nature portfolio

Peer Review File

The valence of client and therapist language reflects symptom changes in messaging-based psychotherapy

Corresponding Author: Ms Henna Vartiainen

This file contains all editorial decision letters in order by version, followed by all author rebuttals in order by version.

Version 0:

Decision Letter:

** Please ensure you delete the link to your author homepage in this e-mail if you wish to forward it to your coauthors **

Dear Ms Vartiainen,

Thank you for your patience during the peer-review process. Your manuscript titled "Diverse approaches to sentiment analysis reliably reflect and explain symptom changes in psychotherapy" has now been seen by 3 reviewers, whose comments are appended below. You will see that they find your work of some potential interest. However, they have raised quite substantial concerns that must be addressed. In light of these comments, we cannot accept the manuscript for publication, but would be interested in considering a revised version that fully addresses these serious concerns.

We hope you will find the Reviewers' comments useful as you decide how to proceed. Should additional work allow you to address these criticisms, we would be happy to look at a substantially revised manuscript. If you choose to take up this option, please highlight all changes in the manuscript text file, and provide a detailed point-by-point reply to the reviewers.

Please bear in mind that we will be reluctant to approach the reviewers again in the absence of substantial revisions.

Editorially, we consider it crucial that the reviewers concern about sentiment analysis methods, including the treatment of sentiment and its potential limitations against other existing methods, are thoroughly addressed. Please also make sure that reviewer #1's request, separating validation of the approach from addressing the conceptual research question with validated measures, is addressed in the revised manuscript.

Please ensure you follow our statistical guidelines when reporting statistics (https://www.nature.com/commspsychol/submit/submission-guidelines#statistical-guidelines). Please note in particular our requirements for the reporting and interpretation of null-results. Non-significant findings derived from null-hypotheses significance tests should be reported in full, but may not be interpreted. Where you interpret null results, this interpretation must be based on Bayes Factors or equivalence tests.

I am attaching a checklist that details critical reporting requirements for the revised manuscript. Please attend to each item and ensure your manuscript is fully compliant. We are requesting that your manuscript aligns with these requirements as this facilitates the evaluation of your manuscript, reducing delays in re-review and potential future acceptance. If your revised manuscript is not aligned with these requests on major issues, such as those concerning statistics, it may be returned to you for further revisions without re-review. Additional information can be found in our style and formatting guide Communications Psychology formatting guide.

If the revision process takes significantly longer than five months, we will be happy to reconsider your paper at a later date, provided it still presents a significant contribution to the literature at that stage.

We would appreciate it if you could keep us informed about an estimated timescale for resubmission, to facilitate our planning.

We are committed to providing a fair and constructive peer-review process. Please do not hesitate to contact us if you wish to discuss the revision in more detail.

Please use the following link to submit your
- revised manuscript,
- point-by-point response to the referees' comments,
- cover letter (as a separate document),
- the Reporting Summary (see below), and
- the completed Editorial Request Table (attached):

Link Redacted

** This url links to your confidential home page and associated information about manuscripts you may have submitted or be reviewing for us. If you wish to forward this email to co-authors, please delete the link to your homepage first **

Thank you for the opportunity to review your work.

Best regards,

Troby Lui, on behalf of

Haiyan Wu

Troby Lui, PhD
Associate Editor
Communications Psychology

Haiyan Wu
External Editor
Communications Psychology

REVIEWER EXPERTISE:
Reviewer #1: mental health, digital intervention
Reviewer #2: mental health
Reviewer #3: mental health, intervention, sentiment analysis

REVIEWER REPORTS:

Reviewer #1 (Remarks to the Author):

This paper proposed to use sentiment analysis to serve as a marker of psychopathology and therapeutic progress, and 11 commonly used sentiment analysis measures were used on both client and therapist. The results indicated that the sentimanet assessments cane be used to measure therapeutic progress.
My only concern is that 11 sentiment analysis measures used different dictionaries, and some of them have tested validity, such as LIWC-22, but others have not. So I suggest the authors test the validity of each sentiment analysis systems, that is, the positive/negative prediction.

Reviewer #2 (Remarks to the Author):

This manuscript presents an impressive effort to examine sentiment in psychotherapy transcripts as it relates to internalizing symptoms over time. Using a large dataset (N = 6,229 clients) from a digital therapy platform, the authors compare 11 sentiment analysis methods and examine whether changes in sentiment—by client, therapist, or their divergence—are associated with symptom trajectories. The use of preregistration, a validation sample, and a rich text corpus are notable strengths. The work addresses a timely and important question: how language in therapy might help monitor or explain psychological change.
However, the manuscript suffers from two serious and related coherence issues that undermine the interpretability of the findings and weaken the overall contribution.
The first major issue is that the manuscript conflates two distinct research aims:(1) the validation and comparative evaluation of sentiment analysis methods, and (2) the use of sentiment as a theoretically meaningful psychological process variable to explain symptom change over time. These are both valuable goals—but they are methodologically and conceptually distinct and should not be collapsed into a single analysis pipeline.
• The validation goal asks: Which sentiment measures most reliably reflect valence in text messages and align with mental health outcomes? This is a methodological inquiry requiring attention to psychometric properties, predictive consistency, and construct validity—ideally with a clear criterion measure.
• The process-outcome goal asks: Does sentiment expressed in therapy track or drive change in client symptoms? This is a substantive psychological question about mechanisms of change, requiring theoretical justification and careful modeling of

temporal relationships.
In the current manuscript, these aims are not clearly distinguished in the rationale, methods, or discussion. This blurs the line between evaluating measurement quality and testing theoretical models of change, without justifying how the measure is sufficiently validated to be used in such a role. Moreover, conclusions about emotion expression, therapist positivity, and attunement are drawn without any theoretically justification (See Atzil Slonim et al., 2018, 2019).
This lack of distinction not only makes the analytic framework difficult to follow, but also undermines the interpretability of the findings. If the sentiment measures differ in what they capture—as the study itself shows—then how can we confidently treat one as a process variable in outcome mediation models? Conversely, if we are testing mediation, why are these measures not validated more rigorously first? (or why they validated only against symptoms measures?)
Closely tied to the issue above is a deeper conceptual inconsistency: the paper wavers between treating sentiment as a proxy for internalizing symptoms and as a measure of emotion or emotional expression. At times, the authors present sentiment as a marker or proxy of internalizing symptoms—an outcome measure that may offer scalable, unobtrusive estimates of distress. This framing suggests that language is reflective of psychopathology and might be used to assess symptoms. Indeed, the comparison of sentiment methods and their correlation with PHQ/GAD scores appears grounded in this logic of measurement validation.
However, elsewhere in the manuscript—most notably in the mediation analyses—sentiment is treated as a mechanism of change, implying it reflects emotional processes that cause or mediate symptom improvement. For instance, therapist positivity is interpreted as a potential therapeutic factor, and sentiment divergence is positioned as an indicator of emotional alignment or attunement. This shifts the conceptual role of sentiment from outcome marker to process variable—a major leap that is neither clearly acknowledged nor justified.
This dual treatment—sentiment as both a dependent variable (a proxy for internal symptoms) and an independent mediator (a driver of change)—is deeply problematic unless explicitly framed and supported. It raises serious concerns about the theoretical coherence and interpretability of the findings:
• If sentiment is a measure of symptoms, how can it be used to explain those same symptoms in mediation models?
• If sentiment is a measure of emotion, what justifies that interpretation, given the absence of independent emotion ratings?
• If it is both, where and how is this duality addressed analytically?
Without clarity on these points, the mediation findings risk circularity or tautological interpretation—where language is used to explain itself. The results cannot simultaneously validate sentiment as a measure of outcome and claim that the same sentiment explains that outcome via mediation without a clearer articulation of what sentiment represents and how it is grounded.
Additional Concerns
• Framing ambiguity: The manuscript weaves together multiple aims—method comparison, emotion expression, relational attunement, process-outcome research—without anchoring the reader in a single coherent narrative.
• Conceptual overreach: Claims about sentiment as reflecting emotional attunement or emotion regulation are speculative in the absence of emotion-specific measures.
• Terminological confusion: The results section describes a dataset of 6,229 text messages, while the methods clarify it includes 6,229 clients. This should be corrected.
Recommendations
To strengthen the manuscript, I strongly recommend the authors:
1. Clearly distinguish in the introduction and methods whether sentiment is being treated as a measure of symptoms, a measure of emotion, or a mechanism of change—and why.
2. Separate the validation of sentiment methods from theoretical modeling of therapeutic mechanisms, rather than treating them as one continuous analysis stream.
3. I also strongly encourage the authors to better situate their work within the rich and well-established literature on emotion in psychotherapy, emotional divergence between clients and therapists, and text-based analysis of therapeutic processes. There is a substantial body of work—both theoretical and empirical—that could help anchor the current study's framing, justify its constructs, and clarify its contributions.

Reviewer #3 (Remarks to the Author):

This manuscript presents an ambitious and large-scale investigation of sentiment analysis in online psychotherapy, comparing multiple sentiment analysis techniques across longitudinal psychotherapy data. With a primary sample of 3,729 units and a validation sample of 2,500, the study is well-powered and provides insights into the development of sentiment over time and between clients and therapists.

The manuscript’s major contribution lies in its comparative scope, examining a wide range of sentiment analysis methods in parallel. This is a significant strength, as it offers a rare opportunity to evaluate different approaches under identical conditions and yields valuable insights into their relative performance. However, this breadth also introduces considerable complexity, which makes it more difficult to follow the results at times. Nonetheless, the richness of this analysis is informative and valuable for future research aiming to select or refine sentiment-based approaches in psychotherapy contexts. However, the manuscript would benefit from important clarifications, methodological transparency, and a more up-to-date positioning in the context of current NLP developments.

- Transformer-Based Sentiment Models: A key limitation of the manuscript is the exclusive use of dictionary-based or machine-learning sentiment analysis methods. Recent advances in Natural Language Processing, particularly transformer-based models and large language models (LLMs), have led to significant improvements in semantic and affective text understanding. None of the evaluated approaches appear to use such architectures, which limits the relevance and innovation of the study. Ideally, the authors could include at least one transformer-based/LLM sentiment method to compare performance. If re-analysis is not feasible, the manuscript should more clearly acknowledge this limitation in the discussion, and critically reflect on how recent advances may outperform traditional approaches. Suggestions for future research in this direction would strengthen the paper.

- Please Clarify the Nature of the Dataset: The term “text-based psychotherapeutic exchanges” is a bit imprecise. Only later is it clarified that the data stems from text messages, not full session transcripts or spoken interactions. Please state this clearly in the abstract and introduction. Additionally, please add essential dataset information: How many patients and therapists participated? How many messages per case or session? Are the messages part of structured online therapy programs?

- Please Review Title and Conclusions: The title suggests a level of precision and causality that is not fully supported by the analysis. Specifically: “Reliably Reflect” implies a psychometric evaluation of reliability, which is not conducted. “Explain Symptom Changes” risks being interpreted as causal inference, which the design does not permit. The term “Psychotherapy” is too broad for a study limited to text-based interactions. A more accurate title could reference “online chat-based psychotherapy” or similar.

- Please Acknowledge Key Prior Work: Please consider citing relevant studies on sentiment analysis in actual psychotherapy contexts (e.g., Eberhardt et al., 2024; Lalk et al., 2025; Provoost et al., 2019; Shapira et al., 2021; Tanana et al., 2021). The claim that most prior work has focused on “laboratory or everyday contexts” is therefore a bit misleading. Including these works would situate the current study more accurately within the field. (1) https://doi.org/10.1080/10503307.2024.2322522 (2) http://dx.doi.org/10.3389/fpsyt.2025.1504306 (3) https://doi.org/10.3389/fpsyg.2019.01065 (4) https://doi.org/10.1037/cou0000440 (5) https://doi.org/10.3758/s13428-020-01531-z

- Please add Methodological Details: Were sentiment analyses run locally or using third-party servers? How were texts de-identified? How was missing data handled? Which software or packages were used for the mixed-effects models?

- Partial Accounting for Data Nesting: While the authors account for the nesting of text messages within clients (e.g., using mixed-effects models), they do not appear to model the nesting of clients within therapists. Please clarify whether therapists treated multiple clients. If so, please mention this as a limitation and unmodeled therapist effects could bias the results or inflate significance levels.

- Please Add Context on Treatment: Further information on the types of psychotherapy delivered (e.g., CBT, psychodynamic, blended care) would provide useful background.

- Descriptive Statistics: Please add a table summarizing basic statistics (N, M, SD, Min, Max, Skewness, Kurtosis, Alpha or Omega, Percentage Missing) for sentiment and outcome variables in the manuscript or supplement.

- Please Clarify Effect Size Interpretation: Effect sizes are described as “small,” but no criteria or benchmarks are provided. It would be helpful to reference standard conventions.

- Please Report Sentiment Measure Correlations: The manuscript does not report correlations between the different sentiment measures. Including such information would allow readers to assess conceptual overlap or divergence among the tools.

- Optional Visualization of Sentiment and Outcome Trajectories: It could be nice to include illustrative visualizations of how sentiment develops over time, ideally in relation to outcome measures.

