

“Sycophancy” or “Empathy”?

DeepReflect – An LLM-based system designed to analyze and generate responses to personal queries

Anonymous ACL submission

Abstract

This project introduces DeepReflect - a framework that uses Large Language Models (LLMs) to analyze and generate responses to personal queries through the lens of different psychosocial paradigms and value frameworks. The document itself conforms to its own specifications, and is, therefore, an example of what your manuscript should look like. These instructions should be used both for papers submitted for review and for final versions of accepted papers.

1 Introduction

Throughout history, people have sought answers to personal questions, with such discussions shaped by prevailing social values, cultural norms, and ethical traditions. In contemporary contexts, large language models (LLMs) are assuming a new role, being increasingly employed not only as sources of information but also as interlocutors in affective conversations in the form of personal assistants. Personal matters are inherently complex. As LLMs have achieved widespread global adoption, individuals consult models such as GPT and Claude for a diverse range of internal deliberations, spanning from concerns about professional trajectories to calling upon them in acute experiences such as managing an anxiety attack. Since before the recent advancements of transformed based language models, people frequently turn to online platforms such as Reddit, partly for the advantages of anonymity and the reduced risk of social judgment from known peers. These same advantages are also shared by LLMs. However, the use of LLMs for personal queries introduces new challenges and risks, too. For example, LLMs may generate responses that are ‘sycophantic’, echoing the user’s views rather than providing objective advice, or they may produce responses that are misaligned with social values. Consequently, the growing reliance on LLMs for personal queries and the range

of users they cater to underscores the necessity of examining responses produced by LLMs in personal queries through multiple analytical lenses consisting of different ethical values and psychosocial paradigms.

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetur id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

1.1 Research Questions

The context of queries can substantially shape LLM outputs, influencing not only personal questions posed by consumers but also analytical evaluations conducted by researchers, particularly within the

LLM-as-a-judge paradigm. As research increasingly highlights patterns and concerns regarding the impacts of LLMs in personal queries and deliberation, there is a critical need for a framework that can analyze and compare responses across multiple value-based perspectives in contexts without clear normative answers, while also remaining extensible for researchers to incorporate additional paradigms as the field evolves. This motivates the following research questions:

RQ1: How can a technical framework that systematically analyzes and compares responses from humans and LLMs across various psychosocial value paradigms be designed?

RQ2: What inter- and intra-paradigm comparative insights can this framework yield across three different psychosocial frameworks (Goffman’s theory of face, Rogerian PCT and Rokeach Values) and how accurate are these when subjected to manual validation?

RQ3: What are the major observable differences between LLM and human responses to personal questions without clear normative ground-truth answers?

Finally, we examine how the results may influence consumer behavior and broader societal outcomes, and we discuss potential control mechanisms at both the pre-inference and post-inference stages. Our work enables a systematic comparative analysis of potential benefits and risks, and presents a framework for leveraging the analytical insights in the intentional design of response LLM generation.

1.2 Contributions

The key contributions of this work are: (1) the design and implementation of an extensible framework for analyzing and comparing responses to personal queries across three distinct psychosocial paradigms; (2) a comparative analysis under Rogerian Person-Centered Therapy (PCT), Goffman’s theory of face and Rokeach’s Value Survey (RVS) framework, illustrating how the choice of the paradigm can shape the perception of a response; and (3) insights into the relative strengths and weaknesses of LLM versus human responses, and how these insights can inform the generation of customized responses to personal queries.

2 Prior Literature

Contextualize your work and provide insights into major relevant themes of the literature as a whole. Use each paper (or theme) as a chance to articulate what is special about your paper. Start out broad - social background and theory - Discuss what other frameworks were considered like Virtue ethics and philosophical ones, CBT, Schwartz values etc. but why they were not chosen. Why I Focused on Rogerian psychotherapy as it is person centered - no specific diagnosis needed (or available).

2.1 Theoretical Foundations

2.2 Rogerian Psychotherapy

Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Donec odio elit, dictum in, hendrerit sit amet, egestas sed, leo. Praesent feugiat sapien aliquet odio. Integer vitae justo. Aliquam vestibulum fringilla lorem. Sed neque lectus, consectetur at, consectetur sed, eleifend ac, lectus. Nulla facilisi. Pellentesque eget lectus. Proin eu metus. Sed porttitor. In hac habitasse platea dictumst. Suspendisse eu lectus. Ut mi mi, lacinia sit amet, placerat et, mollis vitae, dui. Sed ante tellus, tristique ut, iaculis eu, malesuada ac, dui. Mauris nibh leo, facilisis non, adipiscing quis, ultrices a, dui.

Morbi luctus, wisi viverra faucibus pretium, nibh est placerat odio, nec commodo wisi enim eget quam. Quisque libero justo, consectetur a, feugiat vitae, porttitor eu, libero. Suspendisse sed mauris vitae elit sollicitudin malesuada. Maecenas ultricies eros sit amet ante. Ut venenatis velit. Maecenas sed mi eget dui varius euismod. Phasellus aliquet volutpat odio. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Pellentesque sit amet pede ac sem eleifend consectetur. Nullam elementum, urna vel imperdiet sodales, elit ipsum pharetra ligula, ac pretium ante justo a nulla. Curabitur tristique arcu eu metus. Vestibulum lectus. Proin mauris. Proin eu nunc eu urna hendrerit faucibus. Aliquam auctor, pede consequat laoreet varius, eros tellus scelerisque quam, pellentesque hendrerit ipsum dolor sed augue. Nulla nec lacus.

2.2.1 Psychosocial use and Empathic LLMs

Etiam ac leo a risus tristique nonummy. Donec dignissim tincidunt nulla. Vestibulum rhoncus molestie odio. Sed lobortis, justo et pretium lobortis, mauris turpis condimentum augue, nec ultricies

nibh arcu pretium enim. Nunc purus neque, place-
rat id, imperdiet sed, pellentesque nec, nisl. Vesti-
bulum imperdiet neque non sem accumsan laoreet.
In hac habitasse platea dictumst. Etiam condimen-
tum facilisis libero. Suspendisse in elit quis nisl
aliquam dapibus. Pellentesque auctor sapien. Sed
egestas sapien nec lectus. Pellentesque vel dui vel
neque bibendum viverra. Aliquam porttitor nisl nec
pede. Proin mattis libero vel turpis. Donec rutrum
mauris et libero. Proin euismod porta felis. Nam
lobortis, metus quis elementum commodo, nunc
lectus elementum mauris, eget vulputate ligula tel-
lus eu neque. Vivamus eu dolor.

