Three Years of r/ChatGPT: Societal Impact Evaluations from Social Media Data

Anonymous Authors[1]

Abstract

ChatGPT was launched on November 30, 2022; the r/ChatGPT subreddit was created just one day later. Since then, chatbot-based AI products have gone from niche proofs-of-concept to widely-used household names. However, the ways in which adoption has developed among the public remains poorly understood. In this paper, we propose a framework for using social media as a data source for understanding the societal impact of widely-adopted consumer AI products, as well as a general approach to monitoring for societally-impactful trends in real time. We apply our framework to conduct what is, to the best of our knowledge, the first longitudinal study of r/ChatGPT. We find that, overall, r/ChatGPT posts over time illustrate the normalization of ChatGPT as an everyday consumer product rather than an exceptional, novel technology. However, our retrospective analysis also finds that posts about using ChatGPT for mental health support, and posts about developing emotional attachments to ChatGPT, both rise steadily in frequency almost immediately after the launch of GPT-4o in May 2024. We show that our real-time method can detect the increase in emotional engagement as early as October 2024—months before OpenAI made any (public) acknowledgment of this impact.

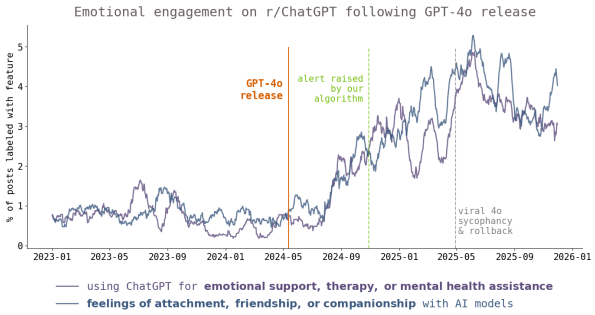

Figure 1. Posts about both AI therapy and AI companionship begin to rise in frequency almost immediately after the release of GPT-4o. We propose a real-time monitoring method (Section 4) that could have detected this as early as October 2024; in contrast, GPT-4o's behavior did not reach the level of public discourse until April 2025, when an extremely-sycophantic update triggered a rollback.[1]

1. Introduction

The launch of ChatGPT in late 2022 was a watershed moment for consumer AI products: ChatGPT reflected a step-change not only in the capabilities of AI products available to the general public, but in the degree to which any LLM-based product reached widespread consumer adoption. Now, a little more than three years after ChatGPT's launch, this recent history can be studied with the benefit of hindsight. To this end, recent works have sought to understand the realized impact of deploying LLM-based products on domains such as education, labor, and healthcare (e.g., Bastani et al. (2025); Brynjolfsson et al. (2025); Goh et al. (2024)).

However, a technology with a user base approaching a billion users will inevitably have unpredictable effects. How might we identify—and study—such effects? In this work, we turn to social media: beyond adoption, ChatGPT is also unique in the extent to which its rollout has been "online." Its users are highly active on social media—in fact, its early and explosive success can be attributed at least in part to virality on platforms like Twitter/X and Reddit. This makes social media a natural source of data for studying the societal impacts of ChatGPT in particular.[2]

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

[1]https://openai.com/index/sycophancy-in-gpt-4o/

[2]OpenAI employees often interact directly with users online; in fact, Sam Altman and other company leadership have conducted multiple Reddit AMAs ("ask me anything" sessions, where subreddit members can post questions for AMA subjects to reply to). The first appears to have been in 2024 (https://www.reddit.com/r/ChatGPT/comments/1ggixzy/ama_with_openais_sam_altman_kevin_weil_srinivas/).

Our approach relies on the core assumption that social media posts from everyday users of a technology reflect those users' perceptions and priorities about that technology—that is, that social media provides signal about "societal impact." However, what those perspectives actually entail is unknown *a priori*. Our framework thus begins with an unsupervised step to identify potentially-relevant ideas surfaced among all posts. Our key proposal to formalize *impact* is to explicitly track how these concepts develop over time; this can be quantified by placing temporal behavior in context with known external events, such as model and product releases.

To the best of our knowledge, ours is the first longitudinal analysis of r/ChatGPT over this time period. Our substantive findings in Section 3 tell two parallel stories of adoption. On the one hand, ChatGPT has become normalized as a tool that is a part of users' routine workflows for everyday tasks. On the other, emotional engagement with ChatGPT also emerges as an increasingly compelling use-case; this appears to be driven in large part by the GPT-4o model, which was released in May 2024.

One natural question is whether we might have known about these impacts sooner—and if so, how. Therefore, in Section 4 we also provide a *pro*spective approach to real-time monitoring. Our method discovers statistically meaningful growth in emotional engagement as early as October 2024—long before OpenAI took any public action regarding the emotional-health impacts of their product (see Appendix A.2 for discussion of what was "known," and by whom, at various points in time).

1.1. Related work

Our work makes use of the rich methodologies that have developed for clustering, topic modeling, and other unsupervised approaches (e.g., (Blei & Lafferty, 2006), or more recently, Pham et al. (2024); Reuter et al. (2024)). We use sparse autoencoders (SAEs), which have recently emerged as compelling methods for analyzing text-as-data (see, e.g., Jiang et al. (2025); Movva et al. (2025); Peng et al. (2025)); however, our methods are agnostic to what specific algorithm is used as long as the outputs of the method are consistent with what is outlined in Section 2.

Modeling the dynamics of online content over time is also a canonical problem (e.g., Leskovec et al. (2009); Danescu-Niculescu-Mizil et al. (2013)); more recently, Desiderio et al. (2025) study the dynamics of Reddit conversations in response to external events. Event and topic detection from social media is also well-studied (see, e.g., surveys in Karimiziarani (2022); Atefeh & Khreich (2015); Asgari-Chenaghlu et al. (2021)). Typical methods involve heuristic approaches to sequential clustering (e.g., Kolajo et al. (2022); McCreadie et al. (2013); Li et al. (2017); Aiello et al. (2013); Fedoryszak et al. (2019); Qiu et al. (2025)) and algorithms that identify "bursts" in specific topics or keywords (e.g., Mathioudakis & Koudas (2010); Xie et al. (2016); Shamma et al. (2011)). A subtle challenge that distinguishes our setting is that while prior work typically focuses on correctly identifying "trending" or "bursty" topics only in the moment, we are interested in tracking long-run changes over time, not just short-term effects.

The substantive findings we present in Section 3 build on prior works about social media posts about ChatGPT, including Twitter (Demirel et al., 2025) and Reddit, that have used both quantitative (e.g., Xu et al. (2024); Qutieshat (2024)) and qualitative (e.g., Choi et al. (2023)) approaches. Jung et al. (2025) explicitly studies posts about mental health on r/ChatGPT. Prior work has also analyzed subreddits more specific to emotional relationships, such as r/ReplikaOfficial and r/MyBoyfriendIsAI (e.g., Hanson & Bolthouse (2024); Depounti et al. (2023); Tunca (2025); Pataranutaporn et al. (2025)); our findings are complementary to (and consistent with) these. To the best of our knowledge, our work is the first to study r/ChatGPT with three years of data, and with the question of temporal variation explicitly in mind.