- Caution on Claims About Psychopathology Tracking: The conclusion that sentiment can track internalizing psychopathology may be overstated. While promising, sentiment scores probably cannot substitute established psychometric questionnaires.

- Please Introduce NLP Terminology: Please consider introducing the term Natural Language Processing (NLP) in the introduction.

- Typo: On page 39, please correct “listed above below”.

EDITORIAL POLICIES

We ask that you ensure your manuscript complies with our editorial policies and reporting requirements.

To that end, we require revised manuscripts to be accompanied by a completed item: a reporting summary that collects information on study design and procedure.

- Nature Research Reporting Summary

Your revised manuscript can only be sent back to the referees if this checklist is completed and uploaded with the revision.

Notes: If you have submitted a Stage 1 Registered Report, Review, Primer, Comment, or Perspective you do not need to submit these forms. If you have already submitted these forms, you may disregard this request.

* TRANSPARENT PEER REVIEW: Communications Psychology uses a transparent peer review system. This means that we publish the editorial decision letters including Reviewers' comments to the authors and the author rebuttal letters online as a supplementary peer review file. However, on author request, confidential information and data can be removed from the published reviewer reports and rebuttal letters prior to publication. If your manuscript has been previously reviewed at another journal, those Reviewers' comments would not form part of the published peer review file.

** Visit Nature Research's author and referees' website at www.nature.com/authors for information about policies, services and author benefits**

Communications Psychology is committed to improving transparency in authorship. As part of our efforts in this direction, we are now requesting that all authors identified as ‘corresponding author’ create and link their Open Researcher and Contributor Identifier (ORCID) with their account on the Manuscript Tracking System prior to acceptance. ORCID helps the scientific community achieve unambiguous attribution of all scholarly contributions. You can create and link your ORCID from the home page of the Manuscript Tracking System by clicking on ‘Modify my Springer Nature account’ and following the instructions in the link below. Please also inform all co-authors that they can add their ORCIDs to their accounts and that they must do so prior to acceptance.
https://www.springernature.com/gp/researchers/orcid/orcid-for-nature-research

For more information please visit http://www.springernature.com/orcid

If you experience problems in linking your ORCID, please contact the Platform Support Helpdesk.

Version 1:

Decision Letter:

** Please ensure you delete the link to your author homepage in this e-mail if you wish to forward it to your coauthors **

Dear Ms Vartiainen,

Your manuscript titled "The valence of client and therapist language reflects symptom changes in messaging-based psychotherapy: Evidence from 13 sentiment measures" has now been seen by our reviewers, whose comments appear below. In light of their advice I am delighted to say that we are happy, in principle, to publish a suitably revised version in Communications Psychology.

We therefore invite you to revise your paper one last time to address the remaining concerns of our reviewers and a list of editorial requests. At the same time we ask that you edit your manuscript to comply with our format requirements and to maximise the accessibility and therefore the impact of your work.

EDITORIAL REQUESTS:

Please review our specific editorial comments and requests regarding your manuscript in the attached "Editorial Requests Table". Please outline your response to each request in the right hand column. Please upload the completed table with your manuscript files as a Related Manuscript file.

If you have any questions or concerns about any of our requests, please do not hesitate to contact me.

SUBMISSION INFORMATION:

In order to accept your paper, we require the files listed here https://www.nature.com/documents/commsj-file-checklist.pdf .

OPEN ACCESS:

Communications Psychology is a fully open access journal. Articles are made freely accessible on publication. For further information about article processing charges, open access funding, and advice and support from Nature Research, please visit https://www.nature.com/commspsychol/open-access

At acceptance, you will be provided with instructions for completing the open access licence agreement on behalf of all authors. This grants us the necessary permissions to publish your paper. Additionally, you will be asked to declare that all required third party permissions have been obtained, and to provide billing information in order to pay the article-processing charge (APC).

* TRANSPARENT PEER REVIEW: Communications Psychology uses a transparent peer review system. On author request, confidential information and data can be removed from the published reviewer reports and rebuttal letters prior to publication. If you are concerned about the release of confidential data, please let us know specifically what information you would like to have removed. Please note that we cannot incorporate redactions for any other reasons.

* CODE AVAILABILITY: All Communications Psychology manuscripts must include a section titled "Code Availability" at the end of the methods section. We require that the custom analysis code supporting your conclusions is made available in a publicly accessible repository at this stage; please choose a repository that generates a digital object identifier (DOI) for the code; the link to the repository and the DOI must be included in the Code Availability statement. Publication as Supplementary Information will not suffice.

* DATA AVAILABILITY:
All Communications Psychology manuscripts must include a section titled "Data Availability" at the end of the Methods section. More information on this policy, is available in the Editorial Requests Table and at http://www.nature.com/authors/policies/data/data-availability-statements-data-citations.pdf.

Please use the following link to submit the above items:
Link Redacted
** This url links to your confidential home page and associated information about manuscripts you may have submitted or be reviewing for us. If you wish to forward this email to co-authors, please delete the link to your homepage first **

We hope to hear from you within two weeks; please let us know if you need more time.

Best regards,

Troby Lui

Troby Lui, PhD
Associate Editor
Communications Psychology

REVIEWERS' COMMENTS:

Reviewer #2 (Remarks to the Author):

First, I want to sincerely commend the authors for their extraordinary effort in responding to the my and the other reviewers' comments. The response letter is thorough, thoughtful, and genuinely impressive, the additions of DistilBERT and Llama, the validation analyses, the RI-CLPMs, and the reconceptualization of sentiment as a marker rather than a mechanism all represent substantial work.

I have a few more comment, though the revised manuscript is clearly improved.
The manuscript itself still feels difficult to follow, likely because the revisions were integrated into the existing structure rather than used as an opportunity to reorganize. I'd strongly encourage the authors to consider restructuring the introduction around the three goals they articulate so well in their response letter (1) validating sentiment measures using prelabeled sentences, 2)
comparing how different sentiment measures track symptom changes in therapy, and 3) testing
whether changes in sentiment precede changes in symptoms (or vice versa)). This would give the reader a much clearer roadmap before entering the methods.

As one concrete example: the second paragraph currently moves from clinical settings to non-clinical examples and back to clinical settings, and the section on therapist language mixes several distinct rationales, therapist perception, empathic responding, and intervention techniques, without clearly distinguishing between them. These feel like artifacts of successive revisions rather than a coherent argument. Similarly, the paragraph discussing whether sentiment precedes or mirrors symptoms opens and closes with almost the same sentence, which makes it read as circular. A cleaner structure would help the reader understand what question is being asked and why before being walked through the evidence.

On a methodological point: the paper collapses PHQ-8 and GAD-7 into a single internalizing composite throughout all analyses. While this is justified by their correlation in this dataset, depression and anxiety have meaningfully different affective profiles, particularly regarding positive affect, which is more specifically disrupted in depression than anxiety. Given that the paper examines both positive and negative sentiment separately, it seems worth at least a brief sensitivity analysis or discussion of whether the sentiment-symptom relationships look different for depression versus anxiety specifically.

Finally, a concern about how divergence is operationalized. The authors frame divergence as capturing therapist responsiveness (attunement, empathic responding) but because it's calculated by subtracting a client's 3-week average from a therapist's 3-week average, the metric completely loses the temporal dynamics that would make it meaningful. A therapist who is wildly out of sync, highly positive when the client is most distressed, more negative when the client is doing better, produces the exact same aggregate divergence score as a therapist who closely tracks their client throughout. The averaging erases the very information the construct is supposed to capture.
One approach worth considering would be to calculate divergence at the message or session level first and then aggregate to the 3-week bin, which would at least preserve the dynamic information. If re-analysis isn't feasible, the authors need to be much more explicit than they currently are, the current divergence metric reflects a macro-level difference in overall tone over three weeks, not anything resembling moment-to-moment attunement or responsiveness, and the theoretical framing should be adjusted accordingly.

Reviewer #3 (Remarks to the Author):

Thank you for the opportunity to review the revised version of the manuscript. Overall, I appreciate the substantial revisions the authors have made in response to the previous reviews. In particular, I find it very positive that the authors addressed the earlier comment regarding the reliance on dictionary-based sentiment analysis methods. The inclusion of more recent approaches, including transformer-based models and a large language model (LLaMA), substantially strengthens the methodological scope of the study and provides an interesting comparison across different types of sentiment analysis techniques.

From my perspective, the manuscript has improved considerably and the authors have addressed the major concerns raised in the previous review. The analyses are clearly presented and the comparison of different sentiment analysis approaches is informative and relevant for the field.

I only have one minor suggestion regarding the interpretation of the LLM-based sentiment analysis. While I appreciate the inclusion of a large language model in the analyses, it may be helpful to briefly acknowledge in the discussion that the performance of general-purpose LLMs in such tasks depends strongly on several factors, including the specific model used, the model size (in this case an 8B model), and the prompt design. As indicated in the supplement, the prompt used for classification is relatively simple (e.g., “Classify the following messages as Positive (1) or negative (0)”). It might therefore be useful to briefly note that alternative prompting strategies or larger models could potentially yield different results. A short sentence acknowledging this dependency would help contextualize the LLM results without requiring additional analyses.

Apart from this minor point, I believe the manuscript is suitable for publication.

Henna Vartiainen
PhD Candidate
hennavartiainen@princeton.edu

Princeton University
Department of Psychology
Peretsman Scully Hall
Princeton, NJ 08540

February 19, 2026

Dr. Marike Schiffer
Chief Editor
Communications Psychology

Thank you very much for your invitation to revise and resubmit our paper, which has been renamed to "The valence of client and therapist language reflects symptom changes in messaging-based psychotherapy: Evidence from 13 sentiment measures". We sincerely appreciate the reviewers' comments, which we believe have considerably strengthened the manuscript. In particular, these revisions strengthen rigor and interpretability through validation against human-labeled sentiment, increase the manuscript's contemporary relevance by adding with more modern transformer-based and LLM-based methods, and reduce concerns about circularity through a more appropriate modeling framework. The reviewers' suggestions have also helped us better situate our work within the broader literature on emotion and language-based methods in therapeutic settings.

Attached you will find a detailed response to all comments and a revised version of the manuscript. We have noted all the changes and new additions in red text. We hope you find this revised paper suitable for publication in Communications Psychology. Please do not hesitate to contact us with any questions.

Sincerely,

Henna I. Vartiainen
PhD Candidate
Princeton University

Erik C. Nook
Assistant Professor of Psychology
Princeton University

Reviewer #1

1. This paper proposed to use sentiment analysis to serve as a marker of psychopathology and therapeutic progress, and 11 commonly used sentiment analysis measures were used on both client and therapist. The results indicated that the sentiment assessments can be used to measure therapeutic progress. My only concern is that 11 sentiment analysis measures used different dictionaries, and some of them have tested validity, such as LIWC-22, but others have not. So I suggest the authors test the validity of each sentiment analysis systems, that is, the positive/negative prediction.

We are grateful for the reviewer's positive response to our paper and agree that demonstrating the validity of our sentiment measures would make an important addition to the paper. We have validated our measures against a publicly available dataset of human-labeled sentences extracted from movie reviews (N = 68,219, Stanford Sentiment Treebank v2 (SST2)). We applied each of the sentiment measures we used in our study to this dataset, and we compared each measure's classification against human-annotated positive and negative labels. This dataset is frequently cited in sentiment analysis research, and so it is a good choice to address the reviewer's concern (Liu et al, 2023; Wankhade et al, 2022). We report these results in the revised manuscript (see updated Methods, Validation results, and Table 1), where we present classification metrics (precision, recall, F1-score) for each measure. Table 1 is presented below for convenience.

Table 1. Validation performance of sentiment analysis measures using human labeled sentences.