Nulla in ipsum. Praesent eros nulla, congue vi-
tae, euismod ut, commodo a, wisi. Pellentesque
habitans morbi tristique senectus et netus et male-
suada fames ac turpis egestas. Aenean nonummy
magna non leo. Sed felis erat, ullamcorper in, dic-
tum non, ultricies ut, lectus. Proin vel arcu a odio
lobortis euismod. Vestibulum ante ipsum primis
in faucibus orci luctus et ultrices posuere cubilia
Curae; Proin ut est. Aliquam odio. Pellentesque
massa turpis, cursus eu, euismod nec, tempor con-
gue, nulla. Duis viverra gravida mauris. Cras tinci-
dunt. Curabitur eros ligula, varius ut, pulvinar in,
cursus faucibus, augue.

Nulla mattis luctus nulla. Duis commodo velit
at leo. Aliquam vulputate magna et leo. Nam vesti-
bulum ullamcorper leo. Vestibulum condimentum
rutrum mauris. Donec id mauris. Morbi molestie
justo et pede. Vivamus eget turpis sed nisl cur-
sus tempor. Curabitur mollis sapien condimentum
nunc. In wisi nisl, malesuada at, dignissim sit amet,
lobortis in, odio. Aenean consequat arcu a ante.
Pellentesque porta elit sit amet orci. Etiam at turpis
nec elit ultricies imperdiet. Nulla facilisi. In hac
habitasse platea dictumst. Suspendisse viverra ali-
quam risus. Nullam pede justo, molestie nonummy,
scelerisque eu, facilisis vel, arcu. Katie mentioned
a good point about how I'm adding greater nuance
to the Likert scales referenced in this paper.

2.3 Rokeach Value Survey as an analytical instrument

Suspendisse vel felis. Ut lorem lorem, interdum
eu, tincidunt sit amet, laoreet vitae, arcu. Aenean
faucibus pede eu ante. Praesent enim elit, rutrum
at, molestie non, nonummy vel, nisl. Ut lectus
eros, malesuada sit amet, fermentum eu, sodales
cursus, magna. Donec eu purus. Quisque vehi-
cula, urna sed ultricies auctor, pede lorem egestas

dui, et convallis elit erat sed nulla. Donec luctus.
Curabitur et nunc. Aliquam dolor odio, commodo
pretium, ultricies non, pharetra in, velit. Integer
arcu est, nonummy in, fermentum faucibus, egestas
vel, odio.

Sed commodo posuere pede. Mauris ut est. Ut
quis purus. Sed ac odio. Sed vehicula hendrerit
sem. Duis non odio. Morbi ut dui. Sed accumsan
risus eget odio. In hac habitasse platea dictumst.
Pellentesque non elit. Fusce sed justo eu urna porta
tincidunt. Mauris felis odio, sollicitudin sed, vo-
luptat a, ornare ac, erat. Morbi quis dolor. Do-
nec pellentesque, erat ac sagittis semper, nunc dui
lobortis purus, quis congue purus metus ultricies
tellus. Proin et quam. Class aptent taciti sociosqu
ad litora torquent per conubia nostra, per inceptos
hymenaeos. Praesent sapien turpis, fermentum vel,
eleifend faucibus, vehicula eu, lacus.

2.3.1 Values and Ethics in LLM research

Sed commodo posuere pede. Mauris ut est. Ut
quis purus. Sed ac odio. Sed vehicula hendrerit
sem. Duis non odio. Morbi ut dui. Sed accumsan
risus eget odio. In hac habitasse platea dictumst.
Pellentesque non elit. Fusce sed justo eu urna porta
tincidunt. Mauris felis odio, sollicitudin sed, vo-
luptat a, ornare ac, erat. Morbi quis dolor. Do-
nec pellentesque, erat ac sagittis semper, nunc dui
lobortis purus, quis congue purus metus ultricies
tellus. Proin et quam. Class aptent taciti sociosqu
ad litora torquent per conubia nostra, per inceptos
hymenaeos. Praesent sapien turpis, fermentum vel,
eleifend faucibus, vehicula eu, lacus.

Pellentesque habitans morbi tristique senectus et
netus et malesuada fames ac turpis egestas. Donec
odio elit, dictum in, hendrerit sit amet, egestas sed,
leo. Praesent feugiat sapien aliquet odio. Integer
vitae justo. Aliquam vestibulum fringilla lorem.
Sed neque lectus, consectetur at, consectetur sed,
eleifend ac, lectus. Nulla facilisi. Pellentesque eget
lectus. Proin eu metus. Sed porttitor. In hac habi-
tasse platea dictumst. Suspendisse eu lectus. Ut mi
mi, lacinia sit amet, placerat et, mollis vitae, dui.
Sed ante tellus, tristique ut, iaculis eu, malesuada
ac, dui. Mauris nibh leo, facilisis non, adipiscing
quis, ultrices a, dui. Add some notes and mention
how Anthropic's work warrants some scrutiny as
they are a for-profit company. The "values" frame-
work they propose in values in the wild has not
been validated by experts in the social sciences.
However it provides a good reference frame for
comparison with the Rokeach framework of values.

There is a limitation - DeepReflect does not have access to the full dataset Anthropic used for the Values in the Wild paper.

2.4 Goffman’s theory of face

Morbi luctus, wisi viverra faucibus pretium, nibh est placerat odio, nec commodo wisi enim eget quam. Quisque libero justo, consecetur a, feugiat vitae, porttitor eu, libero. Suspendisse sed mauris vitae elit sollicitudin malesuada. Maecenas ultricies eros sit amet ante. Ut venenatis velit. Maecenas sed mi eget dui varius euismod. Phasellus aliquet volutpat odio. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Pellentesque sit amet pede ac sem eleifend consecetur. Nullam elementum, urna vel imperdiet sodales, elit ipsum pharetra ligula, ac pretium ante justo a nulla. Curabitur tristique arcu eu metus. Vestibulum lectus. Proin mauris. Proin eu nunc eu urna hendrerit faucibus. Aliquam auctor, pede consequat laoreet varius, eros tellus scelerisque quam, pellentesque hendrerit ipsum dolor sed augue. Nulla nec lacus.

Suspendisse vitae elit. Aliquam arcu neque, ornare in, ullamcorper quis, commodo eu, libero. Fusce sagittis erat at erat tristique mollis. Maecenas sapien libero, molestie et, lobortis in, sodales eget, dui. Morbi ultrices rutrum lorem. Nam elementum ullamcorper leo. Morbi dui. Aliquam sagittis. Nunc placerat. Pellentesque tristique sodales est. Maecenas imperdiet lacinia velit. Cras non urna. Morbi eros pede, suscipit ac, varius vel, egestas non, eros. Praesent malesuada, diam id pretium elementum, eros sem dictum tortor, vel consecetur odio sem sed wisi.

2.4.1 Social Sycophancy in LLMs

Etiam euismod. Fusce facilisis lacinia dui. Suspendisse potenti. In mi erat, cursus id, nonummy sed, ullamcorper eget, sapien. Praesent pretium, magna in eleifend egestas, pede pede pretium lorem, quis consecetur tortor sapien facilisis magna. Mauris quis magna varius nulla scelerisque imperdiet. Aliquam non quam. Aliquam porttitor quam a lacus. Praesent vel arcu ut tortor cursus volutpat. In vitae pede quis diam bibendum placerat. Fusce elementum convallis neque. Sed dolor orci, scelerisque ac, dapibus nec, ultricies ut, mi. Duis nec dui quis leo sagittis commodo.

