Some psychotherapies, such as written exposure therapy for posttraumatic stress disorder, utilize "scripts" during parts of treatment, but verifying script adherence to ensure engagement of key mechanisms of change is a time-consuming step for therapy supervisors. Here, we formalize therapy script adherence as an NLP task, and evaluate several simple (text similarity) and more complex (few-shot LLM) approaches. Over 351 annotated therapist utterance-script pairs, we find text similarity approaches to be highly competitive with LLMs and produce fewer false positives. ROUGE-L recall achieves $F1 = 0.973$, and BLEU achieves $F1 = 0.972$ with full precision and zero false positives. GPT-5.2 achieves $F1 = 0.935$ and GPT-4o-mini achieves $F1 = 0.876$. Given that the text similarity techniques are multiple orders of magnitude less complex, our results underscore the ability of simpler NLP techniques to still be effective in the age of LLMs for tasks that are more textual in nature and suggest their utility in clinical training for giving supervisees rapid, privacy-preserving feedback on their delivery of evidence-based psychotherapy scripts.

1 Introduction

Mental health disorders are a leading cause of disability worldwide (GBD 2019 Mental Disorders Collaborators, 2022), and access to evidence-based psychotherapy (EBP) is limited by a shortage of trained providers (Kazdin and Blase, 2011). High-quality training improves implementation of EBPs and patient outcomes (Herschell et al., 2010), but traditional training in EBPs is difficult to scale. Digital training systems are emerging that allow therapists to practice delivering a complete course of an EBP with large language models (LLMs) prompted to behave as a patient receiving that treatment (Stade et al., 2025).

These digital training platforms are an ideal environment for providing high-quality feedback on clinicians' treatment delivery as they learn outside of traditional therapy consultation time. A core component of this feedback is assessing treatment fidelity, or how closely the clinician delivered the treatment as designed. However, fidelity monitoring is extremely time and cost-intensive, and there is a clear need for automated approaches.

Written exposure therapy (WET), an EBP for posttraumatic stress disorder (PTSD), is one such treatment well-suited for this type of digital training system and automated fidelity assessment. One aspect of WET's design is the delivery of prescribed language, or "scripts," to the patient word-for-word at specific points in the session. Thus, detecting verbatim script usage is important for assessing adherence to WET protocol.

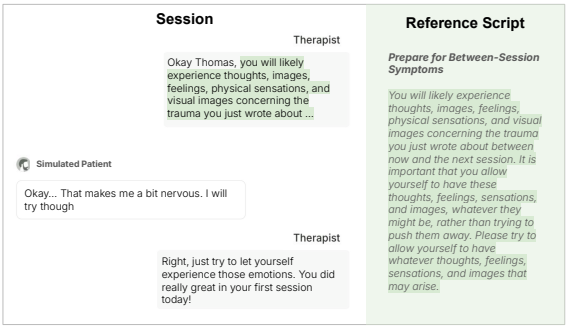

Figure 1: Illustrative example of a clinician delivering a script to a simulated patient. Green highlighting indicates script usage detected. The reference script is on the right.

Detecting correct script usage is not straightforward. As therapists embed these scripts into conversation, natural variations arise. An effective detection method should be able to distinguish between adherent verbatim script delivery and non-adherent paraphrasing or inclusion of proscribed content.

Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2026), pages 250–257

We frame script adherence detection as a binary classification task: given a therapist's utterance and a reference script, did the therapist use the script? In this study, we evaluate several text similarity metrics (ROUGE, BLEU) used in a threshold-based classifier and compare them to LLMs (GPT 5.2 and GPT 4o mini) on 351 therapist utterance-script pairs. We find that the simpler text similarity methods are comparable or superior to LLMs, and each approach has clear error patterns.

Our contributions include (1) formalizing verbatim script adherence detection in written exposure therapy as an NLP task, (2) demonstrating the potential of simple NLP methods for supporting supervisory feedback on script adherence in clinical training contexts, (3) providing evidence that text similarity methods may outperform LLM approaches and produce fewer false positives, and (4) showing that error patterns are linked to the differences between lexical and semantic similarity approaches.

2 Background

Written Exposure Therapy WET has a brief format and protocolized approach (Sloan and Marx, 2025), which makes it well-suited for this type of digital training system. Specifically, WET is a five-session intervention that targets avoidance through repeated exposure to the patient's traumatic event through structured writing sessions. Each session, the clinician completes a list of specific tasks as they guide a patient to write a narrative of their most distressing traumatic event.