	Precision		Recall		F1-score		Mean F1	Ranking
	Negative	Positive	Negative	Positive	Negative	Positive		
AffectR	0.83	0.61	0.22	0.97	0.35	0.75	0.55	13
ANEW	0.81	0.65	0.35	0.94	0.49	0.77	0.63	12
BingLiu	0.79	0.84	0.78	0.84	0.79	0.84	0.82	4
DistilBERT*	0.93	0.93	0.91	0.95	0.92	0.94	0.93	1
LIWC	0.80	0.80	0.66	0.90	0.72	0.85	0.79	5
Llama	0.93	0.81	0.72	0.96	0.81	0.88	0.85	3
NB	0.65	0.75	0.70	0.70	0.67	0.72	0.70	10
NRC	0.77	0.76	0.65	0.85	0.70	0.80	0.75	7
SenticNet	0.78	0.74	0.62	0.82	0.69	0.80	0.75	8
SVM-C	0.68	0.79	0.75	0.72	0.72	0.75	0.74	9

SVM-R*	0.90	0.92	0.89	0.93	0.90	0.92	0.91	2
SWN	0.64	0.70	0.60	0.73	0.62	0.71	0.67	11
VADER	0.81	0.78	0.66	0.89	0.72	0.83	0.78	6

Notes: *F1-scores are presented separately for negative and positive sentiment classifications. They represent the harmonic mean of precision and recall for each class. Rankings were determined using the average of the positive and negative F1-scores (Mean F1) for each sentiment measure. * DistilBERT and our SVM-R model have been partially trained on the validation dataset, and thus those high F1s should be interpreted with caution.*

As you can see, validation performance was overall quite high (F1 > .7 for 10 of 13 measures and F1 > .5 for all 13 measures). However, performance also varied substantially across sentiment measures (min F1 = .55 and max F1 = .93). We summarize our inferences about these results in the paper as follows:

“We first report validation analyses testing how well each sentiment measure aligns with human perceptions of sentiment (see Table 1). Table 1 lists rankings when collapsing across negative and positive F1 scores, with DistilBERT, SVM-R, Llama, BingLiu, and LIWC performing best and SVM-C, Naive Bayes, SWN, ANEW, and AffectR performing worst. For overall model performance for negative sentiment, DistilBERT, Llama, and SVM-R achieved high performance (F1 > 80%); BingLiu, LIWC, NRC, SVM-C, and VADER performed moderately (F1 > 70%); Naive Bayes, SenticNet, and SWN performed poorly (F1 > 50%); and ANEW and AffectR performed very poorly (F1 < 50%). These weak performers had very low recall, suggesting that they tended to miss true negative cases. For positive sentiment, BingLiu, DistilBERT, LIWC, Llama, NRC, SenticNet, SVM-R, and VADER all achieved high performance (F1 > 80%); AffectR, ANEW, Naive Bayes, SVM-C, and SWN, performed moderately well (F1 > 70%). Weaker performance for positive sentiment appeared to be driven by lower precision scores, indicating over-application of positive labels. However, it is important to note that the two best performing models (DistilBERT, SVM-R) were partially trained on the same dataset that was used for these validation analyses. We further discuss this limitation in the Discussion section.“

Importantly, these results are consistent with our main analyses: Sentiment measures that performed poorly in the validation analyses had consistently lower performance in measuring within-subject change over time (SVM-C, Naive Bayes) or produced inconsistent results with the rest of the measures (AffectR, Naive Bayes). Thus, these validation analyses help contextualize our study’s main results: Variation in how well measures track symptoms may be related to how well they detect sentiment itself. We are grateful to the reviewers’ suggestion to include this validation step and believe it has strengthened the paper considerably. We now comment on these patterns and what they mean for interpreting the study’s primary results in our revised discussion:

“Importantly, the sentiment measures that showed the poorest performance in our validation analyses using labeled sentences also showed poor performance in within-subject change over time or produced inconsistent findings with the rest of the measures. Taken together, it appears that how well sentiment measures tracked symptoms in therapeutic conversations were related to how well they are able to detect sentiment itself."

Additionally, we include a short discussion flagging for readers that this validation dataset was used (alongside other datasets) for training two of the models we use (DistilBERT and SVM-R). Although this likely inflates their performance on the validation analyses, we don't believe this has undue influence on the overall inferences we draw from these validation analyses, as we find that these models *also* perform strongly in tracking symptom changes. This overlap in training and test datasets would be concerning if they implied they successfully tracked sentiment but then poorly predicted symptom changes. We nonetheless note this limitation in our revised manuscript:

"Conversely, we were initially concerned that the very high validation performance of the DistilBERT and SVM-R models could reflect the fact that they were partially trained on the dataset we use for validation. However, these models showed some of the strongest within-person correlations with internalizing symptoms. Thus, even if their validation performance is inflated, this does not unduly distort the overall conclusion that models who show high correspondence with human sentiment perceptions also strongly track symptom fluctuations. Nonetheless, we note this as a limitation of the exploratory validation analyses and caution overinterpreting the high F1s for DistilBERT and SVM-R."

Reviewer #2 (Remarks to the Author):

1. This manuscript presents an impressive effort to examine sentiment in psychotherapy transcripts as it relates to internalizing symptoms over time. Using a large dataset (N = 6,229 clients) from a digital therapy platform, the authors compare 11 sentiment analysis methods and examine whether changes in sentiment—by client, therapist, or their divergence—are associated with symptom trajectories. The use of preregistration, a validation sample, and a rich text corpus are notable strengths. The work addresses a timely and important question: how language in therapy might help monitor or explain psychological change.

We thank the reviewer for their positive remarks on the manuscript.

2. However, the manuscript suffers from two serious and related coherence issues that undermine the interpretability of the findings and weaken the overall contribution.The first major issue is that the manuscript conflates two distinct research aims: (1) the validation and comparative evaluation of sentiment analysis methods, and (2) the use of sentiment as a theoretically meaningful psychological process variable to explain symptom change over time. These are both valuable goals—but they are methodologically and conceptually distinct and should not be collapsed into a single analysis pipeline. The validation goal asks: Which sentiment measures most reliably reflect valence in text messages and align with mental health outcomes? This is a

methodological inquiry requiring attention to psychometric properties, predictive consistency, and construct validity—ideally with a clear criterion measure. The process-outcome goal asks: Does sentiment expressed in therapy track or drive change in client symptoms? This is a substantive psychological question about mechanisms of change, requiring theoretical justification and careful modeling of temporal relationships. In the current manuscript, these aims are not clearly distinguished in the rationale, methods, or discussion. This blurs the line between evaluating measurement quality and testing theoretical models of change, without justifying how the measure is sufficiently validated to be used in such a role.

This was a very generative comment that has helped us substantially strengthen our manuscript. We agree that (i) validating sentiment as clinically-relevant measures in therapy and (ii) testing whether sentiment acts as a therapeutic *process* are two very different goals both methodologically and conceptually, and we agree that our original manuscript did not sufficiently separate these goals. We also agree with the reviewer's point here and below that testing these questions simultaneously raises concerns about circular logic (i.e., a putative measure of symptoms is used to explain a putative measure of symptoms).

To remove this circularity, we have substantially revised the manuscript to focus primarily on the first of these goals: Testing how well sentiment measures correlate with symptom measures. We leave open the question of if/how sentiment reflects active therapeutic processes for future researchers to examine. We are also much clearer with the reader about the constructs we believe sentiment measures as well as the noisiness of these measures (i.e., the *multiple* processes that sentiment could reflect beyond those that we theorize to be the primary underlying constructs). We agree that this lack of construct specificity is another reason why we are not yet ready to tackle the latter question about therapeutic mechanisms.

To improve the flow and coherence of the manuscript, we have divided our goals, methods, and results into three sections: 1) validating sentiment measures using prelabeled sentences, 2) comparing how different sentiment measures track symptom changes in therapy, and 3) testing whether changes in sentiment precede changes in symptoms (or vice versa). These analyses converge on a focused argument that sentiment covaries with symptoms and may thus be an innocuous marker of symptoms, and we have removed our prior focus on testing sentiment as a mechanism of therapeutic change. In line with this new framing, we have moved the mediation results to the Supplementary Materials (Figures S1-S2). Because they were preregistered, we continue to report these analyses, but we refrain from interpreting them as evidence that sentiment is a mechanism for symptom change (as we originally argued).

Following Reviewer 1's suggestion above, we first test how well the different sentiment measures cohere with human-annotated valence labels. These validation tests shed light on the best and worst sentiment measures prior to relating them to clinical outcomes, and they provide valuable insights into measurement differences outside therapeutic contexts. These findings show that some measures are more aligned with pre-labeled sentiment ratings than others

(Table 1), and this may explain why some of the findings in the main analyses were inconsistent between measures.

Second, we use the same measures as in our original submission (with the addition of two new measures suggested by Reviewer 3: DistilBERT and Llama) to examine how sentiment changes over time and how well it relates to symptoms at between- and within-subjects levels. These results consistently demonstrate that sentiment is a valid marker of client symptoms, correlating with both overall case severity (with medium effect sizes) and also with changes in symptoms (with small effect sizes).

Third, we use random-intercept cross-lagged panel models (RI-CLPMs) to demonstrate that changes in sentiment can serve as an early marker of changes in internalizing symptoms. These analyses ask (i) whether changes in sentiment reliably precede changes in internalizing symptoms above and beyond between-subjects differences and (ii) whether changes in symptoms instead precede changes in sentiment. We find evidence for the former but not the latter pathway, suggesting that sentiment changes predict symptom scores and not the reverse. We are careful to frame these results as an examination of *temporal ordering* rather than tests of *causal mechanisms*, as we cannot be certain that sentiment itself serves as a mechanism of change (as the reviewer rightly points out). These results provide thus additional evidence that sentiment can be a linguistic marker that may forecast symptom trajectories and clarify questions about temporal ordering.

In the points that follow, we walk through specific adjustments summarized above and provide quotes demonstrating these edits.

2. Moreover, conclusions about emotion expression, therapist positivity, and attunement are drawn without any theoretical justification (See Atzil Slonim et al., 2018, 2019). This lack of distinction not only makes the analytic framework difficult to follow, but also undermines the interpretability of the findings. If the sentiment measures differ in what they capture—as the study itself shows—then how can we confidently treat one as a process variable in outcome mediation models? Conversely, if we are testing mediation, why are these measures not validated more rigorously first? (or why they validated only against symptoms measures?)

We appreciate the reviewer's concern regarding how we interpret the psychological construct underlying sentiment, and we agree that greater theoretical clarity on this point would strengthen our argument considerably. We have made several revisions to address this point (which overlap with Reviewer 2's additional points raised below).

First, we more clearly lay out what psychological processes we believe sentiment likely reflects, drawing on relevant research that provides initial validation of these claims. In particular, we consider client sentiment as an indirect linguistic measure of clients' emotional states, or how individuals experience and make sense of emotional events. The reviewer is correct that we do not have measures of clients' emotions, so we cannot validate this link in our own data.

Nonetheless, prior studies show that sentiment expressed in language is consistently associated with affective experiences (Kahn et al, 2007; Massachi et al, 2020; Munin et al, 2025). There is also substantial evidence showing that internalizing disorders are associated with more negative and less positive emotional states (Conrad et al, 2008; Mata et al, 2012; Joormann, 2016). Together, this literature suggests that sentiment provides a window into emotional states that are reliably related to increased internalizing symptoms.

That said, client sentiment *certainly* reflects multiple psychological processes. A key example is whenever people talk about negative/positive topics but their feelings about those events are not also negative/positive. For example, a client could use positive words when talking about a friend's wonderful news but actually feel envy rather than shared joy. Clients could also discuss a therapeutic skill (e.g., "looking on the bright side") without implementing that skill to feel differently. In addition, sentiment may capture normative expressions: for example, "sounds good!" is likely categorized as positive sentiment across measures, even though the speaker may not be feeling positively at that moment. Indeed, correlations between sentiment and emotional state are overall quite small (Munin et al, 2025), speaking to these gaps in alignment between sentiment and emotional state. Thus we interpret more negative sentiment as a noisy indirect linguistic marker of internalizing symptoms and are clear with the reader about the noisiness of this measure.

We now lay out this part of our theoretical model in the introduction:

Recent work suggests that sentiment (i.e., the valence of one's language) extracted from natural language maps onto a range of measures used to study emotion, such as self-reported and observer-reported affect, as well as facial cues of emotion (7). Indeed, sentiment has been shown to reliably reflect clinically meaningful psychological states. For example, the sentiment of everyday text correlates with self-reports of momentary emotional experiences (8) and self-reported mood (9), indicating that patterns of affective language covary with emotional experiences. Building on classic findings linking more positive language use to better health (10), recent findings have demonstrated that those with higher depressive symptoms tend to use more negative language when describing stressful life events, recent encounters with romantic partners, and collaborative experiences (11, 12, 13, 14). In therapeutic settings, clients' sentiment similarly tracks their emotional distress (15) and session-level sentiment aligns with both self- and therapist-reported emotions and treatment outcomes (16). Taken together, these results suggest that clients' language sentiment could index clients' emotional states (i.e., how they experience and interpret emotional events). Crucially, substantial evidence shows that internalizing disorders are characterized by elevated negative and reduced positive affect (*e.g.* 17, 18). Thus, here we test how strongly sentiment can assess internalizing disorders, presumably via these emotional states.

We also return to these points in our revised Discussion section.

"Recent work highlights that sentiment extracted from natural language is reliably associated with a range of measures used to study emotion, such as self-reported and observer-reported affect, as well as facial cues of emotion (7). However, the effect sizes of these associations vary considerably across datasets (*rs* = 0.07-0.73), and in some studies sentiment is not related with emotional experiences (78). Consequently, our findings should be interpreted with the understanding that sentiment is a noisy and indirect measure, rather than a direct signal of emotional experience. Indeed, there are many reasons why the links in our underlying theoretical model (sentiment measures exacerbated emotional experiences, which in turn reflect internalizing symptoms) may not hold. For example, clients may discuss valenced topics that do not reflect their internal state (e.g. describing a positive event without feeling happy), emotional states may not map onto psychopathology consistently (e.g. momentary increases in negative affect may indicate adaptive emotional processing rather than psychopathology), and language use may be shaped by social norms or communicative goals not directly related to one's emotional states (e.g. using conventional phrases such as "sounds good" to signal agreement without feeling any strong positive affect). As such, it is important to consider sentiment as an inherently imperfect measure of psychopathology or emotional states, likely explaining the modest effect sizes."