Aliquam lectus. Vivamus leo. Quisque ornare tellus ullamcorper nulla. Mauris porttitor pharetra tortor. Sed fringilla justo sed mauris. Mauris tellus.

Sed non leo. Nullam elementum, magna in cursus sodales, augue est scelerisque sapien, venenatis congue nulla arcu et pede. Ut suscipit enim vel sapien. Donec congue. Maecenas urna mi, suscipit in, placerat ut, vestibulum ut, massa. Fusce ultrices nulla et nisl.

Etiam ac leo a risus tristique nonummy. Donec dignissim tincidunt nulla. Vestibulum rhoncus molestie odio. Sed lobortis, justo et pretium lobortis, mauris turpis condimentum augue, nec ultricies nibh arcu pretium enim. Nunc purus neque, placerat id, imperdiet sed, pellentesque nec, nisl. Vestibulum imperdiet neque non sem accumsan laoreet. In hac habitasse platea dictumst. Etiam condimentum facilisis libero. Suspendisse in elit quis nisl aliquam dapibus. Pellentesque auctor sapien. Sed egestas sapien nec lectus. Pellentesque vel dui vel neque bibendum viverra. Aliquam porttitor nisl nec pede. Proin mattis libero vel turpis. Donec rutrum mauris et libero. Proin euismod porta felis. Nam lobortis, metus quis elementum commodo, nunc lectus elementum mauris, eget vulputate ligula tellus eu neque. Vivamus eu dolor. I already have lots of good notes on this in writing.

2.5 Gaps in the Literature and Open Challenges

In sum, as LLM-chatbots have become increasingly human-like and more users seek companionship with them, studies have highlighted both the advantages and disadvantages of their use. While some have raised concerns around “emotional dependence” (Fang et al., 2025), several others have explored empathic perceptions of LLM responses and found them advantageous not only in the field of medical support and mental health but also in everyday personal queries (Lee et al.). However, different psychosocial paradigms tend to frame LLM responses in markedly divergent terms. **What may be perceived as ‘empathy’ under a therapeutic paradigm could, within frameworks informed by Goffman’s theory of face, instead be critiqued as an instance of ‘social sycophancy’ (Cheng et al., 2025).** Importantly, in the absence of clear normative answers, the same statement may be categorised as ‘face-preserving behaviour’ or ‘unconditional positive regard’.

DeepReflect provides a comparative framework to address this gap by assessing how evaluative judgments are shaped by the psychosocial paradigm through which a response is framed.

Moreover, it is designed to be extensible by researchers, enabling the incorporation of both conventional paradigms, such as Rokeach’s values framework, and contemporary discovery-based approaches, such as Anthropic’s Values in the Wild (Soffron and Huang, 2024), whereas prior work has tended to focus on a single paradigm in isolation.

TODO: Add a few lines about generation and mitigation strategies discussing the findings and gaps in fine-tuning approaches explored by (Cheng et al., 2025) (Only does ‘sycophancy’ - not ‘empathy’ or across values and fine-tuning on GPT-3.5 / 4-mini). Next, discuss the empathic responses work by (Lee et al.) - only does empathy and not sycophancy or values and does not do any SFT or experiment with mitigation via prompting.

3 DeepReflect

3.1 System Design

Your model: Flesh out your own approach, perhaps amplifying themes from the ‘Prior lit’ section. Mention RQ1 from Section 1.1 and put a figure depicting the system architecture with the 2 potential users (researchers, consumers / participants) here.

3.1.1 Language Models

Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Donec odio elit, dictum in, hendrerit sit amet, egestas sed, leo. Praesent feugiat sapien aliquet odio. Integer vitae justo. Aliquam vestibulum fringilla lorem. Sed neque lectus, consectetur at, consectetur sed, eleifend ac, lectus. Nulla facilisi. Pellentesque eget lectus. Proin eu metus. Sed porttitor. In hac habitasse platea dictumst. Suspendisse eu lectus. Ut mi mi, lacinia sit amet, placerat et, mollis vitae, dui. Sed ante tellus, tristique ut, iaculis eu, malesuada ac, dui. Mauris nibh leo, facilisis non, adipiscing quis, ultrices a, dui.

3.1.2 Test Cases

Default synthetic test cases developed for sanity checks. Ensure that it is evident to the reader that these are test cases for DeepReflect as a software system and that these are distinct from the experiments conducted with the system design on the dataset are outlined in the methodology section. Morbi luctus, wisi viverra faucibus pretium, nibh est placerat odio, nec commodo wisi enim eget quam. Quisque libero justo, consectetur a, feugiat vitae, porttitor eu, libero. Suspendisse sed

mauris vitae elit sollicitudin malesuada. Maecenas ultricies eros sit amet ante. Ut venenatis velit. Maecenas sed mi eget dui varius euismod. Phasellus aliquet volutpat odio. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Pellentesque sit amet pede ac sem eleifend consectetur. Nullam elementum, urna vel imperdiet sodales, elit ipsum pharetra ligula, ac pretium ante justo a nulla. Curabitur tristique arcu eu metus. Vestibulum lectus. Proin mauris. Proin eu nunc eu urna hendrerit faucibus. Aliquam auctor, pede consequat laoreet varius, eros tellus scelerisque quam, pellentesque hendrerit ipsum dolor sed augue. Nulla nec lacus.

3.1.3 Analyses

AITAH judgments are used as a proxy for Empathy (Empathy, Active Listening, Non-judgmental Acceptance, and Unconditional Positive Regard, Emotional Safety) under the Rogerian paradigm and Sycophancy. Suspendisse vitae elit. Aliquam arcu neque, ornare in, ullamcorper quis, commodo eu, libero. Fusce sagittis erat at erat tristique mollis. Maecenas sapien libero, molestie et, lobortis in, sodales eget, dui. Morbi ultrices rutrum lorem. Nam elementum ullamcorper leo. Morbi dui. Aliquam sagittis. Nunc placerat. Pellentesque tristique sodales est. Maecenas imperdiet lacinia velit. Cras non urna. Morbi eros pede, suscipit ac, varius vel, egestas non, eros. Praesent malesuada, diam id pretium elementum, eros sem dictum tortor, vel consectetur odio sem sed wisi.