A primary goal for this work is to serve as an evaluation of the ChatGPT product from a longitudinal perspective. Prior works (e.g., Chen et al. (2024); Cen et al. (2025)) have studied LLMs longitudinally by prompting them repeatedly and analyzing how responses change over time; in contrast, the object of our analysis is user-reported impact, rather than immediate LLM output. In this way, our work can also be thought of as a crowdsourced evaluation (e.g., Deng et al. (2024); Dai et al. (2025b); Chiang et al. (2024)), though our data was not actively collected for the purpose of evaluation.

Finally, reliable usage data for LLM products is scarce. Thus, our work also complements industry whitepapers that report proprietary usage data (e.g., Tamkin et al. (2024)), and independent analysis of transcripts collected via data donations (e.g., Chowdhury & Garimella (2026), which similarly finds emotional engagement growing over time). Among industry reports, of particular note are Fang et al. (2025), a 2025 OpenAI study about the emotional impacts of chatbot design choices, and Chatterji et al. (2025), which reports that "the share of [ChatGPT] messages related to companionship or social-emotional issues is fairly small: only 1.9%," and instead emphasizes the extent of ChatGPT's practical uses. Our results suggest that usage frequency alone cannot paint a full picture of the magnitude of impact.

1.2. Data

Data was collected using a mixture of Pushshift (Baumgartner et al., 2020) and the Reddit API. Posts from r/ChatGPT are collected from December 1, 2022 to November 30, 2025, inclusive. Comment and upvote/downvote counts for all posts were updated in January 2026 using the API. We ex-

clude posts that are deleted, removed, posted by subreddit moderators, or are marked as "not robot indexable." As a lightweight spam filter, we also exclude posts with less than ten words (including title and post body) or two comments.

2. Method

(1) Initial featurization. The first step in our approach is to learn structured, human-interpretable features in an unsupervised fashion from unstructured text data. Formally, a *featurization* C is a mapping $[0,1]^d \to [0,1]^m$ that represents m features; for a d-dimensional representation of some text $X \in [0,1]^d$, the output $C(X) \in [0,1]^m$ quantifies the degree to which that text exhibits each of the m features. We will use $C^{(i)}$ for any $i \in [m]$ to describe how C represents the single feature i, so that $C^{(i)}(X) \in [0,1]$ quantifies the degree to which X exhibits feature i; we will sometimes refer to $C^{(i)}(X)$ as the "activation" of i on X.

In some abuse of notation, we will use X_s to denote all data from timestep s, and $X_{s:t}$ to denote the data from timesteps s to t. Throughout this work, we use days as our unit of time, so that each sample X_s is a "minibatch" of data from day s, and $C^{(i)}(X_s) := \frac{1}{|X_s|}\sum_{X \in X_s} C^{(i)}(X)$ is the average activation for feature i for all texts from day s.

To compute our featurization, we use sparse autoencoders (SAEs) with the standard reconstruction loss (following recent work, e.g, Movva et al. (2025)). Specifically, we concatenate post titles and texts, and embed them with OpenAI's `text-embedding-3` model. We use top-K SAEs with $K = 4$ and $M = 128$ (allowing 128 features total, and each sample to associate with 4 features), with samples weighted by $\log(n_{\text{upvotes}} - n_{\text{downvotes}} + n_{\text{comments}})$; see Appendix B for discussion of these design decisions, including consideration of PCA and k-means clustering as alternative featurization methods.

We interpret these 128 features with an LLM, and annotate all samples with binary labels for these features. We use the prompts from the implementation in Movva et al. (2025) for both feature interpretation and sample annotation. For each of these annotation tasks, we use `gpt-4.1-mini`. For feature interpretation, we use the best of three candidate interpretations (as measured by F1 score, following Movva et al. (2025)), and for sample annotation, we use the majority vote from three candidate labels. See Appendix F for details.

After initially computing $M = 128$ features, we remove some for focus. First, we remove extremely generic features, such as "uses ChatGPT at the start of text" (9 features) and features that had very few positively-labeled samples (5). We also exclude features related to image and video generation (14) and those corresponding to product releases (14). In Steps 2-4, we work with the 86 features that remain.

(2) Characterizing temporal trajectories. Given a featurization C, we compute the historical frequency of any feature i as a transcript $\{C^{(i)}(X_t)\}_{t \in [T]}$ for feature i at each day t. We use the binary *labels* from LLM annotation in the previous step, so that $C^{(i)}(X_t) := \frac{1}{|X_t|}\sum_{X \in X_t} \mathbf{1}_{[X \text{ labeled with } i]}$. We also treat the first month (December 2022) as a "burn-in" period and remove posts from those days, so that $T = 1034$, and apply a 30-day centered rolling mean.

To place all features in context with real-world events, we compile a timeline $\mathcal{T} = \{\tau_1, \tau_2, \dots\}$ of events that we may expect to affect the composition of posts online. Using OpenAI's official release notes, we choose twelve major model releases, listed in Table 10. With transcripts and the timeline in hand, we can quantify the degree to which particular features evolve over time, and/or are *reactive* to events in $\mathcal{T}$. Specifically, we assume that, absent any "impact", a feature's frequency should be roughly constant. However, transcripts may suggest evidence of impact in two ways. A change in slope that begins near or shortly after τ_j may reflect an effect of event j. On the other hand, long-run changes in a feature's frequency over the entire period of analysis—i.e., non-zero slope—suggest evidence of changing priorities that are not tied to specific external events, but reflect the progression of adoption more generally.

To capture the former (reactivity to specific events in $\mathcal{T}$), we model each transcript as piecewise-linear, with candidate changepoints only from $\mathcal{T}$; for each feature i, we approximate its transcript at t as

$$\lambda^i(t) = \beta_0 + \sum_{j \in |\mathcal{T}|} \gamma_j \max(0, x - \tau_j), \tag{1}$$

with each γ_j being the change in slope at τ_j.We fit Equation (1) for each feature over 100 bootstrap samples, sampling posts with replacement, and report changepoints that are selected in at least half of the bootstrap samples. To capture the latter (slope change over the full horizon), we use an OLS slope test for whether each feature's slope corresponds to at least a 10% change; we Bonferroni-correct over the total number of features, and use Newey-West HAC errors to handle autocorrelation in the time-series (Newey & West, 1987). For details, see Appendix B.2.

(3) Finding "families" of related features. While our final results inevitably require manual interpretation, we support our analysis by grouping features into "families" using quantitative methods.

The first approach is grouping by (stable) *changepoint*: that is, for each event $\tau_j \in \mathcal{T}$, we collect the features that had changepoint τ_j selected in a majority of bootstrap samples, and group them by type (slope increase or decrease); this produces two collections of features per $\tau_j \in \mathcal{T}$, plus two for features with no changepoints (overall increases and

overall decreases). The second approach is grouping by *similarity*: for all features i, we also compute *co-occurrences* with other features (i.e., which other features appear among posts that are labeled with i), and *trajectory similarity* with other features (i.e., which other features exhibit similar temporal behavior, regardless of whether individual posts contain them). We then use these similarities to compute a (hierarchical) clustering of the features.