Turning to therapist sentiment, we likewise now explicitly state that we consider therapist sentiment as a measure of the overall valence of therapists' responses to clients' emotional states without characterizing it as any specific therapeutic process. Therapists have many techniques at their disposal (e.g., empathy, validation, encouragement, reappraisal, challenge, problem-solving, exploration). Sentiment cannot identify these techniques but it does capture the overall valence or affective tone of a therapist's response to their client. We consider this measure to be theoretically interesting for two reasons: (i) therapists may 'align' the valence of their language with those of their clients, and if the logic above holds, that would mean that the overall valence of therapists language may reflect their client's symptom severity. This would be a helpful discovery, to demonstrate that one can learn about the severity of a client's symptoms merely by studying the valence of their therapist's language. (ii) The extent of alignment between therapist and client valence may also reflect severity. Even though it is unclear what specific therapeutic/dyadic process this measure tracks, prior work shows that therapists systematically track and respond to clients' emotional valence, supporting the study of therapist-client sentiment alignment as a construct of interest. For example, Atzil-Slonim and colleagues (2019) showed that therapists accurately monitor their clients' positive and negative emotions, that their own emotions covary with clients' emotions, and that inaccuracy in assessing clients' positive emotions had implications for subsequent symptom change. Relatedly, therapists tend to adapt their verbal behavior to respond to clients' emotional states and needs: for example, through empathic and affectively attuned language (Elliot et al, 2011; Lord et al, 2015). Taken together, these findings suggest that therapists' responses to clients' emotional states are reflected, at least partially, in their language.

These two questions align with our study's research questions and analytic approach, focusing on whether client and therapist sentiment indirectly monitor client symptoms. We have less grounding literature to support claims linking therapist sentiment to any underlying psychological construct, and we agree that it likely reflects multiple processes. Thus, this aim is exploratory in nature, aiming to reveal associations that can then be further clarified in future work.

We are now clearer about our conceptualization of what therapist sentiment represents in the introduction:

"Understanding therapist sentiment is particularly valuable because of the dyadic nature of therapy. Previous work has shown that therapists accurately monitor their clients' positive and negative emotions and that their own emotions covary with their clients' (31). Beyond emotional experiences, therapists also tend to adapt their language to respond to clients' emotional states and needs (32, 33). As such, we see therapist sentiment as an overall reflection of how positively or negatively therapists are responding to clients. Therapists have many techniques at their disposal (e.g., reframing, validation, exploration), and although sentiment cannot determine which technique is being used, the overall valence of a therapist's approach is theoretically interesting. Here we test whether (i) therapist sentiment and (ii) alignment between therapist and client sentiment track client symptom severity in therapy."

As noted above, we also add an additional step of validating that our sentiment measures tend to align with a dataset of human-labeled sentences from movie reviews. This provides some assurance that our measures overall align with human-rated positive or negative labels, but they nonetheless leave open the question of whether people are *feeling* positively/negatively when using valenced language. We also find that different sentiment measures varied in their alignment with the benchmark labels in this validation step, which helps guide inferences in the therapy dataset. Thus, our validation findings give additional context to inconsistencies found in the linear mixed effects models while also providing stronger evidence that our sentiment measures are able to capture sentiment before being used to examine their relationship with symptom trajectories.

Finally, we agree that the conceptual confusion regarding what psychological processes sentiment reflects renders mediation analyses too hasty. We still report mediation models because we preregistered them, but we have moved them to the Supplement and caution readers from interpreting them as evidence that sentiment is a mechanism for symptom change.

"The following mediation analyses were preregistered and are reported for transparency. However, as sentiment is a coarse linguistic measure that reflects multiple psychological processes, these results should be interpreted with caution and not as definitive evidence that sentiment is itself a causal mechanism of symptom change (or reflects a single causal therapeutic process). Determining the underlying process(es) that explain these mediation models is an important area of future research."

3. Closely tied to the issue above is a deeper conceptual inconsistency: the paper wavers between treating sentiment as a proxy for internalizing symptoms and as a measure of emotion or emotional expression. At times, the authors present sentiment as a marker or proxy of internalizing symptoms—an outcome measure that may offer scalable, unobtrusive estimates of distress. This framing suggests that language is reflective of

psychopathology and might be used to assess symptoms. Indeed, the comparison of sentiment methods and their correlation with PHQ/GAD scores appears grounded in this logic of measurement validation. However, elsewhere in the manuscript—most notably in the mediation analyses—sentiment is treated as a mechanism of change, implying it reflects emotional processes that cause or mediate symptom improvement. For instance, therapist positivity is interpreted as a potential therapeutic factor, and sentiment divergence is positioned as an indicator of emotional alignment or attunement. This shifts the conceptual role of sentiment from outcome marker to process variable—a major leap that is neither clearly acknowledged nor justified. This dual treatment—sentiment as both a dependent variable (a proxy for internal symptoms) and an independent mediator (a driver of change)—is deeply problematic unless explicitly framed and supported. It raises serious concerns about the theoretical coherence and interpretability of the findings: If sentiment is a measure of symptoms, how can it be used to explain those same symptoms in mediation models? If sentiment is a measure of emotion, what justifies that interpretation, given the absence of independent emotion ratings? If it is both, where and how is this duality addressed analytically? Without clarity on these points, the mediation findings risk circularity or tautological interpretation—where language is used to explain itself. The results cannot simultaneously validate sentiment as a measure of outcome and claim that the same sentiment explains that outcome via mediation without a clearer articulation of what sentiment represents and how it is grounded.

We thank the reviewer for this extremely important critique. We agree that in our previous submission, treating sentiment as both a signal of internalizing and a mechanism of change made our mediation analyses circular.

To make our conceptualizations more consistent, we now more consistently frame sentiment as a marker of internalizing symptoms (via altered emotional states) rather than as a mechanism that causes symptom improvement. Edits reflecting this shift are described above.

To further support this framing, we have added an additional analysis to assess the temporal relationships between language and symptoms. In our original linear mixed effects models, sentiment over a ~3 week bin was used to predict internalizing symptoms reported at the end of that bin. This provides a weak temporal precedent where language → symptoms. However, given that our symptom measures ask for reports “over the last two weeks”, it is completely possible that language is reflecting changing symptoms (symptoms → language).

Thus, we further tackle the temporal relationship between these constructs using RI-CLPMs. These models examine how strongly language/symptoms in one ‘bin’ predict language/symptoms in the following ‘bin’, taking into account the nested structure of the data and autocorrelations within each measure. Interestingly, we find consistent evidence that language temporally precedes symptoms but not the reverse. This adds nuance to our results, as it suggests that linguistic shifts can be an ‘early marker’ or ‘forecast’ for changing symptom severity, and they somewhat reduce the concern that sentiment in one bin is “contaminated” by the prior symptom level. Importantly, we do not interpret temporal precedence as proof that changing one's language changes one's symptoms. Rather, we interpret sentiment as both a concurrent and prospective linguistic marker that may forecast symptom changes. This framing more clearly distinguishes sentiment as an indicator derived from language, and symptoms as more clinically grounded outcome measures.

These changes are reflected in the introduction:

"As discussed above, client sentiment, therapist sentiment, and their divergence may be meaningful indicators of clients' mental health. However, a key open question is whether different measures of sentiment merely reflect psychopathology, or whether changes in sentiment actually precede and potentially predict subsequent changes in symptoms. Previous research shows that sentiment correlates with concurrent symptom severity (11-14), and some evidence suggests that sentiment can predict at least short-term changes in internalizing symptoms in daily life (3, 4). However, it remains unclear whether sentiment, whether expressed by clients, therapists, or through their divergence, precedes or simply mirrors symptom change within therapeutic settings."

4. Framing ambiguity: The manuscript weaves together multiple aims—method comparison, emotion expression, relational attunement, process-outcome research—without anchoring the reader in a single coherent narrative.

We agree that the original framing was too broad. We now focus the manuscript on the question of whether sentiment can assess changes in client symptoms. As noted above, our revised paper focuses only on demonstrating that sentiment measures track and predict symptom changes in therapy. We have dropped the paper's focus on sentiment as a process/explanatory/therapeutic variable.

5. Conceptual overreach: Claims about sentiment as reflecting emotional attunement or emotion regulation are speculative in the absence of emotion-specific measures.

We agree that in the absence of emotion-specific measures, we cannot infer that sentiment reflects emotional attunement or emotion regulation per se. As noted above, we have made it clearer in the manuscript what we believe client and therapist sentiment measure (as well as the fact that these are highly imprecise). However, we agree that our interpretations of therapist language and client-therapist divergence as alignment processes were too hasty and would require validation to be interpretable as such. We more explicitly define what we mean by sentiment divergence with this new framework:

"Under our framing, this divergence reflects differences between the valence of clients' emotional discussions and the valence of therapists' responses to these topics. For example, sentiment divergence may reflect therapists' attunement with their clients' affect (31), empathetic responding (32), or efforts to help clients reinterpret their emotional reactions more positively (37)."

In addition, we discuss the sentiment divergence findings in more detail in our Discussion:

"Although sentiment divergence showed reliable associations with clients' internalizing symptoms using linear mixed effects models, we did not observe these effects in our RI-CLPMs. This pattern suggests that sentiment divergence, as measured in our study, may reflect concurrent symptoms rather than indicate longer-term therapeutic change. One possibility for these findings is that this measure captures multiple short-term conversational and therapeutic processes relating to clients' momentary emotional experiences without necessarily preceding sustained symptom change. For example, therapists' positive regard may temporarily uplift their clients and act as a motivational force (71), or positive affect during exposure therapy may significantly reduce immediate distress (e.g., 72). However, work on behavioral activation emphasizes that symptom reduction depends more on consistent behavioral changes rather than short-term affective improvements (73). In fact, recent work has demonstrated that even linguistic measures of behavioral activation, such as planning and engaging in enjoyable activities, often precede reduction in depressive symptoms (74). In addition, it is important to note that effective therapy involves not just promoting positive affect but also engaging with negative emotional states to facilitate better outcomes (e.g. 72; 75). From this perspective, lower sentiment divergence may reflect therapists' adaptive engagement with clients' negative experiences, rather than emphasizing an immediate shift toward positive affect. Taken together, sentiment divergence is a multifaceted linguistic marker that can reflect multiple psychological processes. Identifying which specific psychological and linguistic processes contribute to sentiment divergence could help us pinpoint how therapists can more effectively use their affective language to help reduce their clients' symptoms."

Finally, we now discuss the ambiguity of what sentiment divergence measures in our Limitations/Future Directions:

"Although we study therapist sentiment and client-therapist sentiment divergence, it is important to note that these findings should be interpreted cautiously. The sentiment constructs examined in our study are inherently noisy and may reflect multiple psychological processes. In the absence of direct measures of emotion, sentiment-based markers cannot be used to draw direct conclusions about emotion regulation, attunement, or experiences, and they need to be further validated in both experimental and naturalistic contexts."

6. Terminological confusion: The results section describes a dataset of 6,229 text messages, while the methods clarify it includes 6,229 clients. This should be corrected.

This has been corrected in the manuscript.

7 Recommendations. To strengthen the manuscript, I strongly recommend the authors: (i). Clearly distinguish in the introduction and methods whether sentiment is being treated as a measure of symptoms, a measure of emotion, or a mechanism of change—and why. (ii) Separate the validation of sentiment methods from theoretical

modeling of therapeutic mechanisms, rather than treating them as one continuous analysis stream. (iii). I also strongly encourage the authors to better situate their work within the rich and well-established literature on emotion in psychotherapy, emotional divergence between clients and therapists, and text-based analysis of therapeutic processes. There is a substantial body of work—both theoretical and empirical—that could help anchor the current study's framing, justify its constructs, and clarify its contributions.

We thank you for the recommendations. We have made substantial changes to our manuscript to address each point.

i) In the revised manuscript, we have more clearly communicated how we conceptualize each sentiment measure. We now treat client sentiment as a noisy linguistic marker of emotional expression that covaries with internalizing symptom severity. Therapist sentiment is considered to be a coarse marker of how therapists respond to this emotional expression (broadly, context-independently). Finally, sentiment divergence represents how the client expression and therapist response differ from one another. We now refrain from interpreting sentiment as a causal mechanism throughout the manuscript.

ii) We have also separated our methods and results into three sections: sentiment validation, sentiment changes and its relationship to concurrent symptom changes, and sentiment as an early marker of symptom changes.

iii) Finally, we have expanded our Introduction and Discussion to better situate our work within the literature on emotion and language-based methods in therapeutic settings (e.g. Atzil-Slonim et al, 2019; Eberhardt et al., 2024; Lalk et al., 2025; Munin et al, 2025). These additions have helped us justify our goals and constructs, and better communicate the contribution of our manuscript within this broader literature.

Reviewer #3

1. This manuscript presents an ambitious and large-scale investigation of sentiment analysis in online psychotherapy, comparing multiple sentiment analysis techniques across longitudinal psychotherapy data. With a primary sample of 3,729 units and a validation sample of 2,500, the study is well-powered and provides insights into the development of sentiment over time and between clients and therapists.

We thank the reviewer for their positive feedback on our manuscript.