Other EBPs provide content for clinicians to deliver in their own words; however, WET is unique in that scripts in the protocol are meant to be delivered to the patient word-for-word, with some trauma-specific details personalized to the patient. In Session 1 of WET, clinicians deliver five scripts at various points in the session. These scripts should be read verbatim because they contain carefully written language that engages key mechanisms of change in the patient that yield therapeutic value.

Treatment Fidelity Treatment fidelity is a core concern in delivery of evidence-based psychological practices (Perepletchikova and Kazdin, 2005). Greater treatment fidelity leads to better treatment outcomes (Farmer et al., 2017; Stirman et al., 2021). Manual fidelity coding requires trained coders to rate therapists' delivery utterance by utterance through session transcripts or recordings, a process that is time-intensive and faces scalability issues (Bacon et al., 2021; Worley et al., 2020).

Some recent advances in automated approaches include transformer models on the Lyssn platform to code fidelity to components of cognitive behavioral therapy (Coleman et al., 2024), and prototypes of LLM-assisted assessment of therapist adherence to cognitive processing therapy protocol (Held et al., 2025). Prior work focuses on broader clinical competence areas rather than narrow lexical tasks such as detection of verbatim script usage.

Text Similarity The task of determining whether a therapist's utterance contains a reference script maps nicely onto lexical text similarity metrics used in NLP evaluation. At first glance, this task may appear easily solvable with exact string-matching algorithms (e.g., Rabin-Karp, Knuth-Morris-Pratt), but these methods would fail to identify adherent script usages that contain appropriately personalized content, interspersed comments, or typical human spelling and punctuation errors. ROUGE (Recall-Oriented Understudy for Gisting Evaluation; Lin 2004) and BLEU (Bilingual Evaluation Understudy; Papineni et al. 2002) were developed for evaluating machine-generated summaries and machine translations. However, at a basic level both measure text overlap between a candidate text and a reference text (or a therapist utterance and a therapy script, respectively). In realistic clinical training settings, these metrics are more robust to the natural variation and personalization expected during script delivery in WET.

ROUGE-1 and ROUGE-2 measure unigram and bigram recall of the candidate text against the reference text. Here, recall is the proportion of the utterance's n-grams that appeared in the reference script. ROUGE-L measures the longest common subsequence (LCS) between the candidate and the reference. LCS is a classic NLP algorithm with well-known applications in bioinformatics and differencing tools. BLEU measures n-gram precision of the candidate against the reference, with a brevity penalty on shorter candidate texts. Here, precision is the proportion of the utterance's n-grams that appeared in the reference script.

In recent work, lexical metrics like ROUGE and BLEU have been replaced with LLM-based judges which can measure semantic equivalence (Kocmi and Federmann, 2023). Script adherence detection asks the opposite question – whether specific words

Category	Description	Example	Adherence
Verbatim delivery	Script read word-for-word with correct personalization where required; minor typos, lead-ins, and lead-outs are permitted	"Survivors of traumatic experiences often go through changes in their physical reactions..."	Yes
Verbatim with personalization issues	Script read word-for-word but with missing, insufficient, or unapproved personalization	"Survivors of assault and other traumatic experiences often go through changes..."	Yes
Partial delivery	Portions of script omitted; retained portions read verbatim (ranging from minor to extreme omission)	[first half of script delivered, second half skipped entirely]	Yes
Paraphrase	Script meaning conveyed in the therapist's own words without verbatim delivery	"After a traumatic event, survivors often experience changes in their thoughts, emotions, and behaviors..."	No
Non-script	Off-script content, general therapy conversation, or thematically related but non-script material	"What do you do when you feel those emotions or experience the flashbacks?"	No

Table 1: Script adherence categories with examples

appeared in the text, not whether the meaning was preserved. Thus, these classic lexical approaches are a natural fit.

3 Data

Scripts Session 1 of written exposure therapy contains five scripts that clinicians deliver at specific points during the session. Table 2 shows each script and its length. The scripts range from 67 to 417 words and cover distinct components of the treatment (e.g., psychoeducation, rationale for treatment, instructions for the writing task). Two of the five scripts (General Writing Instructions and Session 1 Writing Instructions) contain placeholders where clinicians should insert details personalized to the patient. The remaining three scripts are delivered fully verbatim.