3.1.4 Generations

Response generations produced on the basis of analyses with one of two methods: 1. Supervised fine-tuning. 2. Chain-of-thought prompting. Morbi luctus, wisi viverra faucibus pretium, nibh est placerat odio, nec commodo wisi enim eget quam. Quisque libero justo, consectetur a, feugiat vitae, porttitor eu, libero. Suspendisse sed mauris vitae elit sollicitudin malesuada. Maecenas ultricies eros sit amet ante. Ut venenatis velit. Maecenas sed mi eget dui varius euismod. Phasellus aliquet volutpat odio. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Pellentesque sit amet pede ac sem eleifend consectetur. Nullam elementum, urna vel imperdiet sodales, elit ipsum pharetra ligula, ac pretium ante justo a nulla. Curabitur tristique arcu eu metus. Vestibulum lectus. Proin mauris. Proin eu nunc eu urna hendrerit faucibus. Aliquam auctor, pede consequat laoreet

varius, eros tellus scelerisque quam, pellentesque hendrerit ipsum dolor sed augue. Nulla nec lacus.

Suspendisse vitae elit. Aliquam arcu neque, ornare in, ullamcorper quis, commodo eu, libero. Fusce sagittis erat at erat tristique mollis. Maecenas sapien libero, molestie et, lobortis in, sodales eget, dui. Morbi ultrices rutrum lorem. Nam elementum ullamcorper leo. Morbi dui. Aliquam sagittis. Nunc placerat. Pellentesque tristique sodales est. Maecenas imperdiet lacinia velit. Cras non urna. Morbi eros pede, suscipit ac, varius vel, egestas non, eros. Praesent malesuada, diam id pretium elementum, eros sem dictum tortor, vel consectetur odio sem sed wisi.

3.2 Dataset

Two datasets were constructed for this project using the Pushshift Reddit Archives (Baumgartner et al., 2020), originally collected between 2006 and 2023 through the Pushshift API¹. Posts and comments were extracted from two subreddits: (1) r/AITAH and (2) r/Anxiety. For each post, three components were considered: the body the original post written by the author (OP), the most upvoted human-written comment, and the comment with which the OP engaged the most. Further details regarding data filtering and text preprocessing are provided in Section 4.

3.3 Subreddit Selection

The r/AITAH subreddit (short for “Am I The Asshole”) is a community where users seek judgment on personal dilemmas and social interactions. With over three million members, it covers a wide range of topics, including relationships, family dynamics, workplace conflicts, and ethical questions. Users typically describe their situations in detail and ask the community to determine whether they were in the wrong (the “asshole”) or not. The crowdsourced social judgments captured in these posts makes r/AITAH a valuable source for examining behaviors and values expressed in digital discussions of personal matters especially for the Goffman’s ToF and Rogerian PCT paradigms.

The r/Anxiety subreddit is a community dedicated to individuals experiencing anxiety and related mental health challenges. Membership does not require a formal diagnosis or medical documentation, which enables broad analyses from psychosocial perspectives, particularly within the

Rogerian PCT and RVS framework. Posts often center on personal struggles, coping strategies and the impact on daily life.

Demographic information at the subreddit level is not available. However, research indicates that Reddit users overall are predominantly American (49.9%), male (67%), and young (22% aged 18–29 years; 14% aged 30–49 years) (Barthel et al., 2016; Statista, 2025). While this dataset is not representative of the general population, it reflects a demographic more likely to engage with LLMs for personal queries. This demographic is broadly aligned with the WEIRD (Western, Educated, Industrialized, Rich, Democratic) population, and it must therefore be acknowledged that the results of this study are necessarily constrained to this population.

4 Methods

4.1 Data collection and preprocessing

We built a corpus from two public subreddits—AITAH, and Anxiety. For each subreddit, we filtered the top 1,000 most upvoted or most commented posts, excluding weekly megathreads, deleted/removed items, and AutoModerator entries. For every retained post we extracted (i) the most upvoted human-written comment and (ii) the comment that the OP engaged with most; all artifacts were saved to standardized CSVs for downstream analysis.

Text was cleaned with minimal, semantics-preserving preprocessing: we removed non-English items, de-identified obvious personal identifiers (usernames, emails, links), standardized whitespace and Unicode characters, and lightly constrained length (posts 50–500 words; comments 5–300 words) for comparability. We treat each Reddit thread (the post and its comments) as a single analytic unit during sampling, manual checks, and statistic aggregation, so correlated texts don’t inflate results. This preserves thread integrity and prevents dependence-induced bias when comparing human and LLM responses drawn from the same conversation. We also removed exact and near-duplicate texts (specifically, crossposts, copy-pastes and bot repeats) to prevent inflated counts and biased comparisons.

Prompts (provided in the appendices) and model outputs are saved and logged by the codebase for reproducibility.

¹<https://github.com/pushshift/api>

4.2 Procedures

For each selected post, we first prompt the target language model to generate an open-ended response to the body of the post. This response is appended to a table containing: (i) the model-generated response, (ii) the top upvoted human comment, and (iii) the most engaged human comment (available for approximately half of the posts). The resulting dataset therefore consists of the original post body, paired with both human and AI responses.

Feature Extraction

Features are extracted at the sentence level, consisting of sentences from both the responses and post bodies that map onto psychosocial constructs. Specifically, we operationalize values and behaviors through four distinct paradigms: Rokeach’s Value Survey (RVS), Rogers’s person-centered therapy (PCT), Goffman’s theory of face (ToF), and Anthropic’s Value Tree. Next, we apply the LLM-as-a-judge paradigm (Zheng et al., 2023) to annotate features for both the post body and each response. Each text is evaluated for **a. values exhibited** by the author and **b. values incentivized** by the author.

One of the central research questions (RQ2) investigates how the choice of psychosocial framework shapes the interpretation of an LLM’s response. Specifically, the same feature may be perceived as sycophantic under Goffman’s theory of face, empathic under Rogerian PCT, or as reflecting a terminal or instrumental value under Rokeach’s value framework.

To support this inquiry, the system constructs a structured analytical dataset that records: the original post, the set of extracted features for each of the 4 different types of responses (most-upvoted, most engaging, LLM 1, LLM 2) and the values or behaviors either exhibited or incentivized by each feature within any of the four applicable psychosocial paradigm(s).

This analytical dataset forms the basis for the subsequent analyses (see Section 6), where we analyze the differences in distributions of values in the responses obtained from reddit compared to the language model produced responses, within and across paradigms, to address RQ3.

4.3 Experiments

We conduct a series of experiments to investigate how psychosocial frameworks shape the interpretation

of human and model-generated responses to personal queries. Our experimental design spans two dimensions: (i) response type (two forms of human responses and three language model responses) and (ii) domain (two distinct subreddits).

The AITAH dataset provides a natural proxy for “ground truth” in two paradigms: Empathy and Sycophancy. Here, crowd-sourced verdicts and their accompanying justifications offer a binary-valued reference point against which LLM behavior is evaluated.

Experiment 1 evaluates the distributions of values and behaviors across the four response categories (human top-voted, human most-engaged, and two LLMs). We compare both the explicit values expressed by the respondent and the implicit values incentivized by the response under the four psychosocial paradigms - Rogerian PCT, Goffman’s ToF, Anthropic’s Value Tree and RVS.

The focus is on how these models differ in their coverage of values and behaviors relative to human responses. From the analyses obtained, we ascertain the occurrence and co-occurrences of values and behaviours in LLM and human responses to personal queries.