2. The manuscript's major contribution lies in its comparative scope, examining a wide range of sentiment analysis methods in parallel. This is a significant strength, as it offers a rare opportunity to evaluate different approaches under identical conditions and yields valuable insights into their relative performance. However, this breadth also introduces considerable complexity, which makes it more difficult to follow the results at times.

Nonetheless, the richness of this analysis is informative and valuable for future research aiming to select or refine sentiment-based approaches in psychotherapy contexts. However, the manuscript would benefit from important clarifications, methodological transparency, and a more up-to-date positioning in the context of current NLP developments.

We thank the reviewer for their notes regarding our study's key advances, and we agree that (i) additional clarity regarding analyses and results and (ii) more focus on current NLP advances would strengthen the paper. Based on this and other reviewers' comments, we have majorly revised the introduction and analyses. To help with clarity, we have implemented the following revisions: 1) adding a section validating our 13 sentiment measures using prelabeled sentences, 2) better situating our research in the context of recent NLP advances in the introduction, 3) testing whether changes in language sentiment temporally precede symptom changes in therapy, and 4) adding LLM/transformer-based sentiment methods as analytic techniques. Together, these changes provide a more coherent narrative based on the question of how valid and clinically meaningful sentiment measures are in therapeutic settings.

We walk through many of these modifications in the points below. But here we illustrate how we revised the introduction to better situate our research within the growing NLP literature. We now write on pp. 2-3:

"Natural Language Processing (NLP), which is a branch of artificial intelligence focused on understanding and generating human language, has advanced significantly in the past years. These developments now allow us to examine the relationship between linguistic patterns and mental health on an unprecedented scale. For example, recent studies have analyzed millions of social media posts and messages to predict depression, anxiety, and even self-injurous thoughts (3, 4, 5, 6), demonstrating the power of analyzing language in predicting mental health related patterns."

"In therapeutic settings, clients' sentiment similarly tracks their emotional distress (15) and session-level sentiment aligns with both self- and therapist-reported emotions and treatment outcomes (16)."

"Although recent studies show that transformer-based models can perform substantially better than more traditional dictionary based measures in predicting human-rated sentiment from psychotherapy transcripts (22), agreement with human coders using iCBT messages remains only moderate (23) and even fine-tuned large language models achieve modest classification accuracy using psychotherapy transcripts (24)."

3. Transformer-Based Sentiment Models: A key limitation of the manuscript is the exclusive use of dictionary-based or machine-learning sentiment analysis methods. Recent advances in Natural Language Processing, particularly transformer-based models and large language models (LLMs), have led to significant improvements in semantic and affective text understanding. None of the evaluated approaches appear to use such

architectures, which limits the relevance and innovation of the study. Ideally, the authors could include at least one transformer-based/LLM sentiment method to compare performance. If re-analysis is not feasible, the manuscript should more clearly acknowledge this limitation in the discussion, and critically reflect on how recent advances may outperform traditional approaches. Suggestions for future research in this direction would strengthen the paper.

We are grateful that the reviewer raised this point and agree that the paper would benefit from including LLM sentiment methods. We have added both a transformer-based (DistilBERT) and LLM-based (Llama) sentiment measure to our analyses. Both measures showed strong relations with symptoms, aligning with (and at times outperforming) other sentiment measures. Neither measure diverged from the consistent pattern of results in our longitudinal analyses.

Example edits reflecting these updates include new parts of the methods section:

"We compared 13 different sentiment measures to examine which ones reflect changes in sentiment and clients' internalizing symptoms most reliably over time in therapy: 1) AffectR (44); 2) ANEW (Affective Norms for English Words; 45); 3) Bing Liu (46); 4) DistilBERT (47); 5) LIWC (Linguistic Inquiry and Word Count; LIWC-22, 48); 6) Llama (LLaMA-3.1-8B-Instruct model with 8-billion parameters) (49); 7) Naive Bayes; 8) NRC (50); 9) SenticNet (51); 10) SVM-C (Support Vector Machine - Classification); 11) SVM-R (Support Vector Machine - Regression); 12) SentiWordNet (52); 13) VADER (Valence Aware Dictionary and sEntiment Reasoner; 53). These sentiment measures take different theoretical approaches to sentiment (e.g. whether positive and negative emotions can co-occur), granularity (i.e., word- or sentence-level sentiment), and differ in their use of traditional lexicon-based approaches versus more advanced machine learning, transformer-based, or large language model methods. More detailed descriptions and information about obtaining sentiment scores for specific sentiment measures are described in Supplementary Materials."

In addition, the results have been updated to reflect these changes. The performance in of these models can be visualized in our updated Figure 1:

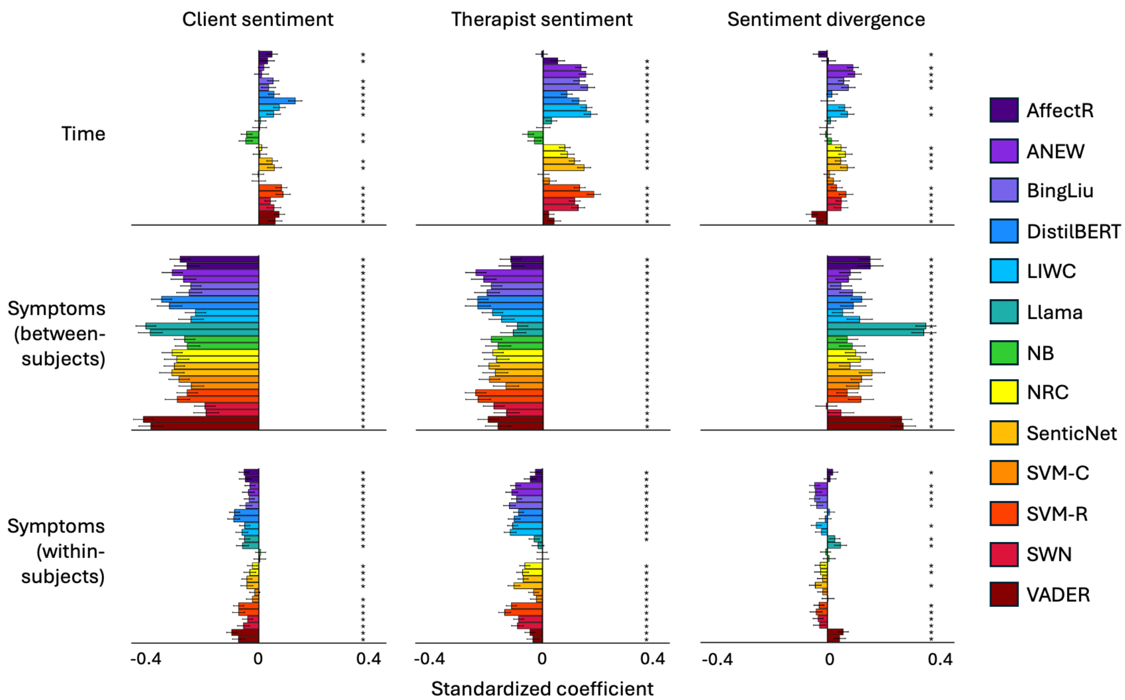


Figure 1. Results for mixed-effects models testing relations between sentiment measures and time (first row), between-subjects variance in internalizing symptoms (second row), and within-subjects variance in internalizing symptoms (third row). Colored bars denote different sentiment measures (see legend). Each sentiment method is presented twice: the upper bar shows results from the exploratory dataset, and the lower bar represents the validation dataset. The error bars represent 95% confidence intervals. If an error bar includes zero, the result is not statistically significant. Error bars and asterisks indicate significant findings after applying the Benjamini-Hochberg correction for multiple comparisons.

In light of these results, we have added a future directions subsection to the Discussion that elaborates on these modern sentiment methods. We note that although transformer-based and LLM-based approaches offer flexibility and contextual modeling advantages, our findings should be interpreted with caution due to non-independent validation dataset. We suggest that future work should examine if fine-tuning LLMs and training machine learning models directly on clinical data improves performance.

We now write on p. 15:

"Although recent developments in transformer- and LLM-based approaches offer improved flexibility and context-sensitivity that simpler lexicon-based approaches do not, our findings indicating their improved out-of-box performance on psychotherapy transcripts should be interpreted with caution. In particular, our transformer-based measure was partially trained on our validation data, making direct comparisons with well-validated lexicon-based methods inconclusive. Future work should test whether using fully independent test sets, fine-tuning models or training them on clinical data yields substantially improved performance."

4. Please Clarify the Nature of the Dataset: The term "text-based psychotherapeutic exchanges" is a bit imprecise. Only later is it clarified that the data stems from text messages, not full session transcripts or spoken interactions. Please state this clearly in the abstract and introduction.

We have replaced "text-based psychotherapeutic exchanges" with "messaging-based psychotherapeutic exchanges" in the abstract and introduction. To further ensure that readers are aware that Talkspace provides messaging-based psychotherapy, we have added a sentence accentuating this point in the introduction on p. 4: "To achieve this, we utilized a large dataset (N = 6,229) of text messages sent between clients and therapists using a messaging-based therapy platform (talkspace.com).".

5. Additionally, please add essential dataset information: How many patients and therapists participated? How many messages per case or session? Are the messages part of structured online therapy programs?

We agree this would be helpful for readers to see in the current paper. We originally did not report these descriptive statistics, as they are presented in the first paper on this publication (Nook et al., 2022). However, in our revised manuscript this information is provided in the newly added Descriptives table, which can be accessed in Supplementary Materials (Table S1).

Table S1. Sample and platform description.

	Full sample	Exploratory subsample	Validation subsample
Gender			
Female	4,742 (77.4%)	2,857 (77.7%)	1,885 (77.0%)
Male	1,306 (21.3%)	766 (20.8%)	540 (22.1%)
Transgender female	11 (0.2%)	8 (0.2%)	3 (0.1%)
Transgender male	11 (0.2%)	5 (0.1%)	6 (0.2%)
Gender queer	27 (0.4%)	21 (0.6%)	6 (0.2%)
Gender variant	6 (0.1%)	3 (0.1%)	3 (0.1%)
Other	20 (0.3%)	16 (0.4%)	4 (0.2%)
No response	106	53	53
Age[a]			
18-25	1,031 (22.2%)	606 (21.8%)	425 (22.8%)
26-35	2,536 (54.6%)	1,517 (54.6%)	1,019 (54.8%)
36-49	871 (18.8%)	526 (18.9%)	345 (18.5%)
50+	203 (4.4%)	131 (4.7%)	72 (3.9%)
No response	1,588	949	639
Race			
Caucasian	1,172 (60.4%)	698 (60.3%)	474 (60.5%)
African American	284 (14.6%)	179 (15.5%)	105 (13.4%)
Asian	140 (7.2%)	82 (7.1%)	58 (7.4%)
Hispanic	120 (6.2%)	66 (5.7%)	54 (6.9%)
Native American	5 (0.3%)	2 (0.2%)	3 (0.4%)
Other	195 (10.1%)	115 (9.9%)	80 (10.2%)

Declined to identify	24 (1.2%)	15 (1.3%)	9 (1.1%)
No response	4,289	2,572	1,717
Education level			
Less than high school	28 (0.5%)	14 (0.4%)	14 (0.7%)
High school	808 (15.7%)	477 (15.3%)	331 (16.2%)
Associate’s degree	78 (1.5%)	40 (1.3%)	38 (1.9%)
Some college no degree	200 (3.9%)	126 (4.0%)	74 (3.6%)
Bachelor’s degree	3,683 (71.4%)	2,238 (71.9%)	1,445 (70.7%)
Master’s degree	260 (5.0%)	163 (5.2%)	97 (4.7%)
Professional degree	43 (0.8%)	23 (0.7%)	20 (1.0%)
Doctoral degree	56 (1.1%)	31 (1.0%)	25 (1.2%)
No response	1,073	617	456
Symptom measures			
Baseline internalizing symptoms	M=22.21, SD=9.90	M_e=21.99, SD_e=9.89	M_v=22.54, SD_v=9.92
Final internalizing symptoms	M=15.16, SD=9.87	M_e=14.95, SD_e=9.79	M_v=15.48, SD_v=10.00
Baseline depression symptoms	M=11.03, SD=5.82	M_e=10.93, SD_e=5.84	M_v=11.18, SD_v=5.80
Final depression symptoms	M=7.56, SD=5.58	M_e=7.44, SD_e=5.54	M_v=7.73, SD_v=5.62
Baseline anxiety symptoms	M=11.18, SD=5.04	M_e=11.06, SD_e=5.02	M_v=11.36, SD_v=5.06
Final anxiety symptoms	M=7.60, SD=4.90	M_e=7.50, SD_e=4.83	M_v=7.75, SD_v=4.99
Therapy and text qualities			
Text-only subscription	6,108 (98.1%)	3,659 (98.1%)	2,449 (98.0%)
Number of client messages	759,706	455,379	304,327
Length of client messages (words)	M=80.81, SD=146.80	M_e=81.61, SD_e=147.02	M_v=79.61, SD_v=146.46
Number of therapist messages	461,911	273,208	188,703
Length of therapist messages (words)	M=82.69, SD=107.30	M_e=84.32, SD_e=108.00	M_v=80.33, SD_v=106.23
Present-tense verbs per message	M=8.73, SD=15.10	M_e=8.81, SD_e=15.11	M_v=8.61, SD_v=15.09
Past-tense verbs per message	M=3.87, SD=9.34	M_e=3.89, SD_e=9.34	M_v=3.83, SD_v=9.34
Future-tense verbs per message	M=0.82, SD=1.80	M_e=0.83, SD_e=1.81	M_v=0.81, SD_v=1.79
First-person singular pronouns per message	M=8.13, SD=14.50	M_e=8.17, SD_e=14.49	M_v=8.06, SD_v=14.51
Other pronouns per message	M=3.75, SD=9.38	M_e=3.78, SD_e=9.38	M_v=3.71, SD_v=9.38
Number of therapists	M=1.21, SD=0.51	M_e=1.21, SD_e=0.51	M_v=1.22, SD_v=0.52
Number of symptom measures	M=3.45, SD=0.51	M_e=3.45, SD_e=0.51	M_v=3.45, SD_v=0.51
Exactly 3 measures	3,453 (55%)	2,058 (55%)	1,395 (56%)
Exactly 4 measures	2,741 (44%)	1,651 (44%)	1,090 (44%)
Exactly 5 measures	35 (1%)	20 (1%)	15 (1%)
Days between start of therapy and final symptom measure	M=63.93, SD=11.90	M_e=63.94, SD_e=11.89	M_v=63.91, SD_v=11.92

Notes: Percentages ignore clients who did not respond to each demographic question. [a]Participants selected from pre-defined age bins.