Scripts	Words
Psychoeducation about PTSD	246
Avoidance + Introducing WET	417
General Writing Instructions	176
Session 1 Writing Instructions	242
End of Session Instructions	67

Table 2: WET script name and word count.

There are varying types of adherent script delivery in WET. These categories were developed in consultation with a written exposure therapy expert trainer (SL), who identified distinct categories of adherence reflecting prototypical therapist behavior during script delivery portions. Adherent categories include verbatim delivery, verbatim delivery with personalization issues, and partial delivery. Non-adherent categories include paraphrasing and off-script content. See Table 1

Dataset We use data from transcripts of two clinical psychologists delivering Session 1 of WET using TherapyTrainer, a digital training platform that allows clinicians to practice delivering full sessions of WET in a chat with a simulated patient. Transcript data was formatted into 145 unique pairs of therapist utterances and reference scripts by pairing each utterance with each of the five scripts. This included negative examples formed by pairing utterances with other WET scripts, which share therapeutic language and thus pose challenging cases of lexical overlap. This approach reflects how an automated fidelity monitoring system would operate, with each utterance evaluated against every script to detect usage. Two WET experts with prior treatment fidelity training (BBS and CW) labeled each utterance in the transcripts as adherent or non-adherent to Session 1 scripts. Their inter-rater reliability was perfect ($\kappa = 1.00$), unsurprising for the simplicity of this narrowly-defined labeling task.

In high-fidelity delivery of WET, each script is delivered once per session, making examples of script usage relatively sparse despite the richness of the transcript data. This scarcity is compounded by the time-intensive nature of manual labeling. To ensure adequate coverage of linguistic variation and to introduce additional challenging cases, we augmented the dataset with examples of script usage exhibiting varying degrees of adherence (i.e., verbatim delivery, verbatim with personalization issues, partial delivery, paraphrasing; see Table 1) for each of the five scripts. These prototypical examples were authored and labeled by two clinicians trained in WET (BBS and ECS), yielding

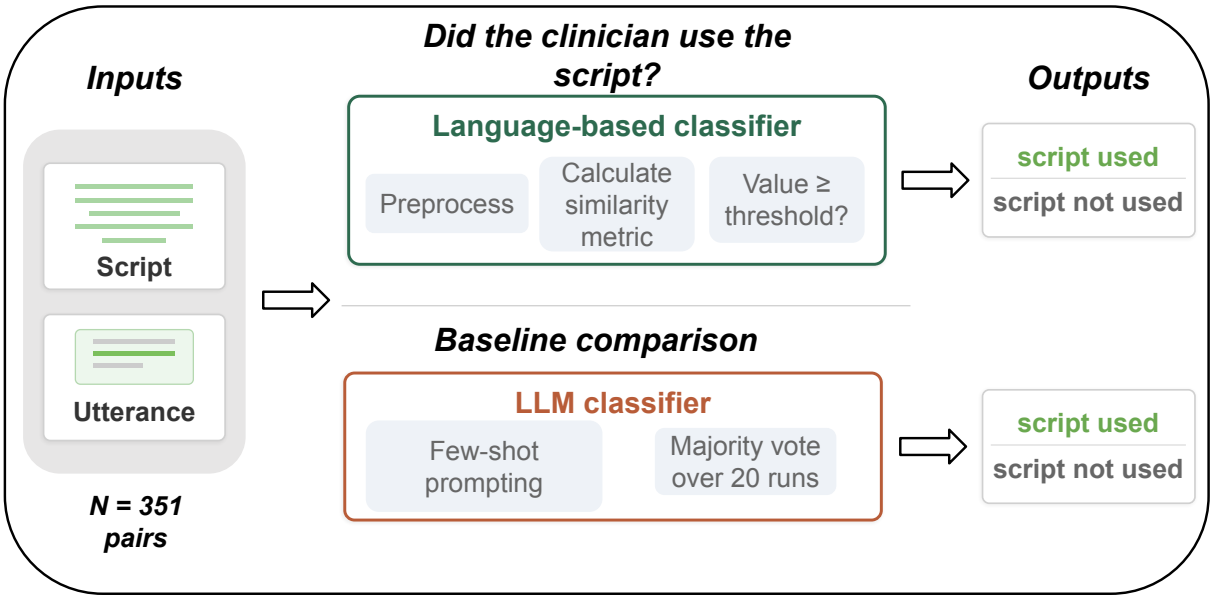

Figure 2: Overview of classification methods

206 additional therapist utterance–script pairs.