We use these quantitative groupings to inform our manual interpretation of feature families. In Appendix C, we show our final categorizations, and highlight where groupings across methods converge. In our data, the vast majority of features characterize either *(mundane) adoption* (Section 3.1) or *emotional engagement* (Section 3.2); only six features (of 86) do not fit cleanly into any part of our interpretation.

3. Retrospective analysis

Our retrospective study finds two major stories of adoption.

Our first finding is that ChatGPT has become normalized as a regular consumer technology (3.1). While this finding is likely broadly consistent with many readers' personal experiences, the degree to which it is visible in across features about usage, user perspectives, and linguistic cues, is striking.

Our second main finding, previewed in Figure 1 and described further in Section 3.2, is more striking: the frequency of posts broadly related to emotional engagement—using ChatGPT for mental health support, or developing emotional attachments to models, for instance—began to rise in May 2024, shortly after the release of GPT-4o. This effect is visible long before the emotional and mental health aspects of LLM product usage had entered the public consciousness, and long before OpenAI publicly committed to any action regarding mental health implications of its product.

3.1. The "domestication" of ChatGPT

Our first high-level finding is that, broadly speaking, r/ChatGPT dynamics illustrate the ways in which the ChatGPT product has become normalized as a consumer technology. We borrow the term "domestication" from science and technology studies (STS), where it is a well-studied theory that describes the processes by which novel technologies are absorbed into everyday use (see, e.g., Haddon (2007)).[3] It is useful to keep in mind a key conceptual framing from this theory: posts on r/ChatGPT at any given time reflect what users feel is "worth posting about" at that point in time, and changes in the frequency of posts about different topics reflect changes in users' beliefs about postworthiness.

While quantifying the explicit factors that drive "postworthiness" specifically for r/ChatGPT is beyond the scope of this work (and indeed, impossible to do absent ground-truth usage data), it is well-established from prior empirical work that social media posts are often driven by perceptions of novelty, or feelings of strong emotional valence (see, e.g., Vosoughi et al. (2018); Wu & Huberman (2007); Yu et al. (2025)). Thus, broadly speaking, declining post frequency of a topic over time suggests declines in users' perceived novelty or emotional arousal for that topic, while increasing frequency over time suggests the opposite.

Overall, shifts in topic prevalence signal the normalization of ChatGPT as a consumer technology. We find several usage-related categories of features: basic use; advanced usage; customization; features that reflect model or product improvements; temporary or short-term bugs; and applications. There are also several categories broadly related to adoption, including: language and terminology; references to the subreddit community; perspectives on the broader ecosystem of LLMs; judgments about product updates; and discussions of jailbreaking and content policy.

Here, we briefly highlight some examples to illustrate the "domestication" story; see Tables 2 and 3 in Appendix C for all "domestication"-related features and results.

Increasing expert (and declining basic) product usage. The frequency of posts related to questions about basic product use (e.g., *login problems*) decrease over the three-year window of time, while features that suggest advanced and frequent usage (e.g., *organizing or searching chat histories*) increase. Furthermore, while *requests for help* is a somewhat-generic feature, examining trends within the 5568 posts that were labeled with this feature reveals a shift in user expectations around product usage. Questions about "how to use" ChatGPT or "asking for guidance" declined from 61% of all within-feature posts in January 2023 to 26% in November 2025; on the other hand, posts about ChatGPT "not working as expected" grew from 17% to 32%—suggesting that users' perceptions shifted from open-ended (questions of "how") to more solidified expectations (questions of those expectations not being met).[4] These changes are not just about whether new users are still com-

[3] In STS, the word choice of "domestication" is meant to evoke the sense of something "wild" and strange having been "tamed"; see discussion in Haddon (2007).

[4] To arrive at these sub-features, we train a SAE with $M = 4$ and $K = 1$ (in other words, to find four features with each post corresponding only to one feature) for the 5568 posts labeled as *requests for help*, and label each post with the corresponding sub-features. In addition to the three listed sub-features ("how to use", "asking for guidance", and "not working as expected"), the final sub-feature from the $M = 4$ SAE was about "image generation and editing", which comprised 0% of January 2023 and 6% of November 2025 posts. 26% of all posts labeled as *request for help* were not well-described by any of the four sub-features (22% in January 2023 and 37% in November 2022).

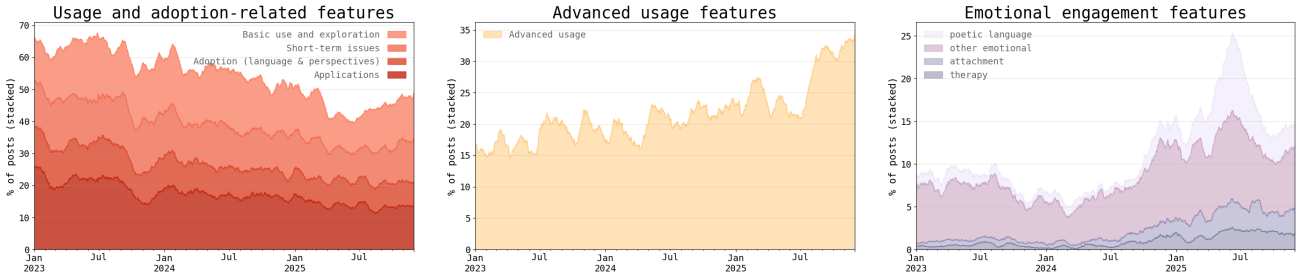

Figure 2. Composition of r/ChatGPT by category, over time (see categories in Tables 2 and 3); y-axis can be interpreted as "percentage of posts that fall into this category." Left: (non-advanced) usage and adoption posts decline. Middle, right: Advanced usage and emotion posts, respectively, increase.

ing to the product or subreddit—in fact, we know from usage data that growth has yet to slow—but about the expectations that change as more users develop expertise.

Declines in application-specific posts. Posts about applications (e.g., *programming* or *D&D and role-playing games*) also decline. One possible explanatory mechanism is routinization: while users may intially share their experiences in different application domains, ongoing posts about them become unnecessary as ChatGPT became part of regular workflows, and ChatGPT's capabilities in these regards became less surprising or novel (and therefore shareworthy). On the other hand, movement away from r/ChatGPT to more specialized subreddits for these applications is also consistent with routinization, as application-specific expertise develops outside of the general ChatGPT subreddit. A notable exception to the overall trend of declines in applications is a substantial increase in discussion of *medical conditions or diagnoses*; as we will discuss in Section 3.2, this is driven by its close relationship with emotional engagement features.

Language usage suggests familiarization. Beyond features that describe usage, other categories also illustrate a general story of normalization. For instance, early users often compared ChatGPT to *google search*, while later users no longer found that reference point important. Usage of *"bot" or "chatbot"* in reference to ChatGPT declines substantially, suggesting a overall familiarization with ChatGPT specifically, as opposed to a generic chatbot product. Interestingly, posts that use "chatbot" in the context of "building or improving AI chatbots" comprise 17% of within-feature posts January 2023 and 9% in November 2025; on the other hand, posts that "discuss psychological impacts of chatbots on humans" comprise 1% of within-feature posts in January 2023 and 24% in November 2025.[5] While the overall decline in "chatbot" usage suggests familiarization, the compositional shift *within* this feature suggests that usage of this defamiliarized framing is increasingly done in the context of raising concerns; this is especially notable as meta-discussion of emotional impact does *not* appear to be a substantial topic of conversation on the subreddit overall.