Regarding the question about ‘structured online therapy programs’, therapists on Talkspace are licensed clinicians who provide psychotherapy. Talkspace does not prescribe a common structured therapeutic approach for all clients but instead provides a platform through which clients and therapists can work together via video and messaging technology. Thus, their practice on Talkspace likely reflects the approaches they would take if they worked face-to-face with clients. This likely varies across therapists, with some deploying a structured treatment, and others integrating many approaches.

To make this clearer to readers we have added the following information to our description of Talkspace on p. 4:

"Thus, therapist practice on Talkspace likely reflects the various approaches they would take if they worked face-to-face with clients (see Supplemental Materials for proportions therapist reporting expertise on common therapeutic approaches)"

Finally, we now report in Table S2 details of what proportion of therapists reported expertise in different modalities of psychotherapy (this is pasted in response to your point 10).

6. Please Review Title and Conclusions: The title suggests a level of precision and causality that is not fully supported by the analysis. Specifically: "Reliably Reflect" implies a psychometric evaluation of reliability, which is not conducted. "Explain Symptom Changes" risks being interpreted as causal inference, which the design does not permit. The term "Psychotherapy" is too broad for a study limited to text-based interactions. A more accurate title could reference "online chat-based psychotherapy" or similar.

We appreciate the reviewer's suggestions for ensuring the paper's title accurately reflects its contents. Based both on this comment and other reviewers' comments, we have revised the title to "**The valence of client and therapist language reflects symptom changes in messaging-based psychotherapy: Evidence from 13 sentiment measures**" We have maintained the use of the word "psychotherapy" (but specified that it is messaging-based), as Talkspace does indeed provide treatment with licensed therapists. However, we have revised the title, discussion, and conclusions to clarify that our results reflect messaging-based therapy, and may not extend to synchronous, in-person therapy.

For example, in the discussion section we now write:

"Importantly, our findings were obtained from messaging-based psychotherapy and should be interpreted within this context. Although messaging-based therapy is conducted by licensed therapists and includes aspects of in-person psychotherapy, it is important to note that its asynchronous and text-based format can shape both language and therapeutic processes differently. As such, future work should examine whether these results generalize to synchronous in-person or video-based therapy."

7. Please Acknowledge Key Prior Work: Please consider citing relevant studies on sentiment analysis in actual psychotherapy contexts (e.g., Eberhardt et al., 2024; Lalk et al., 2025; Provoost et al., 2019; Shapira et al., 2021; Tanana et al., 2021). The claim that most prior work has focused on "laboratory or everyday contexts" is therefore a bit

misleading. Including these works would situate the current study more accurately within the field.

We are grateful to the reviewer for listing these very relevant papers. We now include them in our literature review, and we have removed the claim that most prior work focuses on laboratory or everyday contexts. Relevant revisions include:

"In therapeutic settings, clients' sentiment similarly tracks their emotional distress (15) and session-level sentiment aligns with both self- and therapist-reported emotions and treatment outcomes (16)."

"Although recent studies show that transformer-based models can perform substantially better than more traditional dictionary based measures in predicting human-rated sentiment from psychotherapy transcripts (22), agreement with human coders using iCBT messages remains only moderate (23) and even fine-tuned large language models achieve modest classification accuracy using psychotherapy transcripts (24)."

8. Please add Methodological Details: Were sentiment analyses run locally or using third-party servers? How were texts de-identified? How was missing data handled? Which software or packages were used for the mixed-effects models?

We thank the reviewer for these important points. We have updated our Methods to clarify that all analyses were run locally (no third-party servers included), that data were provided fully deidentified from Talkspace, and that missing data were handled by the analytical framework itself (lmer, which we used, accommodates empty values, and we used the default maximum likelihood with robust standard errors for RI-CLPMs). We also specified the R version (version 4.5.1) and packages used for the mixed effects models (lme4; Bates et al, 2015), standardized coefficients and confidence intervals (parameters; Ludecke et al, 2021) and RI-CLPMs (lavaan; Rosseel, 2012) (pages 4-9).

For example, we now write:

"We received fully deidentified transcripts of interactions between clients and their therapists via text messages from Talkspace, the clients' scores on depression and anxiety measures, and clients' self-reported demographics for our analyses." (p. 4)

"All analyses were run locally without use of third-party servers, and missing data for the linear mixed effects models were handled by the analytical framework itself (lme4, version 1.1.37, 62). The linear mixed effects models were run in R (version 4.5.1) and standardized coefficients were obtained using the R package parameters (version 0.28.2, 63)." (p. 7)

"We implemented an exploratory Random-Intercept Cross-Lagged Panel Model (RI-CLPM) using the *lavaan* package in R (version 0.6.19, 64) to study the relationship between language sentiment and internalizing symptoms over time in therapy." (p. 7)

“Handling missing data for RI-CLPMs was achieved by using maximum likelihood with robust standard errors, which is robust to non-independent data, randomly missing data, and normality deviations (67, 68).“ (p. 8)

9. Partial Accounting for Data Nesting: While the authors account for the nesting of text messages within clients (e.g., using mixed-effects models), they do not appear to model the nesting of clients within therapists. Please clarify whether therapists treated multiple clients. If so, please mention this as a limitation and unmodeled therapist effects could bias the results or inflate significance levels.

This is an excellent point. Indeed, some of the therapists treated multiple clients. To make sure that this does not influence our results, we reran our analyses adding a random intercept for therapists. This addition did not change the significance or direction of any results presented in the main text of the original submission, with the exception of turning two marginally significant results to significant (therapist sentiment becoming more positive over time; SVM-C, VADER). We provide these findings in Supplementary Materials, Table S12.

We now report each of these points in the revised manuscript. On p. 10 we write: “Because approximately 56% of therapists treated more than one client, we reran our analyses including therapist as a random effect. This adjustment did not change the direction or the significance of our results, with the exception of turning two marginally significant results to significant (therapist sentiment becoming more positive over time for SVM-C and VADER, see Supplemental Materials).”

10. Please Add Context on Treatment: Further information on the types of psychotherapy delivered (e.g., CBT, psychodynamic, blended care) would provide useful background.

We appreciate the suggestion to provide additional context surrounding the therapeutic orientation of providers represented in our dataset. Therapists had the option to report their therapeutic expertise. We reviewed these data and found that 34% of therapists did not respond to this question, and the others reported a variety of approaches, including Cognitive Behavioral, Mindfulness-based, Dialectical, Psychodynamic, Motivational Interviewing, Somatic, Integrative, and Family Systems Therapy. It was also common for therapists to report using more than one approach (43% did so). We now report the percentage of therapists who reported using each modality (and multiple modalities) in Table S2.

Table S2. Therapist-reported expertise across datasets.

	N (%)
Unique therapists	2392
Expertise reported	1570 (65.64%)
Expertise missing	822 (34.36%)

More than 1 expertise	1023 (42.77%)
Cognitive behavioral	1175 (49.12%)
Mindfulness-based	436 (18.23%)
Motivational Interviewing	330 (13.80%)
Dialectical	316 (13.21%)
Family Systems	275 (11.50%)
Humanistic	236 (9.87%)
Psychodynamic	199 (8.32%)
Integrative	143 (5.98%)
Somatic	46 (1.92%)

11. Descriptive Statistics: Please add a table summarizing basic statistics (N, M, SD, Min, Max, Skewness, Kurtosis, Alpha or Omega, Percentage Missing) for sentiment and outcome variables in the manuscript or supplement.

We have added these descriptive measures for all sentiment measures and internalizing symptoms in both exploratory and validation datasets (please see Tables S3-S5).

Table S3. Descriptive statistics of internalizing symptoms.

Measure	N	M	SD	Min	Max	Skewness	Kurtosis	% missing
Exploratory dataset								
Symptoms	12823	17.56	10.11	0.00	45.00	0.50	-0.46	0.00
Validation dataset								
Symptoms	9130	17.85	10.26	0.00	45.00	0.46	-0.56	0.00

Table S4. Descriptive statistics of client unidimensional sentiment.

Measure	N	M	SD	Min	Max	Skewness	Kurtosis	% missing
Exploratory dataset								
AffectR	12650	0.15	0.02	-0.26	0.40	-0.33	20.70	1.35
ANEW	12643	5.90	0.21	2.10	7.77	-0.14	20.20	1.40
BingLiu	12650	0.012	0.03	-0.50	0.53	2.38	48.50	1.35

DistilBERT	12672	-0.88	0.40	-1.00	1.00	0.34	3.33	1.18
LIWC	12650	2.56	3.68	-22.22	100.00	4.90	71.60	1.35
Llama	12668	0.76	0.18	0.00	1.00	-1.00	4.77	1.21
NB	12650	0.40	0.22	0.00	1.00	0.47	3.32	1.35
NRC	12650	0.02	0.02	-0.50	0.32	-0.03	53.30	1.35
SenticNet	12650	0.20	0.1	-0.88	0.97	-0.06	8.65	1.35
SVM-C	12650	0.66	0.20	0.00	1.00	-0.42	3.58	1.35
SVM-R	12650	0.44	0.06	0.06	0.82	-0.23	4.69	1.35
SWN	12650	0.01	0.03	-0.63	0.39	0.17	23.20	1.35
VADER	12650	0.35	0.32	-1.00	1.00	-0.43	3.78	1.35
Validation dataset								
AffectR	9016	0.16	0.02	-0.08	0.57	1.34	24.30	1.25
ANEW	9015	5.90	0.21	3.61	7.82	0.26	12.90	1.26
BingLiu	9018	0.02	0.03	-0.33	0.67	3.88	56.00	1.23
DistilBERT	9062	-0.18	0.40	-1.00	1.00	0.58	3.65	0.75
LIWC	9017	2.42	3.71	-32.60	100.00	5.33	88.60	1.24
Llama	9039	0.79	0.17	0.00	1.00	-1.16	5.44	1.00
NB	9018	0.40	0.22	0.00	1.00	0.46	3.33	1.23
NRC	9018	0.02	0.02	-0.21	0.33	0.98	21.20	1.23
SenticNet	9018	0.23	0.12	-0.86	0.90	0.38	7.18	1.23
SVM-C	9017	0.63	0.21	0.00	1.00	-0.35	3.47	1.24
SVM-R	9017	0.44	0.06	0.11	0.82	-0.15	4.48	1.24
SWN	9018	0.01	0.04	-0.40	0.63	1.38	24.70	1.23
VADER	9018	0.34	0.33	-0.99	1.00	-0.37	3.61	1.23

Table S5. Descriptive statistics of therapist unidimensional sentiment

Measure	N	M	SD	Min	Max	Skewness	Kurtosis	% missing
Exploratory dataset								
AffectR	12618	0.18	0.02	-0.03	0.39	-0.65	11.30	1.60
ANEW	12612	6.01	0.18	4.54	7.73	0.29	8.48	1.65
BingLiu	12619	0.03	0.03	-0.38	1.00	5.07	1.66	1.59

DistilBERT	12672	0.32	0.37	-1.00	1.00	-0.24	3.26	1.18
LIWC	12618	4.19	3.21	-8.33	100.00	4.06	75.40	1.60
Llama	12657	0.92	0.10	-1.00	1.00	-3.86	39.80	1.29
NB	12619	0.56	0.21	0.00	1.00	-0.05	3.13	1.59
NRC	12619	0.03	0.02	-0.13	0.40	2.18	29.5	1.59
SenticNet	12619	0.25	0.11	-0.86	0.90	-0.13	6.73	1.59
SVM-C	12619	0.75	0.17	0.00	1.00	-0.77	4.56	1.59
SVM-R	12619	0.51	0.05	0.14	0.84	0.33	5.41	1.59
SWN	12619	0.03	0.04	-0.29	0.56	1.22	17.60	1.59
VADER	12619	0.59	0.25	-0.99	1.00	-0.82	5.05	1.59
Validation dataset								
AffectR	9041	0.19	0.02	0.06	0.35	0.16	7.03	0.98
ANEW	9040	6.01	0.19	4.90	7.34	0.22	6.24	0.99
BingLiu	9046	0.03	0.02	-0.13	0.34	1.71	15.30	0.92
DistilBERT	9062	0.24	0.38	-0.99	1.00	-0.12	3.30	0.75
LIWC	9046	3.90	3.20	-10.90	100.00	5.02	104.00	0.92
Llama	9061	0.96	0.07	0.00	1.00	-4.08	36.00	0.76
NB	9046	0.55	0.22	0.00	1.00	-0.05	3.14	0.92
NRC	9046	0.03	0.02	-0.17	0.50	2.29	46.40	0.92
SenticNet	9046	0.28	0.12	-0.72	0.90	0.13	5.69	0.92
SVM-C	9046	0.73	0.19	0.00	1.00	-0.72	4.36	0.92
SVM-R	9046	0.51	0.05	0.23	0.99	0.43	5.72	0.92
SWN	9046	0.03	0.04	-0.36	0.49	0.83	17.50	0.92
VADER	9046	0.58	0.25	-0.99	1.00	-0.66	4.37	0.92

12. Please Clarify Effect Size Interpretation: Effect sizes are described as "small," but no criteria or benchmarks are provided. It would be helpful to reference standard conventions.