The final dataset contained 93 adherent (26%) and 258 non-adherent (74%) therapist utterance-script pairs with ground-truth script adherence labels. Average therapist utterance length is 122 words ($SD = 115$).

4 Methods

4.1 Text Similarity Based Classification

Given a reference script and a clinician utterance, we applied three steps: preprocessing, text similarity calculation, and threshold-based classification (Figure 2).

Preprocessing Both the reference script and the clinician utterance were normalized and tokenized. Each text similarity metric applied its own preprocessing through its respective Python library: the ROUGE metrics lowercase and strip punctuation, while BLEU preserves case and treats punctuation as separate tokens.

Text Similarity Calculation We computed text similarity between the reference script and the clinician utterance using four metrics. ROUGE-1 and ROUGE-2 recall measured the proportion of the script's unigrams and bigrams that appear in the utterance. ROUGE-L recall measured the proportion of the script recovered from the longest common subsequence in the utterance. BLEU measured the proportion of the utterance's n-grams (up to 4-grams) that appeared in the script, with a brevity penalty for short candidates. All four metrics produced scores ranging from 0 to 1 for each script-utterance pair, with higher values indicating greater textual overlap.

Classification Using Threshold To classify a clinician utterance as adherent the similarity score must have exceeded a decision threshold. We selected this threshold using 5-fold stratified cross-validation grouped by script. We stratified folds by script to ensure that utterances linked to the same script did not appear in both the training set and test set in a fold and prevent data leakage across folds. This design also allowed us to test whether the threshold generalized across scripts of different lengths and content.

Within each fold, a grid search was performed over 100 values from 0 to 1 on the training set only. The threshold that maximized training set *F1* was selected and applied to that fold's held-out test set – a threshold was optimized separately within each fold. This process repeated for each fold, and we computed final performance metrics by aggregating predictions from the held-out test sets. We evaluated each text similarity metric independently using this cross-validation procedure.

4.2 LLM Baselines

We compared against GPT-5.2 (OpenAI, 2026), a leading reasoning model, and GPT-4o-mini (OpenAI, 2024), a smaller model commonly used in annotation tasks (Tan et al., 2024). We prompted

each model with few-shot examples and instructions to determine whether a clinician utterance contains adherent script usage and return a binary label. For few-shot learning, we selected four examples spanning adherence categories shown in Table 1, including verbatim delivery, partial delivery, paraphrasing, and non-script content. We specified that some word substitutions and minor typographical errors are acceptable, but paraphrasing or unrelated therapy content constitutes non-adherence. To account for stochasticity, each utterance-script pair was evaluated across 20 independent runs. The final prediction was determined by majority vote.

4.3 Evaluation

We evaluate all methods against ground-truth adherence labels established in dataset creation. For the text similarity methods, we report performance metrics aggregated across held-out test folds from each metric's cross-validation process. For the LLM baselines, we report majority-vote metrics across 20 runs. We compare methods using McNemar's exact test.

5 Results

Classification performance of all methods is reported in Table 3. We include precision (how often detected script usage is correct), recall (how often true script usage is detected), and *F1* (the harmonic mean of the precision and recall).

Method	Precision	Recall	F1
ROUGE-L	0.989	0.957	0.973
BLEU	1.000	0.946	0.972
ROUGE-2	0.967	0.957	0.962
GPT-5.2	0.877	1.000	0.935
ROUGE-1	0.930	0.860	0.894
GPT-4o-mini	0.880	0.871	0.876

Table 3: Performance metrics. Text similarity metrics report held-out test fold metrics from 5-fold cross-validation. LLM methods report majority-vote metrics across 20 runs.

Text Similarity Metrics Three of the four text similarity metrics had higher *F1* than both LLM baselines. ROUGE-L had the highest *F1* (0.973) with a precision of 0.989 and a recall of 0.957. BLEU had comparable F1 (0.972) with full precision and zero false positives, though with slightly lower recall (0.946). ROUGE-2 followed closely (*F1* = 0.962), and ROUGE-1 was the weakest text similarity metric (*F1* = 0.894). The top text similarity metrics achieved near-perfect precision, rarely misclassifying non-adherent utterances as adherent.