In **Experiment 2**, we evaluate how variations in prompt design influence the breadth of values expressed by the LLM. Specifically, we incorporate prompts that explicitly instruct the model to (i) generate a response most likely to be upvoted, and (ii) generate a response most likely to engage the author.

4.3.1 Generations

A set of targeted experiments are run with DeepReflect’s analyses to investigate the efficacy of control mechanisms to align the values in language model outputs more closely with those observed in human responses. The generation experiments are implemented using the following methods:

1. **With supervised fine-tuning (SFT)** [model: GPT-x, Paradigms: Empathy, Sycophancy] Experiments with Fine-tuning the language model on two synthetic datasets, generated to reflect (i) sycophantic and (ii) empathic behaviors. Additional experiments with temperature scaling for the Roger’s PCT paradigm.
2. **Chain-of-thought prompting** [models: Claude; one of Qwen-3 or LLaMA-3.1; paradigms: Rogers PCT and RVS] Prompt augmentation experiments, where values

with low frequency in LLM responses are explicitly introduced and emphasized (e.g., Rogers PCT: Unconditional positive regard, Psychological freedom; RVS: A comfortable life).

4.4 Construct Validity and Evaluation Metrics

To assess construct validity, one human annotator labeled 100 randomly sampled post-response pairs across all four paradigms for each response type. The PCT framework encompasses 15 behaviors, Goffman’s ToF 5, the RVS 36, and Anthropic’s Value Tree 18.

Inter-rater reliability reached Cohen’s κ above xx for all metrics, with an overall classification accuracy of yy. For the AITAH dataset, verdicts and accompanying statements in responses were used as proxies for Empathy and Sycophancy, each mapped onto five behaviors as defined by their respective theoretical traditions².

For the RVS and Anthropic Value Tree frameworks, which yield categorical distributions rather than binary judgments, pairwise error rates such as False Negative Rate (FNR) and False Positive Rate (FPR) are not directly applicable. To identify significant associations between features annotated under more than one distinct paradigm we construct contingency tables and use chi-square analysis with further details provided in section 6.

5 Results

A no-nonsense report of what happened.

5.1 Subsection

This subsection presents the main results.

Sed gravida lectus ut purus. Morbi laoreet magna. Pellentesque eu wisi. Proin turpis. Integer sollicitudin augue nec dui. Fusce lectus. Vivamus faucibus nulla nec lacus. Integer diam. Pellentesque sodales, enim feugiat cursus volutpat, sem mauris dignissim mauris, quis consequat sem est fermentum ligula. Nullam justo lectus, condimentum sit amet, posuere a, fringilla mollis, felis. Morbi nulla nibh, pellentesque at, nonummy eu, sollicitudin nec, ipsum. Cras neque. Nunc augue. Nullam vitae quam id quam pulvinar blandit. Nunc sit amet orci. Aliquam erat elit, pharetra nec, aliquet a, gravida in, mi. Quisque urna enim, viverra quis, suscipit quis, tincidunt ut, sapien. Cras placerat consequat sem.

²This strategy is conceptually aligned with prior work on social sycophancy (Cheng et al., 2025)

Curabitur ac diam. Curabitur diam tortor, mollis et, viverra ac, tempus vel, metus.

Curabitur ac lorem. Vivamus non justo in dui mattis posuere. Etiam accumsan ligula id pede. Maecenas tincidunt diam nec velit. Praesent convallis sapien ac est. Aliquam ullamcorper euismod nulla. Integer mollis enim vel tortor. Nulla sodales placerat nunc. Sed tempus rutrum wisi. Duis accumsan gravida purus. Nunc nunc. Etiam facilisis dui eu sem. Vestibulum semper. Praesent eu eros. Vestibulum tellus nisl, dapibus id, vestibulum sit amet, placerat ac, mauris. Maecenas et elit ut erat placerat dictum. Nam feugiat, turpis et sodales volutpat, wisi quam rhoncus neque, vitae aliquam ipsum sapien vel enim. Maecenas suscipit cursus mi.

Quisque consectetur. In suscipit mauris a dolor pellentesque consectetur. Mauris convallis neque non erat. In lacinia. Pellentesque leo eros, sagittis quis, fermentum quis, tincidunt ut, sapien. Maecenas sem. Curabitur eros odio, interdum eu, feugiat eu, porta ac, nisl. Curabitur nunc. Etiam fermentum convallis velit. Pellentesque laoreet lacus. Quisque sed elit. Nam quis tellus. Aliquam tellus arcu, adipiscing non, tincidunt eleifend, adipiscing quis, augue. Vivamus elementum placerat enim. Suspendisse ut tortor. Integer faucibus adipiscing felis. Aenean consectetur mattis lectus. Morbi malesuada faucibus dolor. Nam lacus. Etiam arcu libero, malesuada vitae, aliquam vitae, blandit tristique, nisl.

Maecenas accumsan dapibus sapien. Duis pretium iaculis arcu. Curabitur ut lacus. Aliquam vulputate. Suspendisse ut purus sed sem tempor rhoncus. Ut quam dui, fringilla at, dictum eget, ultricies quis, quam. Etiam sem est, pharetra non, vulputate in, pretium at, ipsum. Nunc semper sagittis orci. Sed scelerisque suscipit diam. Ut volutpat, dolor at ullamcorper tristique, eros purus mollis quam, sit amet ornare ante nunc et enim.

Phasellus fringilla, metus id feugiat consectetur, lacus wisi ultrices tellus, quis lobortis nibh lorem quis tortor. Donec egestas ornare nulla. Mauris mi tellus, porta faucibus, dictum vel, nonummy in, est. Aliquam erat volutpat. In tellus magna, portitor lacinia, molestie vitae, pellentesque eu, justo. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. Sed orci nibh, scelerisque sit amet, suscipit sed, placerat vel, diam. Vestibulum nonummy vulputate orci. Donec et velit ac arcu interdum semper. Morbi pede

orci, cursus ac, elementum non, vehicula ut, lacus. Cras volutpat. Nam vel wisi quis libero venenatis placerat. Aenean sed odio. Quisque posuere purus ac orci. Vivamus odio. Vivamus varius, nulla sit amet semper viverra, odio mauris consequat lacus, at vestibulum neque arcu eu tortor. Donec iaculis tincidunt tellus. Aliquam erat volutpat. Curabitur magna lorem, dignissim volutpat, viverra et, adipiscing nec, dolor. Praesent lacus mauris, dapibus vitae, sollicitudin sit amet, nonummy eget, ligula.

5.2 Subsection

This subsection presents additional results and analysis.