To strengthen the interpretation of our results, we have clarified that for linear mixed effects models we use Cohen's (1988) guidelines as benchmarks: "According to previous work (61), we used benchmark values of 0.1 representing a small effect, 0.3 representing a medium effect, and 0.5 representing a large effect" (p. 7). For RI-CLPMs, we reference the benchmark values recommended for cross-lagged effects: “According to previous work and guidelines on cross-lagged effects, benchmark values of 0.03 (small effect), 0.07 (medium effect), and 0.12 (large effect) were used.” (p. 8).

13. Please Report Sentiment Measure Correlations: The manuscript does not report correlations between the different sentiment measures. Including such information would allow readers to assess conceptual overlap or divergence among the tools.

This is another excellent suggestion. We have included a table reflecting both between- and within-subjects correlations in the Supplementary Materials (Tables S6-S9). All measures showed significant positive correlations, but the magnitude varied significantly between measures (ranging from .17 to .80 at the between-subject level and .12 to .70 at the within-subject level).

Table S6. Between-subjects correlations between measures of client unidimensional sentiment (all measures $p < .001$)

Measure	AffectR	ANEW	Bing Liu	Distil BERT	LIWC	Llama	NB	NRC	Sentic Net	SVM-C	SVM-R	SWN	VADER
Exploratory dataset													
AffectR	1.00	0.57	0.42	0.43	0.27	0.26	0.40	0.49	0.36	0.49	0.25	0.34	0.54
ANEW		1.00	0.71	0.51	0.67	0.36	0.25	0.63	0.50	0.29	0.52	0.52	0.55
BingLiu			1.00	0.50	0.72	0.40	0.25	0.59	0.47	0.23	0.58	0.58	0.43
DistilBERT				1.00	0.52	0.61	0.45	0.42	0.38	0.48	0.54	0.45	0.56
LIWC					1.00	0.42	0.26	0.45	0.39	0.20	0.56	0.45	0.36
Llama						1.00	0.38	0.34	0.36	0.37	0.52	0.37	0.49
NB							1.00	0.34	0.32	0.61	0.52	0.21	0.33
NRC								1.00	0.48	0.28	0.46	0.48	0.51
SenticNet									1.00	0.33	0.38	0.37	0.52
SVM-C										1.00	0.31	0.24	0.53
SVM-R											1.00	0.47	0.32
SWN												1.00	0.41
VADER													1.00
Validation dataset													
AffectR	1.00	0.64	0.51	0.47	0.54	0.29	0.42	0.55	0.44	0.41	0.39	0.39	0.52
ANEW		1.00	0.68	0.52	0.66	0.32	0.24	0.62	0.57	0.26	0.54	0.51	0.54
BingLiu			1.00	0.51	0.80	0.34	0.25	0.54	0.57	0.22	0.57	0.48	0.39
DistilBERT				1.00	0.55	0.52	0.45	0.41	0.47	0.43	0.60	0.46	0.48
LIWC					1.00	0.35	0.29	0.47	0.61	0.24	0.57	0.45	0.37

Llama						1.00	0.36	0.35	0.40	0.38	0.50	0.35	0.45
NB							1.00	0.35	0.37	0.62	0.52	0.17	0.34
NRC								1.00	0.48	0.29	0.48	0.47	0.52
SenticNet									1.00	0.37	0.50	0.36	0.50
SVM-C										1.00	0.35	0.23	0.51
SVM-R											1.00	0.48	0.34
SWN												1.00	0.40
VADER													1.00

Table S7. Within-subjects correlations between measures of client unidimensional sentiment (all measures $p < .001$)

Measure	AffectR	ANEW	Bing Liu	Distil BERT	LIWC	Llama	NB	NRC	Sentic Net	SVM-C	SVM-R	SWN	VADER
Exploratory dataset													
AffectR	1.00	0.57	0.53	0.28	0.45	0.10	0.24	0.42	0.29	0.28	0.32	0.351	0.36
ANEW		1.00	0.63	0.32	0.62	0.16	0.18	0.52	0.39	0.20	0.44	0.38	0.40
BingLiu			1.00	0.32	0.70	0.18	0.16	0.51	0.34	0.17	0.38	0.45	0.33
DistilBERT				1.00	0.34	0.32	0.20	0.22	0.19	0.19	0.42	0.26	0.34
LIWC					1.00	0.19	0.15	0.39	0.29	0.15	0.46	0.38	0.30
Llama						1.00	0.13	0.13	0.14	0.10	0.29	0.16	0.23
NB							1.00	0.15	0.19	0.43	0.38	0.12	0.18
NRC								1.00	0.33	0.14	0.36	0.41	0.34
SenticNet									1.00	0.20	0.32	0.28	0.34
SVM-C										1.00	0.32	0.13	0.24
SVM-R											1.00	0.36	0.39
SWN												1.00	0.31
VADER													1.00
Validation dataset													
AffectR	1.00	0.61	0.58	0.27	0.63	0.15	0.25	0.45	0.40	0.25	0.39	0.36	0.34
ANEW		1.00	0.62	0.30	0.60	0.19	0.18	0.56	0.48	0.19	0.47	0.40	0.42
BingLiu			1.00	0.30	0.70	0.16	0.17	0.49	0.45	0.20	0.46	0.39	0.34
DistilBERT				1.00	0.32	0.25	0.20	0.21	0.24	0.17	0.43	0.26	0.30
LIWC					1.00	0.18	0.16	0.39	0.45	0.17	0.45	0.34	0.29
Llama						1.00	0.11	0.15	0.18	0.12	0.29	0.18	0.23

NB							1.00	0.17	0.19	0.46	0.39	0.12	0.22
NRC								1.00	0.29	0.16	0.38	0.36	0.35
SenticNet									1.00	0.20	0.40	0.24	0.32
SVM-C										1.00	0.35	0.15	0.28
SVM-R											1.00	0.37	0.43
SWN												1.00	0.30
VADER													1.00

Table S8. Between-subjects correlations between measures of therapist unidimensional sentiment (all measures p < .001)

Measure	AffectR	ANEW	Bing Liu	Distil BERT	LIWC	Llama	NB	NRC	Sentic Net	SVM-C	SVM-R	SWN	VADER
Exploratory dataset													
AffectR	1.00	0.47	0.40	0.44	0.29	0.14	0.38	0.47	0.28	0.22	0.42	0.33	0.47
ANEW		1.00	0.71	0.47	0.67	0.24	0.14	0.62	0.49	0.17	0.55	0.51	0.43
BingLiu			1.00	0.42	0.76	0.23	0.15	0.62	0.50	0.15	0.62	0.58	0.41
DistilBERT				1.00	0.44	0.37	0.48	0.41	0.41	0.55	0.63	0.44	0.61
LIWC					1.00	0.27	0.15	0.57	0.48	0.14	0.58	0.50	0.38
Llama						1.00	0.15	0.20	0.24	0.26	0.27	0.22	0.35
NB							1.00	0.26	0.20	0.28	0.42	0.15	0.36
NRC								1.00	0.43	0.17	0.54	0.47	0.41
SenticNet									1.00	0.16	0.42	0.34	0.41
SVM-C										1.00	0.24	0.13	0.31
SVM-R											1.00	0.51	0.39
SWN												1.00	0.34
VADER													1.00
Validation dataset													
AffectR	1.00	0.55	0.48	0.44	0.42	0.25	0.33	0.52	0.31	0.14	0.47	0.36	0.40
ANEW		1.00	0.71	0.47	0.67	0.14	0.11	0.59	0.52	0.13	0.54	0.49	0.43
BingLiu			1.00	0.45	0.76	0.10	0.14	0.62	0.53	0.13	0.61	0.56	0.42
DistilBERT				1.00	0.48	0.31	0.42	0.42	0.42	0.46	0.65	0.44	0.50
LIWC					1.00	0.08	0.15	0.56	0.57	0.13	0.59	0.50	0.39
Llama						1.00	0.27	0.16	0.16	0.33	0.20	0.11	0.37

NB							1.00	0.25	0.18	0.25	0.40	0.12	0.32
NRC								1.00	0.39	0.14	0.52	0.47	0.43
SenticNet									1.00	0.15	0.44	0.31	0.39
SVM-C										1.00	0.22	0.12	0.28
SVM-R											1.00	0.48	0.38
SWN												1.00	0.32
VADER													1.00

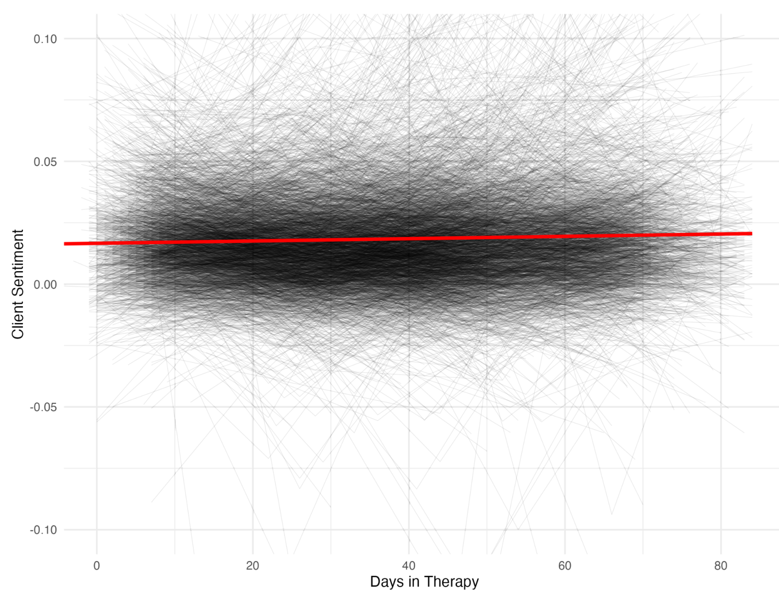
Table S9. Within-subjects correlations between measures of therapist unidimensional sentiment (all correlations p < .001 except one correlation marked with *, where p = 0.675)

Measure	AffectR	ANEW	Bing Liu	Distil BERT	LIWC	Llama	NB	NRC	Sentic Net	SVM-C	SVM-R	SWN	VADER
Exploratory dataset													
AffectR	1.000	0.532	0.438	0.285	0.364	0.080	0.263	0.452	0.291	0.234	0.409	0.393	0.412
ANEW		1.000	0.640	0.318	0.615	0.127	0.126	0.594	0.429	0.160	0.496	0.485	0.439
BingLiu			1.000	0.309	0.613	0.112	0.109	0.530	0.371	0.179	0.495	0.479	0.382
DistilBERT				1.000	0.293	0.160	0.221	0.246	0.225	0.251	0.423	0.285	0.339
LIWC					1.000	0.140	0.083	0.471	0.318	0.120	0.446	0.410	0.315
Llama						1.000	0.004*	0.010	0.124	0.042	0.108	0.091	0.154
NB							1.000	0.147	0.141	0.456	0.327	0.110	0.235
NRC								1.000	0.329	0.137	0.413	0.397	0.374
SenticNet									1.000	0.130	0.345	0.255	0.368
SVM-C										1.000	0.326	0.164	0.261
SVM-R											1.000	0.446	0.395
SWN												1.000	0.327
VADER													1.000
Validation dataset													
AffectR	1.00	0.59	0.57	0.31	0.48	0.16	0.22	0.50	0.32	0.24	0.47	0.40	0.41
ANEW		1.00	0.66	0.36	0.63	0.13	0.11	0.58	0.48	0.17	0.51	0.46	0.44
BingLiu			1.00	0.35	0.69	0.10	0.13	0.57	0.46	0.19	0.54	0.50	0.41
DistilBERT				1.00	0.32	0.14	0.20	0.25	0.28	0.27	0.46	0.29	0.32
LIWC					1.00	0.09	0.10	0.57	0.43	0.17	0.46	0.41	0.32

Llama	1.00	0.08	0.09	0.09	0.09	0.12	0.10	0.15
NB		1.00	0.14	0.13	0.49	0.33	0.09	0.18
NRC			1.00	0.30	0.14	0.42	0.41	0.37
SenticNet				1.00	0.21	0.39	0.27	0.34
SVM-C					1.00	0.36	0.15	0.24
SVM-R						1.00	0.43	0.40
SWN							1.00	0.29
VADER								1.00

14. Optional Visualization of Sentiment and Outcome Trajectories: It could be nice to include illustrative visualizations of how sentiment develops over time, ideally in relation to outcome measures.