LLM Baselines The LLM baselines showed a different pattern. GPT-5.2 achieved perfect recall, detecting every instance of script usage in the dataset, but produced 13 false positives, yielding *F1* = 0.935 and a precision of 0.877. GPT-4o-mini achieved lower overall performance (*F1* = 0.876), with errors in both directions (11 false positives, 12 false negatives). The two approaches thus exhibited a clear tradeoff: text similarity metrics favored precision while LLMs favored recall.

Statistical Comparisons The top two text similarity metrics, ROUGE-L and BLEU, both significantly outperformed GPT-4o-mini (McNemar's exact test, both $p < .001$). Neither ROUGE-L nor BLEU significantly outperformed GPT-5.2, though the text similarity and LLM approaches achieved comparable *F1* through markedly different error profiles, which we address in the "Error Analysis" section that follows.

6 Error Analysis

The text similarity metrics and LLM baselines showed distinct error profiles (Table 4).

Method	False Positives	False Negatives
ROUGE-L	1	4
BLEU	0	5
ROUGE-2	3	4
GPT-5.2	13	0
ROUGE-1	6	13
GPT-4o-mini	11	12

Table 4: False positive and false negative error breakdown of each approach.

Text Similarity Metrics Across all four text similarity metrics, each false negative was a case of partial delivery, in which the clinician read only a portion of the script (i.e., used only a fraction of the script, skipping entire paragraphs). Because so much text is missing, the similarity metrics do not cross the decision threshold even though the delivered portions of the script were verbatim. None of the text similarity metrics misclassified a fully verbatim or near-verbatim case. Paraphrases and unrelated therapy content never produced false positives

for ROUGE-L or BLEU – this follows logically from the underlying algorithms, as paraphrased content shared few n-grams or subsequences with the reference script.

LLM Baselines GPT-4o-mini produced errors in both directions. The 12 false negatives split into two failure modes. One was the misclassification of appropriately personalized scripts (6 cases). Clinicians read the script verbatim but replaced a few generic placeholders like "your trauma" with specific terms such as "the accident" or "the suicide," or added personalized examples of sights, sounds, smells, thoughts, or emotions to the script. This is considered high-quality clinical behavior, and the written exposure therapy protocol instructs therapists to personalize these segments of the script. The remaining false negatives were partial script deliveries (6 cases). The 11 false positives were all paraphrases or cases of thematically related but non-script content. The LLM detected semantic similarity and over-classified, treating utterances that conveyed only the script's meaning as verbatim delivery. This is the core failure mode one might expect from a model trained to encode meaning rather than surface form.

GPT-5.2 eliminated the false negative problem entirely (zero false negatives) but did not fix the false positives. The 13 false positives were paraphrases (6 cases) and unrelated therapy content (7 cases). This suggests even a more advanced LLM could not distinguish paraphrased content from verbatim delivery.

The core distinction here is that text similarity metrics failed only in ambiguous cases (extreme omission), whereas LLMs failed to distinguish between paraphrase and verbatim content.

7 Discussion

Results indicated that simple algorithms for text matching were competitive with frontier LLMs on a clinically meaningful task. Script usage detection superficially may appear to require a deep level of language understanding. The method must handle personalization, minor rewordings, partial delivery, inserted conversational speech, and the full variability of how clinicians deliver script content. However, the underlying question is lexical (whether specific words appeared in a specific order), and simple methods designed around that lexical question performed better.

The error profiles shed light on the strong performance of the text similarity approaches. Lexical text similarity algorithms operate on surface text, so they never mistake paraphrased text for verbatim delivery (paraphrased content shares only scattered single-word overlaps with the scripts). However, they fail to recognize partial delivery in some cases (e.g., a clinician reading 40% of a script verbatim and skipping the rest is still considered script use). LLMs operate on semantic meaning, which, in principle, should help them recognize personalized script deliveries that include appropriate patient-specific substitutions. However, the smaller LLM, GPT-4o-mini, misclassified these as non-adherent. Both LLMs also struggle with distinguishing between paraphrased scripts and verbatim scripts – which may arguably be what they are optimized to do. This suggests a mismatch for the task; LLMs are optimized to capture semantic equivalence while this task requires sensitivity to exact lexical text.