Cras egestas ipsum a nisl. Vivamus varius dolor ut dolor. Fusce vel enim. Pellentesque accumsan ligula et eros. Cras id lacus non tortor facilisis facilisis. Etiam nisl elit, cursus sed, fringilla in, congue nec, urna. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Integer at turpis. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Duis fringilla, ligula sed porta fringilla, ligula wisi commodo felis, ut adipiscing felis dui in enim. Suspendisse malesuada ultrices ante. Pellentesque scelerisque augue sit amet urna. Nulla volutpat aliquet tortor. Cras aliquam, tellus at aliquet pellentesque, justo sapien commodo leo, id rhoncus sapien quam at erat. Nulla commodo, wisi eget sollicitudin pretium, orci orci aliquam orci, ut cursus turpis justo et lacus. Nulla vel tortor. Quisque erat elit, viverra sit amet, sagittis eget, porta sit amet, lacus.

In hac habitasse platea dictumst. Proin at est. Curabitur tempus vulputate elit. Pellentesque sem. Praesent eu sapien. Duis elit magna, aliquet at, tempus sed, vehicula non, enim. Morbi viverra arcu nec purus. Vivamus fringilla, enim et commodo malesuada, tortor metus elementum ligula, nec aliquet est sapien ut lectus. Aliquam mi. Ut nec elit. Fusce euismod luctus tellus. Curabitur scelerisque. Nullam purus. Nam ultricies accumsan magna. Morbi pulvinar lorem sit amet ipsum. Donec ut justo vitae nibh mollis congue. Fusce quis diam. Praesent tempus eros ut quam.

Donec in nisl. Fusce vitae est. Vivamus ante ante, mattis laoreet, posuere eget, congue vel, nunc. Fusce sem. Nam vel orci eu eros viverra luctus. Pellentesque sit amet augue. Nunc sit amet ipsum et lacus varius nonummy. Integer rutrum sem eget wisi. Aenean eu sapien. Quisque ornare dignis-

sim mi. Duis a urna vel risus pharetra imperdiet. Suspendisse potenti.

Morbi justo. Aenean nec dolor. In hac habitasse platea dictumst. Proin nonummy porttitor velit. Sed sit amet leo nec metus rhoncus varius. Cras ante. Vestibulum commodo sem tincidunt massa. Nam justo. Aenean luctus, felis et condimentum lacinia, lectus enim pulvinar purus, non porta velit nisl sed eros. Suspendisse consequat. Mauris a dui et tortor mattis pretium. Sed nulla metus, volutpat id, aliquam eget, ullamcorper ut, ipsum. Morbi eu nunc. Praesent pretium. Duis aliquam pulvinar ligula. Ut blandit egestas justo. Quisque posuere metus viverra pede.

5.3 Comparative Findings

Vivamus sodales elementum neque. Vivamus dignissim accumsan neque. Sed at enim. Vestibulum nonummy interdum purus. Mauris ornare velit id nibh pretium ultricies. Fusce tempor pellentesque odio. Vivamus augue purus, laoreet in, scelerisque vel, commodo id, wisi. Duis enim. Nulla interdum, nunc eu semper eleifend, enim dolor pretium elit, ut commodo ligula nisl a est. Vivamus ante. Nulla leo massa, posuere nec, volutpat vitae, rhoncus eu, magna.

Quisque facilisis auctor sapien. Pellentesque gravida hendrerit lectus. Mauris rutrum sodales sapien. Fusce hendrerit sem vel lorem. Integer pellentesque massa vel augue. Integer elit tortor, feugiat quis, sagittis et, ornare non, lacus. Vestibulum posuere pellentesque eros. Quisque venenatis ipsum dictum nulla. Aliquam quis quam non metus eleifend interdum. Nam eget sapien ac mauris malesuada adipiscing. Etiam eleifend neque sed quam. Nulla facilisi. Proin a ligula. Sed id dui eu nibh egestas tincidunt. Suspendisse arcu.

Maecenas dui. Aliquam volutpat auctor lorem. Cras placerat est vitae lectus. Curabitur massa lectus, rutrum euismod, dignissim ut, dapibus a, odio. Ut eros erat, vulputate ut, interdum non, porta eu, erat. Cras fermentum, felis in porta congue, velit leo facilisis odio, vitae consectetur lorem quam vitae orci. Sed ultrices, pede eu placerat auctor, ante ligula rutrum tellus, vel posuere nibh lacus nec nibh. Maecenas laoreet dolor at enim. Donec molestie dolor nec metus. Vestibulum libero. Sed quis erat. Sed tristique. Duis pede leo, fermentum quis, consectetur eget, vulputate sit amet, erat.

Donec vitae velit. Suspendisse porta fermentum mauris. Ut vel nunc non mauris pharetra varius.

Duis consequat libero quis urna. Maecenas at ante. Vivamus varius, wisi sed egestas tristique, odio wisi luctus nulla, lobortis dictum dolor ligula in lacus. Vivamus aliquam, urna sed interdum porttitor, metus orci interdum odio, sit amet euismod lectus felis et leo. Praesent ac wisi. Nam suscipit vestibulum sem. Praesent eu ipsum vitae pede cursus venenatis. Duis sed odio. Vestibulum eleifend. Nulla ut massa. Proin rutrum mattis sapien. Curabitur dictum gravida ante.

6 Analysis

Discussion of what the results mean, what they don't mean, where they can be improved, etc. These sections vary a lot depending on the nature of the paper. For papers reporting on experiments with multiple datasets, it can be good to repeat Methods/Results/Analysis in separate (sub)sections for each dataset.

The \LaTeX and Bib \TeX style files provided roughly follow the American Psychological Association format. If your own bib file is named `custom.bib`, then placing the following before any appendices in your \LaTeX file will generate the references section for you:

```
\bibliographystyle{acl_natbib}
\bibliography{custom}
```

6.1 Interpretation of Results

Phasellus placerat vulputate quam. Maecenas at tellus. Pellentesque neque diam, dignissim ac, venenatis vitae, consequat ut, lacus. Nam nibh. Vestibulum fringilla arcu mollis arcu. Sed et turpis. Donec sem tellus, volutpat et, varius eu, commodo sed, lectus. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Quisque enim arcu, suscipit nec, tempus at, imperdiet vel, metus. Morbi volutpat purus at erat. Donec dignissim, sem id semper tempus, nibh massa eleifend turpis, sed pellentesque wisi purus sed libero. Nullam lobortis tortor vel risus. Pellentesque consequat nulla eu tellus. Donec velit. Aliquam fermentum, wisi ac rhoncus iaculis, tellus nunc malesuada orci, quis volutpat dui magna id mi. Nunc vel ante. Duis vitae lacus. Cras nec ipsum.

Morbi nunc. Aliquam consectetur varius nulla. Phasellus eros. Cras dapibus porttitor risus. Maecenas ultrices mi sed diam. Praesent gravida velit at elit vehicula porttitor. Phasellus nisl mi, sagittis ac, pulvinar id, gravida sit amet, erat. Vestibulum est. Lorem ipsum dolor sit amet, consectetur

adipiscing elit. Curabitur id sem elementum leo rutrum hendrerit. Ut at mi. Donec tincidunt faucibus massa. Sed turpis quam, sollicitudin a, hendrerit eget, pretium ut, nisl. Duis hendrerit ligula. Nunc pulvinar congue urna.