We agree that visualizing the sentiment trajectories can be informative. However, with our dataset there's several features that limit the interpretability of such a figure. Our large number of participants makes these trajectories extremely cluttered, making it difficult to see any pattern. Additionally, because sentiment is bound within a narrow range (e.g. between -1 and 1) but time is on a much larger scale (0 to over 80 days), the change is hard to observe in a plot even though the pattern is statistically there. As such, we have opted to visualize the change in sentiment over time in therapy by showing the standardized effects for each measure (first panel in our main results). However, we have included an example lineplot below for your reference.



15. Caution on Claims About Psychopathology Tracking: The conclusion that sentiment can track internalizing psychopathology may be overstated. While promising, sentiment scores probably cannot substitute established psychometric questionnaires.

We agree with the reviewer's caution and have clarified in the discussion that although sentiment measures appear to be promising markers of psychopathology, they cannot replace established clinical evaluations or psychometric questionnaires. This is especially true given our small within-person effect sizes, and we further emphasize this by pointing out that sentiment measures lack specificity and likely reflect multiple dissociable constructs (e.g. emotional expression, social norms, emotional experiences, cognitive framing). Therefore, we see sentiment as a measure that compliments, rather than replaces, traditional psychopathology assessment methods.

We now write:

"Although language-based measures show promise for scalable assessments for mental health, it is important to note that effect sizes for within-person relations were small, and sentiment lacks the validity and reliability of established psychometric questionnaires. Recent work emphasizes that affective language reflects multiple dissociable social and psychological components, and that language measures are best viewed as complimentary tools rather than substitutes for validated clinical instruments (21; 76)."

16. Please Introduce NLP Terminology: Please consider introducing the term Natural Language Processing (NLP) in the introduction.

As noted in our response to point 2 above, we have revised our introduction to introduce the term Natural Language Processing (NLP) and briefly discuss its relevance to mental health research.

"Natural Language Processing (NLP), which is a branch of artificial intelligence focused on understanding and generating human language, has advanced significantly in the past years. These developments now allow us to examine the relationship between linguistic patterns and mental health on an unprecedented scale. For example, recent studies have analyzed millions of social media posts and messages to predict depression, anxiety, and even self-injurous thoughts (3, 4, 5, 6), demonstrating the power of analyzing language in predicting mental health related patterns."

17. Typo: On page 39, please correct "listed above below".
We thank the reviewer for catching this typo. It has been corrected in the revised manuscript.

Reviewer #2 (Remarks to the Author):

1. First, I want to sincerely commend the authors for their extraordinary effort in responding to the my and the other reviewers' comments. The response letter is thorough, thoughtful, and genuinely impressive, the additions of DistilBERT and Llama, the validation analyses, the RI-CLPMs, and the reconceptualization of sentiment as a marker rather than a mechanism all represent substantial work.

We appreciate the reviewer's positive response to our revisions!

2. I have a few more comment, though the revised manuscript is clearly improved. The manuscript itself still feels difficult to follow, likely because the revisions were integrated into the existing structure rather than used as an opportunity to reorganize. I'd strongly encourage the authors to consider restructuring the introduction around the three goals they articulate so well in their response letter (1) validating sentiment measures using prelabeled sentences, 2) comparing how different sentiment measures track symptom changes in therapy, and 3) testing whether changes in sentiment precede changes in symptoms (or vice versa)). This would give the reader a much clearer roadmap before entering the methods.

As one concrete example: the second paragraph currently moves from clinical settings to non-clinical examples and back to clinical settings, and the section on therapist language mixes several distinct rationales, therapist perception, empathic responding, and intervention techniques, without clearly distinguishing between them. These feel like artifacts of successive revisions rather than a coherent argument. Similarly, the paragraph discussing whether sentiment precedes or mirrors symptoms opens and closes with almost the same sentence, which makes it read as circular. A cleaner structure would help the reader understand what question is being asked and why before being walked through the evidence.

Thank you for these suggestions, which we have incorporated into the revised introduction. We have reorganized the introduction so it introduces our aims in the same order as the three aims of our overall paper, as suggested by the reviewer. We specifically attended to the paragraph the reviewer mentions to ensure it does not read as circular, and we have also added headers to help make our aims and organization crystal clear. We believe the revised introduction is much improved and are grateful for their suggestions. In the main text's revised introduction, sentences that have been moved, added, or reframed are shown in red.

3. On a methodological point: the paper collapses PHQ-8 and GAD-7 into a single internalizing composite throughout all analyses. While this is justified by their correlation in this dataset, depression and anxiety have meaningfully different affective profiles, particularly regarding positive affect, which is more specifically disrupted in depression than anxiety. Given that the paper examines both positive and negative sentiment separately, it seems worth at least a brief sensitivity analysis or discussion of whether

the sentiment-symptom relationships look different for depression versus anxiety specifically.

Thank you for this point. We have added separate analyses for anxiety and depression. These new analyses can be found in Supplemental Materials (Tables S26 and S27 for unidimensional sentiment; tables S28 and S29 for positive sentiment; and S30 and S31 for negative sentiment). We did not find clearly dissociable patterns for anxiety and depression: Higher depressive and anxious symptoms were associated with more negative and less positive client and therapist sentiment. We note these results in the methods section:

"Although we collapsed the PHQ-8 and GAD-7 scores to make a combined measure of internalizing symptoms in our main analyses, we also conducted non-preregistered sensitivity analyses examining depressive and anxious symptoms separately. We found a consistent pattern of results for both sets of symptoms (see Supplementary Materials for more information)."

4. Finally, a concern about how divergence is operationalized. The authors frame divergence as capturing therapist responsiveness (attunement, empathic responding) but because it's calculated by subtracting a client's 3-week average from a therapist's 3-week average, the metric completely loses the temporal dynamics that would make it meaningful. A therapist who is wildly out of sync, highly positive when the client is most distressed, more negative when the client is doing better, produces the exact same aggregate divergence score as a therapist who closely tracks their client throughout. The averaging erases the very information the construct is supposed to capture.

One approach worth considering would be to calculate divergence at the message or session level first and then aggregate to the 3-week bin, which would at least preserve the dynamic information. If re-analysis isn't feasible, the authors need to be much more explicit than they currently are, the current divergence metric reflects a macro-level difference in overall tone over three weeks, not anything resembling moment-to-moment attunement or responsiveness, and the theoretical framing should be adjusted accordingly.

Thank you for this comment. To clarify, we calculated the divergence for each 'turn' between client and therapist and then calculated the average of these differences within each 3-week bin. We did not average all client and all therapist sentiment in a bin and then take the difference between them (as the reviewer has understood). We apologize that our initial writing did not clarify this point, but we are pleased to hear that the results we have presented actually align with the reviewer's suggested modification. We now state our method more explicitly in the methods section:

"Text message data
Text messages were obtained with details including their delivery date and time, and whether they were sent by a client or therapist. If the client or therapist wrote multiple messages during their turn, these messages were combined to a single text block so subsequent analyses

focused on the sentiment of each individual's 'turn'. We then analyzed the sentiment of individual messages exchanged during each three-week period between symptom assessments and calculated average sentiment scores for these time periods. These averages were then used to predict subsequent symptom measures. The average length of text messages in our dataset was approximately 80 words (see Supplementary Materials, Table 1 for more information)."

"*Sentiment divergence*

We assessed client-therapist sentiment divergence for unidimensional, negative, and positive sentiment scores. We calculated divergence by subtracting the sentiment of a client's turn from the sentiment of the therapist's subsequent turn, and we averaged these scores for each of the 3-week bins. When the conversation was initiated by the therapist (58.19% in the exploratory dataset; 58.32% in the validation dataset), the first message sent by the therapist was excluded. As such, divergence scores represent how well therapists' language aligns with the sentiment of the client's preceding language. For the unidimensional score and positive sentiment, this yielded a one-dimensional score where positive values indicate the therapist's sentiment is more positive than the client's, and negative values indicate the opposite. For negative sentiment, we multiplied each score by -1, so that higher scores mean the therapist uses less negative language than the client, while negative values indicate the client is less negative than the therapist. This ensures all divergence scores can be interpreted such that more positive values indicate that the therapist's language is more positive / less negative than the client's."

In reviewing the details of our methods, we learned that our original submission did not exclude the first message if therapists spoke first, meaning that for some dyads, the divergence score assessed client-aligned-to-therapist divergence rather than therapist-aligned-to-client divergence. Fortunately, we analyzed our data and found that this distinction makes essentially no difference for our conclusions. However, we decided that it would be best to ensure divergence scores reflect the same construct for all participants (i.e., the gap between a client's sentiment and the therapist's response). We have thus recomputed scores to make this consistent across dyads in our final report.

Making this metric consistent across dyads has produced some very small changes to results in Figure 1, Table 4, and Table 5 (e.g., some associations slid from null to positive/negative or vice versa). To make this very clear, we provide a table below summarizing these minor shifts. Critically: (i) no relationships flip from significantly positive to significantly negative and (ii) the overall conclusions remain the same (i.e., divergence tracks increased symptoms at between-subjects level, and decreased symptoms at within-subjects level). If anything, our results are now more consistent across measures, with a stronger majority showing a positive relationship with time and a negative within-subject relationship with symptoms. Thus, we believe these adjustments do not alter primary takeaways and ensure our measures are more true to the construct of interest (i.e., how well therapists align in valence with what the client just said), so we report them in the final version of the paper.

	Change over time			Between-Ss symptom effect			Within-Ss symptom effect		
	+	null	-	+	null	-	+	null	-
Prior submission	**7/7**	4/**5**	**2**/1	12/13	1/0	0/0	**3**/2	4/5	**6**/6
Current submission	**8/8**	4/**4**	**1**/1	12/13	1/0	0/0	**2**/2	4/5	**7**/6

Table Note: Values indicate the number of sentiment measures (out of 13) that showed each direction of effect in the prior submission (which mixed client-aligned-to-therapist and therapist-aligned-to-client divergence) vs. the current submission (only therapist-aligned-to-client divergence). The value to the left of the slash in each cell represents the number of models in the exploratory dataset showing that effect, and the value to the right represents the validation dataset. Changes across submissions are marked in bold.

Reviewer #3 (Remarks to the Author):

1. Thank you for the opportunity to review the revised version of the manuscript. Overall, I appreciate the substantial revisions the authors have made in response to the previous reviews. In particular, I find it very positive that the authors addressed the earlier comment regarding the reliance on dictionary-based sentiment analysis methods. The inclusion of more recent approaches, including transformer-based models and a large language model (LLaMA), substantially strengthens the methodological scope of the study and provides an interesting comparison across different types of sentiment analysis techniques.

From my perspective, the manuscript has improved considerably and the authors have addressed the major concerns raised in the previous review. The analyses are clearly presented and the comparison of different sentiment analysis approaches is informative and relevant for the field.

We appreciate the reviewer's positive response to our revisions!

2. I only have one minor suggestion regarding the interpretation of the LLM-based sentiment analysis. While I appreciate the inclusion of a large language model in the analyses, it may be helpful to briefly acknowledge in the discussion that the performance of general-purpose LLMs in such tasks depends strongly on several factors, including the specific model used, the model size (in this case an 8B model), and the prompt design. As indicated in the supplement, the prompt used for classification is relatively simple (e.g., "Classify the following messages as Positive (1) or negative (0)"). It might therefore be useful to briefly note that alternative prompting strategies or larger models could potentially yield different results. A short sentence acknowledging this dependency would help contextualize the LLM results without requiring additional analyses.

Apart from this minor point, I believe the manuscript is suitable for publication.

Thank you for this excellent point. Our prompt was designed to be simple to allow out-of-box model comparison of across models. This framing represents typical zero-shot prompting for sentiment analysis, which has been shown to be highly reliable across different datasets (e.g. Amjadi et al, 2025; DiPalma et al, 2025). However, we agree that future studies could benefit from alternative strategies, and we have included this point in our discussion section.

“Future work should test whether using fully independent test sets, fine-tuning models or training them on clinical data yields substantially improved performance. In particular, different models and alternative prompting strategies that allow for continuous or bivariate measurement of sentiment may yield different results than our simple binary classification with LlaMA.”