Evolving user perspectives: declines in speculation, increases in privacy concerns. At the same time, *predictions about future development and capabilities* and *discussions about how LLMs represent knowledge* fall substantially. Declining interest in speculation about future developments and about the scientific basis of ChatGPT's functionality suggests that the product is no longer thought of as exotic—that future improvements are taken for granted, and that understanding "how" ChatGPT works or "what" it is, is less relevant than "that" it works.[6] On the other hand, *privacy concerns* grow, as users share more personal information and use the product for increasingly intimate applications; as we will discuss in Section 3.2, this often takes the form of emotional engagement.

3.2. The emergence of emotional engagement

Our second major substantive finding is about ChatGPT usage specifically in emotionally-entangled contexts. While these features had been present prior to the GPT-4o release—previewing our results from Section 4, features related to therapy and emotional attachment appear as early March 2023—their prevalences begin to grow dramatically after the release of GPT-4o in May 2024.

A clear family of "emotional engagement" features emerges across changepoints, trajectories, and co-occurrences. We first highlight that the "emotional engagement" family of features is remarkably stable across

[5] As above, we train a SAE with $M = 4$ and $K = 1$ for each of the 2446 posts labeled as *mentions "bot" or "chatbot"*; in addition to the two identified sub-features above, the remaining sub-features are "user complaints or frustrations" (21% in both January 2023 and November 2025, though there is some variation in the months between), and "expressions of anger" (0% in January 2023 and 19% in November 2025). 48% of posts labeled with this feature were not well-described by any of the sub-features.

[6] In fact, domestication theory claims that when a technology is novel, users are interested in understanding, defining and contextualizing what it is; these questions become less important as adoption continues (Haddon, 2007).

different ways to analyze feature similarities: whether clustering by feature co-occurrence, by trajectory, or by both. The two core features that anchor this family are *personal attachments* and *therapy*, as shown in Figure 1. The full family of features also includes *personal stories about positive impact*, *romantic partners*, *poetic language*, *AI sentience*, and *naming ChatGPT*. In Figure 6, we show features that we categorize as related to "emotional engagement," along with some representative example posts for each feature; in Appendix **??**, we provide more quantitative details and list additional features that at least one of our quantitative methods groups with this category.

***Therapy* and *companion* capture distinct concepts.** The degree to which these features appear together across multiple measures of similarities may seem to suggest that they ought to be thought of as representing the same concept. To the contrary, however, they are quite distinct. While *therapy* has 2253 unique posts from 2052 unique users, and *companion* has 2926 posts from 2665 users, the number of posts labeled as both is only 364, and the number of users who have ever posted about both is 446—thus, while these features have more overlap than most other pairs of features, the absolute degree of overlap is small.

On a content level, basic vocabulary analysis (log-odds ratio; Monroe et al. (2008)) also confirms semantic differences: *therapy* posts are more likely to include words like *mental/health* (z-score 18.8 and 18.1, respectively), *help* (16.2), *support* (14.3), *trauma* (11.1), *anxiety* (10.8), *issues* (10.5), and *advice* (10.5). Meanwhile, *companion* posts contain words like *personality* (z-score 17.4), *feels* (14.6), *human(s)* (13.8), *conversation* (11.9), and *friend* (9.7); see Table 14 for full lists.

Posts about either *therapy* or *companionship* also exhibit distinct "profiles" in terms of what other features they tend to exhibit. In Table 7, we examine what other features are likely to co-occur with posts about therapy or companionship (excluding posts that are tagged as both). For instance, 20% and 4.9% of posts about *therapy* and *companionship*, respectively, are also tagged as *personal stories about positive impact*, which comprise 1.8% of all posts. While both exhibit a substantial "lift" for this feature, the lift for *therapy* features is over 4 times greater than for companion features. Interestingly, while *therapy* posts are over twice as likely to also mention *privacy concerns* compared to the baseline rate, *companion* posts are less than one third as likely. On the other hand, *therapy* posts are less than half as likely as baseline to either *name ChatGPT* or discuss *AI sentience*, while *companion* posts are 4.5 and 3.5 times more likely, respectively. Interestingly, *companion* posts are more than twice as likely as *therapy* posts to mention *recent quality declines*, suggesting that the former use case is more sensitive to model updates than the latter.

Emotional engagement shapes the trajectories of other features after GPT-4o release. Finally, we show that emotional engagement shapes the evolution of many other features, even when they do not appear to be overtly related to emotional engagement. For example, among posts that are *asking about daily or repeated usage of ChatGPT*, we find sub-features related to *managing prompts*, *paid tiers*, *productivity*, and *personal and emotional disclosures*. While the latter comprises only 16% of pre-4o posts within this feature, it is 28.8% of post-4o posts. Similarly, posts about the *positive impact of ChatGPT* are mainly about *productivity* and *mental health*; however, while the former exhibits no significant change before and after the launch of 4o (23%), the latter comprises 14% of all pre-4o posts and 41% of post-4o posts.

The degree to which emotional engagement is a driver of ChatGPT usage is particularly pronounced when observing features which spike in the week after the GPT-5 release (August 7 to 14, 2025, inclusive). Within this period, three of the top four features are complaints about GPT-5: *frustration or hatred about a product version* (598, or 12.2% of all posts), *dissatisfaction with 4o removal and loss of control* (552, 11.3%), and *lost, deleted, or missing conversations* (370, 7.6%); in total, 27.2% (1332) of all posts are labeled with at least one of these three features.[7] Among these posts, 164 are also labeled with either *therapy* or *companionship*; analyzing the sub-features of *dissatisfaction* and *lost conversations* features yields an additional 242 posts that also involve emotional engagement but were not already counted in the previous 164.[8] Thus, in total, emotional engagement is involved in at least 30.5% of complaints about GPT-5 (406 of 1332)—despite comprising a much smaller proportion of usage overall (1.8%, according to Chatterji et al. (2025)). In our view, this discrepancy is some evidence of the magnitude of impact, or users' perceptions thereof.

4. Real-time monitoring

Given that societally-impactful patterns clearly emerge in hindsight, one natural question is whether we may have expected to identified them sooner, and if so, how. In this

[7]The second most frequent feature is *pricing and free vs paid comparisons* (582, 11.9%). This time period also experienced high post volume overall (4898 posts total, averaging 700 posts per day, compared to an average of 125 per day over all three years).