Although ROUGE-L and BLEU did not significantly outperform GPT-5.2, the approaches exhibited qualitatively different error profiles. False positives are the costlier error for this task. In a clinical training setting, a false negative flags a correct delivery for unnecessary review. By contrast, a false positive tells a clinician they delivered the script correctly when they did not, thus the clinician does not receive important feedback to help them learn how to deliver the treatment to patients as intended. The low false positive rates of the text similarity approaches reflect their conservative lexical design. They accept only utterances that closely match the script at the word level.

Our findings concern a specific type of task and do not imply that a lexical approach is universally preferable for fidelity, nor do they assess the competence aspect of treatment fidelity. For clinical behaviors that require understanding intent or meaning (e.g., assessing whether a therapist effectively validated a patient's emotions, or whether Socratic questioning was used), LLMs may be the better tool. There are several natural directions for future work. Structured testing and comparison of prompts and LLM models (e.g., different frontier models, reasoning vs. non-reasoning models) may reveal more advanced LLM capabilities. An agentic architecture in which an LLM agent selects an NLP method to apply could combine the strengths of both LLM and simpler, lexical approaches. Expanding beyond binary classification to multiple levels of adherence (e.g., verbatim, partial, para-

phrased) would likely better reflect how clinicians actually use scripts. While evaluated on written exposure therapy, in principle, the text similarity approaches presented here are not specific to this protocol. Such text similarity metrics might be tested on other protocolized therapies with verbatim script segments. We hypothesize that this method could also be adapted in reverse; for instance, in protocols like cognitive processing therapy, where therapists are expected to deliver content in their own words, a high text similarity score could flag insufficient personalization.

As this work used transcripts from clinicians practicing with simulated patients, an important next step is further validation with more naturalistic therapy transcripts. In therapy sessions with real patients, script delivery would likely include some areas of repetition or other delivery adjustments in response to the patient.

For detecting verbatim script delivery in a protocolized treatment, classic NLP text similarity methods are competitive and may even outperform LLMs, offering simplicity, interpretability, speed, and fewer false positives. Further, they do not require sharing HIPAA-protected data with external providers. Simple, interpretable language-based methods are appropriate for detecting therapist adherence to standardized language, a key aspect of therapist fidelity in some evidence-based treatments.

Limitations

Due to the resource-intensity of labeling transcripts, our dataset consisted of a limited number of therapist utterances. Further the data may not fully represent real-world clinical delivery variability since the transcripts come from practice sessions with a simulated patient rather than naturalistic therapy sessions. Additionally, the language data was provided by clinical psychologists and may be non-representative of therapists with other training backgrounds (e.g., social workers, counselors).

The evaluation was limited to a single treatment protocol, and we evaluated scripts from written exposure therapy Session 1 only; generalization to the remaining sessions or other treatments remains to be tested. We evaluated one LLM provider (OpenAI) on one prompt, thus other models or prompting strategies might yield different results. Finally, our method did not assess the quality of script personalization – it detected only whether the script was delivered, not whether personalization was appropriate.

Ethical Considerations

The dataset consisted of utterances from trained clinicians during encounters with simulated patients. The WET fidelity raters and the clinicians who provided the transcript data are all co-authors on this study. No patient data or actual therapy sessions were used. Therapy scripts were drawn from the published written exposure therapy protocol.

References

Katharine Bacon, Jane Marshall, Anna Caute, Katie Monnelly, Madeline Cruice, Corinne Moutou, and Celia Woolf. 2021. Treatment fidelity of technology-enhanced reading therapy (CommuniCATE) for people with aphasia. *International Journal of Language & Communication Disorders*, 56(6):1114–1131.

Jeremy J. Coleman, Jesse Owen, Jesse H. Wright, Tracy D. Eells, Becky Antle, Markessa McCoy, and Christina Signe Soma. 2024. Using artificial intelligence to identify effective components of computer-assisted cognitive behavioural therapy. *Clinical Psychology & Psychotherapy*, 31(6):e70023.

Courtney C. Farmer, Karen S. Mitchell, Kelly Parker-Guilbert, and Tara E. Galovski. 2017. Fidelity to the cognitive processing therapy protocol: Evaluation of critical elements. *Behavior Therapy*, 48(2):195–206.

GBD 2019 Mental Disorders Collaborators. 2022. Global, regional, and national burden of 12 mental disorders in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. *The Lancet Psychiatry*, 9(2):137–150.