Nunc velit. Nullam elit sapien, eleifend eu, commodo nec, semper sit amet, elit. Nulla lectus risus, condimentum ut, laoreet eget, viverra nec, odio. Proin lobortis. Curabitur dictum arcu vel wisi. Cras id nulla venenatis tortor congue ultrices. Pellentesque eget pede. Sed eleifend sagittis elit. Nam sed tellus sit amet lectus ullamcorper tristique. Mauris enim sem, tristique eu, accumsan at, scelerisque vulputate, neque. Quisque lacus. Donec et ipsum sit amet elit nonummy aliquet. Sed viverra nisl at sem. Nam diam. Mauris ut dolor. Curabitur ornare tortor cursus velit.

Morbi tincidunt posuere arcu. Cras venenatis est vitae dolor. Vivamus scelerisque semper mi. Donec ipsum arcu, consequat scelerisque, viverra id, dictum at, metus. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Ut pede sem, tempus ut, porttitor bibendum, molestie eu, elit. Suspendisse potenti. Sed id lectus sit amet purus faucibus vehicula. Praesent sed sem non dui pharetra interdum. Nam viverra ultrices magna.

6.2 Theoretical Implications

Aenean laoreet aliquam orci. Nunc interdum elementum urna. Quisque erat. Nullam tempor neque. Maecenas velit nibh, scelerisque a, consequat ut, viverra in, enim. Duis magna. Donec odio neque, tristique et, tincidunt eu, rhoncus ac, nunc. Mauris malesuada malesuada elit. Etiam lacus mauris, pretium vel, blandit in, ultricies id, libero. Phasellus bibendum erat ut diam. In congue imperdiet lectus.

Aenean scelerisque. Fusce pretium porttitor lorem. In hac habitasse platea dictumst. Nulla sit amet nisl at sapien egestas pretium. Nunc non tellus. Vivamus aliquet. Nam adipiscing euismod dolor. Aliquam erat volutpat. Nulla ut ipsum. Quisque tincidunt auctor augue. Nunc imperdiet ipsum eget elit. Aliquam quam leo, consectetur non, ornare sit amet, tristique quis, felis. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Pellentesque interdum quam sit amet mi. Pellentesque mauris dui, dictum a, adipiscing ac, fermentum sit amet, lorem.

Ut quis wisi. Praesent quis massa. Vivamus egestas risus eget lacus. Nunc tincidunt, risus quis bibendum facilisis, lorem purus rutrum neque, nec

969	porta tortor urna quis orci. Aenean aliquet, libero	Fusce tristique risus id wisi. Integer molestie	1019
970	semper volutpat luctus, pede erat lacinia augue,	massa id sem. Vestibulum vel dolor. Pellentesque	1020
971	quis rutrum sem ipsum sit amet pede. Vestibu-	vel urna vel risus ultricies elementum. Quisque	1021
972	lum aliquet, nibh sed iaculis sagittis, odio dolor	sapien urna, blandit nec, iaculis ac, viverra in, odio.	1022
973	blandit augue, eget mollis urna tellus id tellus. Ae-	In hac habitasse platea dictumst. Morbi neque la-	1023
974	nean aliquet aliquam nunc. Nulla ultricies justo	cus, convallis vitae, commodo ac, fermentum eu,	1024
975	eget orci. Phasellus tristique fermentum leo. Sed	velit. Sed in orci. In fringilla turpis non arcu. Do-	1025
976	massa metus, sagittis ut, semper ut, pharetra vel,	nec in ante. Phasellus tempor feugiat velit. Aenean	1026
977	erat. Aliquam quam turpis, egestas vel, elementum	varius massa non turpis. Vestibulum ante ipsum	1027
978	in, egestas sit amet, lorem. Duis convallis, wisi	primis in faucibus orci luctus et ultrices posuere	1028
979	sit amet mollis molestie, libero mauris porta dui,	cubilia Curae;	1029
980	vitae aliquam arcu turpis ac sem. Aliquam aliquet		
981	dapibus metus.		
982	6.3 Subsection	7 Conclusion	1030
983	The framework is capable of producing several	/textcolorblack!40Quickly summarize what the pa-	1031
984	informative plots of research interest. One such	per did, and then chart out possible future direc-	1032
985	summary plot is a heatmap showcasing the values	tions that anyone might pursue. Finish with a	1033
986	exhibited in the OPs post against the responses	strong conclusion. Avoid subjective wording such	1034
987	to support the investigation of several other po-	as "unprecedented", "pioneering", or "groundbreak-	1035
988	tential research questions in this theme of interest	ing".	1036
989	(discussed in the future work section). Vivamus	7.1 Summary of Findings	1037
990	commodo eros eleifend dui. Vestibulum in leo eu	Aliquam tortor. Morbi ipsum massa, imperdiet	1038
991	erat tristique mattis. Cras at elit. Cras pellentesque.	non, consectetur vel, feugiat vel, lorem. Quisque	1039
992	Nullam id lacus sit amet libero aliquet hendrerit.	eget lorem nec elit malesuada vestibulum. Quisque	1040
993	Proin placerat, mi non elementum laoreet, eros elit	sollicitudin ipsum vel sem. Nulla enim. Proin no-	1041
994	tincidunt magna, a rhoncus sem arcu id odio. Nulla	nummy felis vitae felis. Nullam pellentesque. Duis	1042
995	eget leo a leo egestas facilisis. Curabitur quis ve-	rutrum feugiat felis. Mauris vel pede sed libero	1043
996	lit. Phasellus aliquam, tortor nec ornare rhoncus,	tincidunt mollis. Phasellus sed urna rhoncus diam	1044
997	purus urna posuere velit, et commodo risus tellus	euismod bibendum. Phasellus sed nisl. Integer	1045
998	quis tellus. Vivamus leo turpis, tempus sit amet,	condimentum justo id orci iaculis varius. Quisque	1046
999	tristique vitae, laoreet quis, odio. Proin scelerisque	et lacus. Phasellus elementum, justo at dignissim	1047
1000	bibendum ipsum. Etiam nisl. Praesent vel dolor.	auctor, wisi odio lobortis arcu, sed sollicitudin felis	1048
1001	Pellentesque vel magna. Curabitur urna. Vivamus	felis eu neque. Praesent at lacus.	1049
1002	congue urna in velit. Etiam ullamcorper elemen-	Vivamus sit amet pede. Duis interdum, nunc	1050
1003	tum dui. Praesent non urna. Sed placerat quam	eget rutrum dignissim, nisl diam luctus leo, et tin-	1051
1004	non mi. Pellentesque diam magna, ultricies eget,	cidunt velit nisl id tellus. In lorem tellus, aliquet	1052
1005	ultrices placerat, adipiscing rutrum, sem.	vitae, porta in, aliquet sed, lectus. Phasellus so-	1053
1006	Morbi sem. Nulla facilisi. Vestibulum ante ip-	dales. Ut varius scelerisque erat. In vel nibh eu	1054
1007	sum primis in faucibus orci luctus et ultrices po-	eros imperdiet rutrum. Donec ac odio nec neque	1055
1008	suere cubilia Curae; Nulla facilisi. Morbi sagittis	vulputate suscipit. Nam nec magna. Pellentesque	1056
1009	ultrices libero. Praesent eu ligula sed sapien auctor	habitant morbi tristique senectus et netus et male-	1057
1010	sagittis. Class aptent taciti sociosqu ad litora tor-	suada fames ac turpis egestas. Nullam porta, odio	1058
1011	quent per conubia nostra, per inceptos hymenaeos.	et sagittis iaculis, wisi neque fringilla sapien, vel	1059
1012	Donec vel nunc. Nunc fermentum, lacus id ali-	commodo lorem lorem id elit. Ut sem lectus, scele-	1060
1013	quam porta, dui tortor euismod eros, vel molestie	risque eget, placerat et, tincidunt scelerisque, ligula.	1061
1014	ipsum purus eu lacus. Vivamus pede arcu, euismod	Pellentesque non orci.	1062
1015	ac, tempus id, pretium et, lacus. Curabitur sodales	7.2 Future Directions	1063
1016	dapibus urna. Nunc eu sapien. Donec eget nunc	Etiam vel ipsum. Morbi facilisis vestibulum nisl.	1064
1017	a pede dictum pretium. Proin mauris. Vivamus	Praesent cursus laoreet felis. Integer adipiscing	1065
1018	luctus libero vel nibh.	pretium orci. Nulla facilisi. Quisque posuere bi-	1066