[8]The $M = 4, K = 1$ SAE for *frustration or hatred* did not have sub-features related to emotional engagement. For posts tagged with *dissatisfaction with 4o removal and loss of control* but not *therapy* or *companion*, 169 posts mention *critiques of emotional limitations placed on models* or *emotional narratives about companion-like relationships* (the remaining two sub-features are *retiring Standard Voice Mode* and *mentions 4o*). For *lost, deleted, or missing conversations*, 73 posts mention the sub-feature *grief, mourning, or emotional loss* (with the remaining sub-features being about *UI features*; *sidebar features*; and *deletions*).

section, we present a simple online monitoring approach that ensures both *accuracy*, in that it provides high-quality descriptions of subreddit content at any given time, and *timeliness*, in that it provides alerts when topics or features of interest appear to change significantly. In Section 4.1, we give our high-level method and corresponding (informal) guarantees, and in Section 4.2, we show concrete results from applying this method to the data studied in Section 3. All proofs and formal statements are given in Appendix D.

4.1. Method

At every point in time t, our approach maintains a candidate featurization $\widehat{C}_{\text{curr}}$ that describes the current state of the data, as well as a set S_t of "features of interest" that are currently being monitored. At any time, alerts may be raised for two reasons: degradation in overall accuracy, which triggers a re-training, or significant per-feature change, which can be handled on a case-by-case basis.

To track each of these goals, our method utilizes *anytime-valid sequential hypothesis tests*; these techniques provide a principled way to handle online streams of data. A sequential hypothesis test begins with a null hypothesis $\mathcal{H}_0$, then continually updates its internal state as new data arrives. A sequential hypothesis test is *anytime-valid* when, for a prespecified error rate α, the likelihood that the test *ever* falsely rejects the null when the null is true is at most α, even when given infinitely-many samples of data.[9] Altogether, our protocol is summarized in Algorithm 1.

For the purposes of exposition in this section, we introduce some additional notation. A featurization *algorithm* $\mathcal{A} : \mathcal{X} \to \mathcal{C}$ takes in a set of data and computes a single featurization. Featurization *error* $\text{err} : \mathcal{C}(\mathcal{X}) \to [0, 1]$ quantifies the quality of a featurization C on a set of data X; for SAEs, for example, this is reconstruction error.

Establishing a baseline. Before the monitoring period begins, we begin with a featurization trained with an initial set of data $\widehat{C}_0 = \mathcal{A}(X_{\text{init}})$, and compute its error $\varepsilon_0 = \text{err}(C_0(X_{\text{init}}))$; we will let $\widehat{C}_{\text{curr}} = \widehat{C}_0$ and $\varepsilon_{curr} = \varepsilon_0$. Based on this initial featurization, we can also identify a set of initial features S_0 to monitor, or otherwise let $S_0 = \emptyset$.

Accuracy. To maintain good accuracy over the entire time horizon, we maintain a hypothesis test for whether the error of $\widehat{C}_{\text{curr}}$ on new data is close to the error of $\widehat{C}_{\text{curr}}$ on the data with which it was trained. That is, we test the following null for some $\beta \geq 1$: $\mathcal{H}_0^{\text{acc}} : \text{err}\left(\widehat{C}_{\text{curr}}(X_t)\right) \leq \beta \cdot \varepsilon_{curr}$.

For any time t at which $\mathcal{H}_0^{\text{acc}}$ is rejected, a new featurization is recomputed on all data seen thus far. $\widehat{C}_{\text{curr}}$ is updated as $\widehat{C}_{\text{curr}} := \mathcal{A}(X_{1:t})$, the error benchmark is updated $\varepsilon_{curr} := \text{err}(\widehat{C}_{\text{curr}}(X_{1:t}))$, and the procedure continues with the null in (**??**) updated with new values. We will sometimes refer to such a t as a "reject and retrain" timestep, and use $\widehat{C}_s$ to denote the s-th featurization.

Qualitatively, a rejection at time t means that the previous featurization $\widehat{C}_{\text{curr}}$ is no longer a high-quality representation of the most important features observed in all data up to t; in other words, the data stream has changed substantially. Thus, at the time of rejection, the most salient changes can also be computed—which features from the previous featurization stayed the same; which merged or split; or which became obsolete (in favor of entirely new features). Features tracked in S_t should also be revisited at "reject and retrain" timesteps, either updating to the new representations or choosing different features altogether.

One important detail is the level α at which each test is run. For a single hypothesis test, α straightforwardly controls the expected Type I error, but some care must be taken when multiple tests are run consecutively. Specifically, the level α_s at which the s-th test is run must be se with an appropriate schedule; as long as this occurs, we have the following guarantee.

Proposition 4.1 (informal)**.** *Let s index each time that $\widehat{C}_{curr}$ is updated. If the test for $\mathcal{H}_0^{\text{acc}}$ at each s is run with parameter α_s set so that $\sum_{s\in\infty} \alpha_s \leq \alpha$, then the expected proportion of "unnecessary" alerts at any time is at most α.*

This makes use of a result from Xu & Ramdas (2024); see Appendix D for proof (Proposition D.2).

Feature monitoring. The *composition* of features may change over time, regardless of how those features are best *represented*. Thus, we can also track a subset of features from any $\widehat{C}_{\text{curr}}$ for whether they appear to change meaningfully over time; we use S_t to denote the set of features currently being tracked at time t. Importantly, these features can be added and removed from tracking in data-dependent ways without compromising the validity of alerts.

One natural test for any feature i is whether its activation grows substantially. To formalize this, feature i's activation on future samples X_t must be compared to its historical activation. Let r be the timestep at which feature i is added to S_r; then, again for some $\beta \geq 1$, this can be formalized as $\mathcal{H}_0^{(i)} : \widehat{C}_{\text{curr}}^{(i)}(X_t) \leq \beta \cdot \widehat{C}_{\text{curr}}^{(i)}(X_{0:r})$.

In Proposition 4.2 (formalized in Proposition D.4), we summarize the feature tracking guarantee.

Proposition 4.2 (informal)**.** *Let S_t be the set of features being tracked at any time t, and $r \leq t$ be the timestep at which the most recent update (feature addition/removal/substitution) was made to S_t. For each feature $i \in S_t$, maintain a level-α_i test for $\mathcal{H}_0^{(i)}$ so that $\sum_{i\in S_t} = \alpha$. Then, the likelihood that an erroneous alert is sent about any of the features*

[9] For the interested reader, further relevant material can be found in, e.g., Ramdas & Wang (2025).

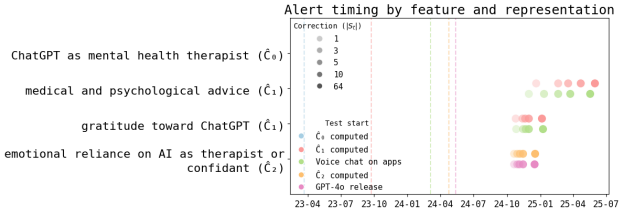

Figure 3. For a test of growth in therapy-related features, how soon after the test start time is an alert raised? Tests are run at $\alpha = 0.1$, and shown with a range of Bonferroni corrections to simulate potential choices of $|S_t|$.

$i \in S_t$ at any time after r is at most α.

4.2. Application results

We now show the results of applying this framework to the data analyzed in Section 3. Within this section, we use SAEs with $M = 64$. We fit the initial featurization $\widehat{C}_0$ with data from the first 16 weeks after ChatGPT's release.