Philip Held, Elizabeth C. Stade, Katherine Dondanville, and Shannon Wiltsey Stirman. 2025. Generative artificial intelligence in posttraumatic stress disorder treatment: Exploring five different use cases. *Journal of Traumatic Stress*, 38(5):813–820.

Amy D. Herschell, David J. Kolko, Barbara L. Baumann, and Abigail C. Davis. 2010. The role of therapist training in the implementation of psychosocial treatments: A review and critique with recommendations. *Clinical Psychology Review*, 30(4):448–466.

Alan E. Kazdin and Stacey L. Blase. 2011. Rebooting psychotherapy research and practice to reduce the burden of mental illness. *Perspectives on Psychological Science*, 6(1):21–37.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. *Preprint*, arXiv:2302.14520.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81. Association for Computational Linguistics.

OpenAI. 2024. GPT-4o system card. Technical report.

OpenAI. 2026. OpenAI GPT-5 system card. Technical report.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Francheska Perepletchikova and Alan E. Kazdin. 2005. Treatment integrity and therapeutic change: Issues and research recommendations. *Clinical Psychology: Science and Practice*, 12(4):365–383.

Denise M Sloan and Brian P Marx. 2025. *Written exposure therapy for PTSD: A brief treatment approach for mental health professionals*, 2nd edition. American Psychological Association.

Elizabeth C. Stade, Johannes C. Eichstaedt, Debra L. Kaysen, Aadesh Salesha, Alanna Greenberger, Shreya Singhvi, and Shannon Wiltsey Stirman. 2025. TherapyTrainer: Using AI to train therapists in written exposure therapy. *Cognitive and Behavioral Practice*.

Shannon Wiltsey Stirman, Cassidy A. Gutner, Jennifer Gamarra, Michael K. Suvak, Dawne Vogt, Clara Johnson, Jennifer Schuster Wachen, Katherine A. Dondanville, Jeffrey S. Yarvis, Jim Mintz, Alan L. Peterson, Stacey Young-McCaughan, and Patricia A. Resick. 2021. A novel approach to the assessment of fidelity to a cognitive behavioral therapy for PTSD using clinical worksheets: A proof of concept with cognitive processing therapy. *Behavior Therapy*, 52(3):656–672.

Zhen Tan, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. 2024. Large language models for data annotation and synthesis: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 930–957. Association for Computational Linguistics.

Courtney B. Worley, Stefanie T. LoSavio, Syed Aajmain, Craig Rosen, Shannon Wiltsey Stirman, and Denise M. Sloan. 2020. Training during a pandemic: Successes, challenges, and practical guidance from a virtual facilitated learning collaborative training program for written exposure therapy. *Journal of Traumatic Stress*, 33(5):634–642.

A Implementation

Experiments with text similarity required minimal compute, running on CPU and completed in less than a minute. The Python libraries used were `rouge-score 0.1.2` and `sacrebleu 2.6.0`. LLM services were accessed through Microsoft Azure OpenAI cloud services. GPT-4o-mini used temperature = 1.0, and GPT-5.2 used reasoning effort = low (temperature not configurable when reasoning is enabled). Code is available on GitHub.

B Prompt

```
# Instructions:
You are evaluating whether a specific script
was delivered verbatim or near-verbatim within
a user's message. The user's message may
contain additional speech before or after the
script. Your task is to determine whether the
script text appears within the message. Ignore
any surrounding content that is not part of
the script.

A match means the user read the script using
its actual words. The exact wording, or very
close to it, should be present somewhere in
their message. Minor typos, a few small word
substitutions, or some slight rewording are
acceptable. If they read part of the script
word-for-word but skipped other parts entirely,
that still counts as a match. Even extreme
omissions are acceptable as long as what they
did say uses the script's words.

A non-match means the user paraphrased the
script or only communicated the general idea
without using the scripted language. If they
restructure sentences, reorder information, or
use their own words instead of the script's
words, it's a non-match, even if the meaning
is accurate.

# Few shot examples
Script: {script1}
User message: {utterance1}
Response: {response1}

Script: {script2}
User message: {utterance2}
Response: {response2}

Script: {script3}
User message: {utterance3}
Response: {response3}

Script: {script4}
User message: {utterance4}
Response: {response4}

# Current case:
Script: {script}

Therapist's Message: {utterance}

Respond with only 1 (match) or 0 (non-match).
```