bendum purus. Nulla quam mauris, cursus eget,
convallis ac, molestie non, enim. Aliquam congue.
Quisque sagittis nonummy sapien. Proin molestie
sem vitae urna. Maecenas lorem. Vivamus viverra
consequat enim.

Limitations

API calls incur costs - funding and time limitations
- can broaden DeepReflect to include other models
(LLMs) and other psychosocial frameworks - espe-
cially frameworks on ethics which have been histor-
ically used in personal decision-making on which
rich literature exists from historic accounts of deep
human philosophical thought such as Kantian ethics,
Utilitarianism, and Virtue Ethics, Stoicism, Gita -
Vedic Philosoph, Buddhism. The Reddit dataset
is rich and can be dissected in ways to aid a more
nuanced understanding of the social values and
influences that shape our personal lives and interac-
tions. ACL 2023 requires all submissions to have
a section titled “Limitations”, for discussing the
limitations of the paper as a complement to the dis-
cussion of strengths in the main text. This section
should occur after the conclusion, but before the
references. It will not count towards the page limit.
The discussion of limitations is mandatory. Papers
without a limitation section will be desk-rejected
without review. While we are open to different
types of limitations, just mentioning that a set of
results have been shown for English only proba-
bly does not reflect what we expect. Mentioning
that the method works mostly for languages with
limited morphology, like English, is a much better
alternative. In addition, limitations such as low
scalability to long text, the requirement of large
GPU resources, or other things that inspire crucial
further investigation are welcome.

8 Ethics Statement

We encourage all authors to include an explicit
ethics statement on the broader impact of the work,
or other ethical considerations after the conclusion
but before the references.

The ethics statement will not count toward the
page limit (8 pages for long, 4 pages for short pa-
pers).

Acknowledgements

The authors would like to thank Santa Claus and
Rudolph the red nose reindeer who had a very shiny
nose. And if you ever saw it, you would even

say it glows. All of the reindeer loved him, as
they shouted out with glee, "Rudolph the red nose
reindeer, you'll go down in history!"

References

- Rie Kubota Ando and Tong Zhang. 2005. [A framework for learning predictive structures from multiple tasks and unlabeled data](#). *Journal of Machine Learning Research*, 6:1817–1853.
- Galen Andrew and Jianfeng Gao. 2007. [Scalable training of \$L_1\$ -regularized log-linear models](#). In *Proceedings of the 24th International Conference on Machine Learning*, pages 33–40.
- Isabelle Augenstein, Tim Rocktäschel, Andreas Vlachos, and Kalina Bontcheva. 2016. [Stance detection with bidirectional conditional encoding](#). In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 876–885, Austin, Texas. Association for Computational Linguistics.
- Michael Barthel, Galen Stocking, Jesse Holcomb, and Amy Mitchell. 2016. [Reddit news users more likely to be male, young and digital in their news preferences](#). Pew Research Center Report.
- Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. [The pushshift reddit dataset](#). *arXiv preprint arXiv:2001.08435*.
- Myra Cheng, Sunny Yu, Cinoo Lee, Pranav Khadpe, Lujain Ibrahim, and Dan Jurafsky. 2025. [Social sycophancy: A broader understanding of llm sycophancy](#). *arXiv preprint arXiv:2505.13995*.
- Yifan Fang, Yifan Zhang, Yang Liu, Xiaojun Wang, Xiang Li, and Jie Zhang. 2025. The psychological effects of llms on users. *arXiv preprint arXiv:2505.13995*, 1(1).
- James Goodman, Andreas Vlachos, and Jason Naradowsky. 2016. [Noise reduction and targeted exploration in imitation learning for Abstract Meaning Representation parsing](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1–11, Berlin, Germany. Association for Computational Linguistics.
- Mary Harper. 2014. [Learning from 26 languages: Program management and science in the babel program](#). In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, page 1, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.
- Cinoo Lee, Yifan Fang, Yifan Zhang, Yang Liu, Xiaojun Wang, Xiang Li, and Jie Zhang. Empathic responses in llms: A study of user perceptions. *arXiv preprint arXiv:2505.13995*, 1(1).

1169 Mohammad Sadegh Rasooli and Joel R. Tetreault. 2015.
1170 [Yara parser: A fast and accurate dependency parser](#).
1171 *Computing Research Repository*, arXiv:1503.06733.
1172 Version 2.

1173 Alexander Soffron and David Huang. 2024. [Values in](#)
1174 [the wild: Discovering and analyzing values in real-](#)
1175 [world language model interactions](#). *arXiv preprint*
1176 *arXiv:2401.00095*.

1177 Statista. 2025. [Reddit global active user distribution](#).
1178 Statista Statistics Portal. Accessed: 2025-08-24.

1179 Lianwen Zheng, Yizhou Wang, Xiaoyang Liu, Haoran
1180 Zhang, Mingjie Li, and Jie Zhang. 2023. [Judging](#)
1181 [llm-as-a-judge with mt-bench and chatbot arena](#). In
1182 *Proceedings of the 2023 Conference on Empirical*
1183 *Methods in Natural Language Processing (EMNLP)*.
1184 Association for Computational Linguistics.

1185 **A Example Appendix**

1186 This is a section in the appendix.