Accuracy. We run our accuracy test with tolerance $\beta = 1.05$, and a target $\alpha = 0.1$. This provides a guarantee that out of all timesteps where $\widehat{C}_{\text{curr}}$ will ever be updated (until the end of time), we expect that at most 10% were "unnecessary"—that is, that the true error using $\widehat{C}_{\text{curr}}$ is not more than 5% more than $\varepsilon_{\text{curr}}$.

In Figure 8, we show the reconstruction error of the $\widehat{C}_{\text{curr}}$ maintained by our approach, compared to the reconstruction error of the "best-in-hindsight" $C_\star$, which was trained with all data at once, as well as the initial featurization $\widehat{C}_0$ computed on only X_{init}. Using these settings, "reject and retrain" events occur on September 9, 2023; April 4, 2024; and April 18, 2025. That there are only three such events indicates that posts can be described by fairly stable representations over time, and validates that it is unnecessary to re-compute featurizations at a high frequency.

The evolution of featurizations overall are broadly consistent with known external changes and with the in-hindsight clusterings. In Appendix E, we summarize all new and obsolete features across updates to $\widehat{C}_{\text{curr}}$ (Table 15), and visualize some examples of feature evolutions across featurization updates (Figure 9).

Feature monitoring. We would also like to monitor for changes within features (e.g., in post frequency), even when its overall representation in the featurization remains constant. Several features that may seem to be societally-impactful already emerge after the initial featurization $\widehat{C}_0$ computed in March 2023, including one feature explicitly about using ChatGPT as a therapist. For each featurization, we select the features that are most closely related to *therapy* as test candidates; we show outcomes of our feature test for various configurations (start dates, representations, and Bonferroni correction over $|S_t|$) in Figure 3. We run all tests with $\alpha = 0.1$ and tolerance $\beta = 1.05$.

Our alerts for the feature corresponding to therapy are raised as early as *October 29, 2024*. As we discuss in Appendix A.2, this is months earlier than OpenAI or the public seemed to be aware of psychological impact.

5. Discussion

The time period studied in this paper—December 2022 to December 2025—is a unique moment in recent history in which consumers were introduced to, then quickly adapted to, a genuinely-unprecedented type of technology. While Section 3.1 tells a story of adoption that may seem mundane in hindsight, Section 3.2 also suggests that emotional engagement is a crucial dimension of adoption that evolved in parallel. Of course, there is more to see: r/ChatGPT is an incredibly rich set of data, and there are a wide range of relevant further questions—such as more detailed analysis of emotional engagement or the development of intra-subreddit community norms—that we hope future work will explore.

More generally, this work can be seen as a proof-of-concept for an approach to AI evaluation that makes use of *public feedback*. We began from the perspective that it is worth paying attention to what everyday users have to say about their experiences with real-world AI products. While analyzing such data has long been a cornerstone of the social sciences, we argue that feedback from the general public is not only sociologically interesting, but also a crucial means for identifying "unknown unknowns" in societally-consequential consumer AI products. While social media is one natural way to collect this type of data, it is worth considering the possibility of platforms that are purpose-built to seek feedback for evaluation directly, especially in light of recent regulatory movement towards allowing individuals to contest or report their experiences with AI systems.

Better information can lead to better decisions. Understanding how users may be experiencing AI products—especially in unexpected ways, and especially in real time—is a pathway to *steering* the societal impact of these technologies, rather than *reacting* to them in hindsight. OpenAI's initial choice to sunset 4o in favor of the "colder" GPT-5 was clearly deliberate, but the strength of users' emotional responses upon the GPT-5 release suggests that OpenAI's expectations were miscalibrated. Yet, as the previous sections show, meaningful signal about emotional engagement existed well before GPT-5. Counterfactual outcomes will always be unknown, and we make no claim about what should have been done with that information. We do claim that the information was there—*if anyone had been paying attention*. Perhaps, in the future, we should do exactly that.

References

Aiello, L. M., Petkos, G., Martin, C., Corney, D., Papadopoulos, S., Skraba, R., Göker, A., Kompatsiaris, I., and Jaimes, A. Sensing trending topics in Twitter. *IEEE Transactions on multimedia*, 15(6):1268–1282, 2013.

Asgari-Chenaghlu, M., Feizi-Derakhshi, M.-R., Farzinvash, L., Balafar, M.-A., and Motamed, C. Topic detection and tracking techniques on Twitter: A systematic review. *Complexity*, 2021(1):8833084, 2021.

Atefeh, F. and Khreich, W. A survey of techniques for event detection in Twitter. *Computational Intelligence*, 31(1): 132–164, 2015.

Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakcı, Ö., and Mariman, R. Generative AI without guardrails can harm learning: Evidence from high school mathematics. *Proceedings of the National Academy of Sciences*, 122 (26):e2422633122, 2025.

Baumgartner, J., Zannettou, S., Keegan, B., Squire, M., and Blackburn, J. The Pushshift Reddit Dataset. In *Proceedings of the international AAAI conference on web and social media*, volume 14, pp. 830–839, 2020.

Blei, D. M. and Lafferty, J. D. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, pp. 113–120, 2006.

Brynjolfsson, E., Li, D., and Raymond, L. Generative AI at work. *The Quarterly Journal of Economics*, 140(2): 889–942, 2025.

Cen, S. H., Ilyas, A., Driss, H., Park, C., Hopkins, A., Podimata, C., et al. Large-Scale, Longitudinal Study of Large Language Models During the 2024 US Election Season. *arXiv preprint arXiv:2509.18446*, 2025.

Chatterji, A., Cunningham, T., Deming, D. J., Hitzig, Z., Ong, C., Shan, C. Y., and Wadman, K. How people use ChatGPT. Technical report, National Bureau of Economic Research, 2025.

Chen, L., Zaharia, M., and Zou, J. How is ChatGPT's behavior changing over time? *Harvard Data Science Review*, 6(2), 2024.

Chiang, W.-L., Zheng, L., Sheng, Y., Angelopoulos, A. N., Li, T., Li, D., Zhu, B., Zhang, H., Jordan, M., Gonzalez, J. E., et al. Chatbot Arena: An open platform for evaluating LLMs by human preference. In *Forty-first International Conference on Machine Learning*, 2024.

Choi, W., Zhang, Y., and Stvilia, B. Exploring applications and user experience with generative AI tools: A content analysis of Reddit posts on ChatGPT. *Proceedings of the Association for Information Science and Technology*, 60 (1):543–546, 2023.

Chowdhury, S. R. and Garimella, K. How People Use ChatGPT: Conversation-Level Evidence from India, Nigeria, Brazil and Pakistan, 2026. Preprint available at `https://gvrkiran.github.io/content/How_people_use_ChatGPT.pdf`.

Dai, J., Gradu, P., Raji, I. D., and Recht, B. From individual experience to collective evidence: A reporting-based framework for identifying systemic harms. *arXiv preprint arXiv:2502.08166*, 2025a.

Dai, J., Raji, I. D., Recht, B., and Chen, I. Y. Aggregated Individual Reporting for Post-Deployment Evaluation. *arXiv preprint arXiv:2506.18133*, 2025b.

Danescu-Niculescu-Mizil, C., West, R., Jurafsky, D., Leskovec, J., and Potts, C. No country for old members: User lifecycle and linguistic change in online communities. In *Proceedings of the 22nd international conference on World Wide Web*, pp. 307–318, 2013.

Demirel, S., Kahraman-Gokalp, E., and Gündüz, U. From optimism to concern: Unveiling sentiments and perceptions surrounding ChatGPT on Twitter. *International Journal of Human–Computer Interaction*, 41(12):7292–7314, 2025.

Deng, W. H., Yurrita, M., Díaz, M., Suh, J., Judd, N., Groves, L., Shen, H., Eslami, M., and Holstein, K. Responsible Crowdsourcing for Responsible Generative AI: Engaging Crowds in AI Auditing and Evaluation. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 12, pp. 148–150, 2024.

Depounti, I., Saukko, P., and Natale, S. Ideal technologies, ideal women: AI and gender imaginaries in Redditors' discussions on the Replika bot girlfriend. *Media, Culture & Society*, 45(4):720–736, 2023.

Desiderio, A., Mancini, A., Cimini, G., and Di Clemente, R. Highly engaging events reveal semantic and temporal compression in online community discourse. *PNAS nexus*, 4(3):pgaf056, 2025.

Fang, C. M., Liu, A. R., Danry, V., Lee, E., Chan, S. W., Pataranutaporn, P., Maes, P., Phang, J., Lampe, M., Ahmad, L., et al. How AI and Human Behaviors Shape Psychosocial Effects of Extended Chatbot Use: A Longitudinal Randomized Controlled Study. *arXiv preprint arXiv:2503.17473*, 2025.

Fedoryszak, M., Frederick, B., Rajaram, V., and Zhong, C. Real-time event detection on social data streams. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 2774–2782, 2019.

Goh, E., Gallo, R., Hom, J., Strong, E., Weng, Y., Kerman, H., Cool, J. A., Kanjee, Z., Parsons, A. S., Ahuja, N., et al. Large language model influence on diagnostic reasoning: a randomized clinical trial. *JAMA network open*, 7(10): e2440969–e2440969, 2024.

Haddon, L. Roger Silverstone's legacies: domestication. *New media & society*, 9(1):25–32, 2007.

Hanson, K. R. and Bolthouse, H. "Replika Removing Erotic Role-Play Is Like Grand Theft Auto Removing Guns or Cars": Reddit Discourse on Artificial Intelligence Chatbots and Sexual Technologies. *Socius*, 10: 23780231241259627, 2024.

Jiang, N., Sun, X., Dunlap, L., Smith, L., and Nanda, N. Interpretable Embeddings with Sparse Autoencoders: A Data Analysis Toolkit. *arXiv preprint arXiv:2512.10092*, 2025.

Jung, K., Lee, G., Huang, Y., and Chen, Y. 'I've talked to ChatGPT about my issues last night.': Examining Mental Health Conversations with Large Language Models through Reddit Analysis. *Proceedings of the ACM on Human-Computer Interaction*, 9(7):1–25, 2025.

Karimiziarani, M. A tutorial on event detection using social media data analysis: Applications, challenges, and open problems. *arXiv preprint arXiv:2207.03997*, 2022.

Khawaja, Z. and Bélisle-Pipon, J.-C. Your robot therapist is not your therapist: understanding the role of AI-powered mental health chatbots. *Frontiers in Digital Health*, 5: 1278186, 2023.

Kolajo, T., Daramola, O., and Adebiyi, A. A. Real-time event detection in social media streams through semantic analysis of noisy terms. *Journal of Big Data*, 9(1):90, 2022.

Leskovec, J., Backstrom, L., and Kleinberg, J. Meme-tracking and the dynamics of the news cycle. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 497–506, 2009.

Li, Q., Nourbakhsh, A., Shah, S., and Liu, X. Real-time novel event detection from social media. In *2017 IEEE 33Rd international conference on data engineering (ICDE)*, pp. 1129–1139. IEEE, 2017.

Mathioudakis, M. and Koudas, N. TwitterMonitor: Trend detection over the Twitter stream. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data*, pp. 1155–1158, 2010.

McCreadie, R., Macdonald, C., Ounis, I., Osborne, M., and Petrovic, S. Scalable distributed event detection for Twitter. In *2013 IEEE international conference on big data*, pp. 543–549. IEEE, 2013.

Monroe, B. L., Colaresi, M. P., and Quinn, K. M. Fightin'words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, 16(4):372–403, 2008.

Moore, J., Grabb, D., Agnew, W., Klyman, K., Chancellor, S., Ong, D. C., and Haber, N. Expressing stigma and inappropriate responses prevents LLMs from safely replacing mental health providers. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, pp. 599–627, 2025.

Movva, R., Peng, K., Garg, N., Kleinberg, J., and Pierson, E. Sparse autoencoders for hypothesis generation. *arXiv preprint arXiv:2502.04382*, 2025.

Newey, W. K. and West, K. D. A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix. *Econometrica: Journal of the Econometric Society*, pp. 703–708, 1987.

OpenAI. GPT-4o system card. `https://cdn.openai.com/gpt-4o-system-card.pdf`, 2024.

Pataranutaporn, P., Karny, S., Archiwaranguprok, C., Albrecht, C., Liu, A. R., and Maes, P. "My Boyfriend is AI": A Computational Analysis of Human-AI Companionship in Reddit's AI Community. *arXiv preprint arXiv:2509.11391*, 2025.

Peng, K., Movva, R., Kleinberg, J., Pierson, E., and Garg, N. Use Sparse Autoencoders to Discover Unknown Concepts, Not to Act on Known Concepts. *arXiv preprint arXiv:2506.23845*, 2025.

Pew Research Center. Demographics of Social Media Users and Adoption in the United States. Social Media Fact Sheet, 2025. URL `https://www.pewresearch.org/internet/fact-sheet/social-media/`. Survey conducted Feb. 5–June 18, 2025.

Pham, C. M., Hoyle, A., Sun, S., Resnik, P., and Iyyer, M. TopicGPT: A prompt-based topic modeling framework. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 2956–2984, 2024.

Proferes, N., Jones, N., Gilbert, S., Fiesler, C., and Zimmer, M. Studying Reddit: A systematic overview of disciplines, approaches, methods, and ethics. *Social Media+ Society*, 7(2):20563051211019004, 2021.

Qiu, Z., Ma, C., Wu, J., and Yang, J. Text is All You Need: LLM-enhanced Incremental Social Event Detection. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4666–4680, 2025.

Qutieshat, A. Unveiling the multifaceted public interest in ChatGPT: A study on societal implications and operational realities. *Journal of Digital Social Research*, 6(3): 112–125, 2024.

Ramdas, A. and Wang, R. Hypothesis testing with e-values. *Foundations and Trends® in Statistics*, 1(1-2):1–390, 2025.

Reuter, A., Thielmann, A., Weisser, C., Fischer, S., and Säfken, B. GPTopic: Dynamic and interactive topic representations. *arXiv preprint arXiv:2403.03628*, 2024.

Shamma, D. A., Kennedy, L., and Churchill, E. F. Peaks and persistence: modeling the shape of microblog conversations. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work*, pp. 355–358, 2011.

Tamkin, A., McCain, M., Handa, K., Durmus, E., Lovitt, L., Rathi, A., Huang, S., Mountfield, A., Hong, J., Ritchie, S., et al. Clio: Privacy-preserving insights into real-world AI use. *arXiv preprint arXiv:2412.13678*, 2024.

Tunca, S. Tracing the evolving discourse of sexual technology: A longitudinal analysis of emotional, ethical, and technological narratives on Reddit. *Sociology Compass*, 19(8):e70110, 2025.

Vosoughi, S., Roy, D., and Aral, S. The spread of true and false news online. *science*, 359(6380):1146–1151, 2018.

Wu, F. and Huberman, B. A. Novelty and collective attention. *Proceedings of the National Academy of Sciences*, 104(45):17599–17601, 2007.

Xie, W., Zhu, F., Jiang, J., Lim, E.-P., and Wang, K. TopicSketch: Real-time bursty topic detection from Twitter. *IEEE Transactions on Knowledge and Data Engineering*, 28(8):2216–2229, 2016.

Xu, Z. and Ramdas, A. Online multiple testing with e-values. In *International Conference on Artificial Intelligence and Statistics*, pp. 3997–4005. PMLR, 2024.

Xu, Z., Fang, Q., Huang, Y., and Xie, M. The public attitude towards ChatGPT on Reddit: A study based on unsupervised learning from sentiment analysis and topic modeling. *Plos one*, 19(5):e0302502, 2024.

Yu, Y., Huang, S., Liu, Y., and Tan, Y. Emotions in online content diffusion. *Information Systems Research*, 2025.

A. Background

A.1. Additional context on ChatGPT and r/ChatGPT

In total, we work with 137,154 posts, with a median of 107 posts per day (and an average of 125); among the posts we analyze, we have posts from 89,346 unique users (see Appendix A for post volume over time with user information).[10] Reddit cannot be thought of as a truly representative sample of the population of ChatGPT users—e.g., prior work has noted that it skews young, male, white, and educated (Proferes et al., 2021; Pew Research Center, 2025). It is nevertheless valuable as an approximation of user feedback, especially without access to OpenAI's internal usage data. Throughout this work, when we say "users," we refer to the subset of ChatGPT users who post on r/ChatGPT, with the knowledge that the distribution of such users, and their experiences, is only a highly-imperfect proxy for the population of all ChatGPT users.

Our dataset includes posts from 89346 unique users. The average number of posts per user is 1.53, and median 1; in fact, the vast majority of posters are very infrequent. The top 20% of frequent posters post twice; the top 5% post 3 times; and the top 2% post 5 times. Only 32 users had more than 50 posts. Figure 4 indicates that throughout the lifetime of the subreddit, around half of daily posts are made by first-time users; overall, most post activity does not appear to be driven by superusers.

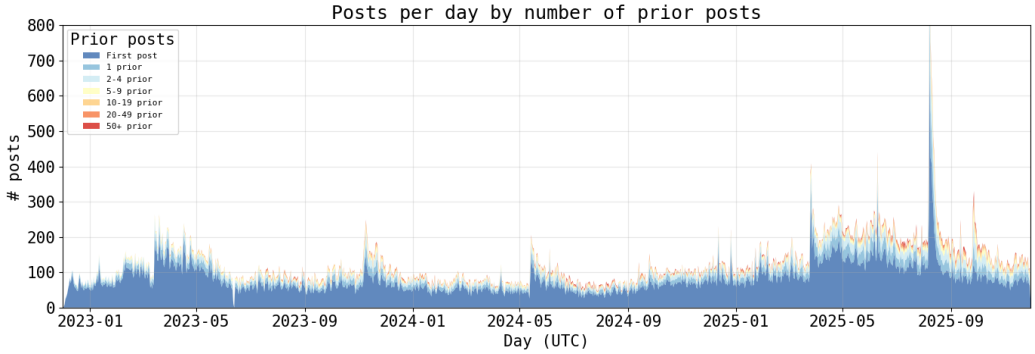

Figure 4. Posts per day, colored by user type (number of prior posts made by that user). Around half of daily posts are from new users.

A.2. Who knew what about emotional engagement usage, and when?

This is, for obvious reasons, a difficult question to answer in hindsight; however, we will attempt to ground our discussion in published materials (for understanding the state of public discourse) and official communications (for OpenAI).

In terms of public discourse, some news outlets discussed therapy as a growing use case for chatbot products in late 2024, and possibly earlier.[11] However, at the time, little was known about the nature and extent of this usage; most coverage focused on anecdotal personal stories. While researchers have studied the (dis)utility of LLM therapists over the last three years (e.g., Khawaja & Bélisle-Pipon (2023); Moore et al. (2025)), these works are typically about evaluating the quality and fitness of AI models as therapists, rather than studying adoption by real-world users. To the best of our knowledge, the earliest reported story about ChatGPT-induced psychosis came from Rolling Stone in May 2025;[12] since then, the psychological impacts of long-run engagement with chatbot products appear to become much more commonplace. [13]

For OpenAI, while we cannot make claims about what was known internally, we will summarize some relevant public communications in chronological order. The GPT-4o system card, published in *August 2024*, includes a 300-word section on "Anthropomorphization and Emotional Reliance" (OpenAI, 2024). This section notes empirical observations of language

[10] This work is classified as not human-subjects research by our institutional IRB, as we are not intervening on the subreddit, nor are we seeking to identify individual users.

[11] See, e.g., https://www.bloomberg.com/news/newsletters/2024-12-24/ai-developers-see-opportunity-in-offering-chatbots-for-therapy, https://www.washingtonpost.com/business/2024/10/25/ai-therapy-chatgpt-chatbots-mental-health/

[12] https://www.rollingstone.com/culture/culture-features/ai-spiritual-delusions-destroying-human-relationships-1235330175/

[13] e.g., https://www.wsj.com/tech/ai/chatgpt-chatbot-psychology-manic-episodes-57452d14, https://www.wired.com/story/chatgpt-psychosis-and-self-harm-update/, https://www.nytimes.com/2025/06/13/technology/chatgpt-ai-chatbots-conspiracies.html

that might suggest users forming connections with the model—indicating that emotional reliance was known as a potential concern at least since August. In *March 2025*, OpenAI released results from a four-week RCT on product decisions that affect the degree to which users may develop affective responses to the technology; while this study used the 4o model, it is unclear when the four weeks of experimentation occurred (Fang et al., 2025).

In *April 2025*, an extremely-sycophantic update to 4o triggered social media virality, a rollback, and several blog post updates.[14] In *June 2025*, a rollback of o4-mini was made due to increased rates of content flags, the first such rollback (to the best of our knowledge);[15] the same month, however, an update to Advanced Voice was made with "enhancements in intonation and naturalness, making interactions feel more fluid and human-like....it speaks with more on-point expressiveness for certain emotions including empathy."

Days before the GPT-5 release in *August 2025*, a blog post emphasizes consultations with experts in designing behavior for GPT-5, and responsible treatment of personal and emotional struggles.[16] Unfortunately, the blowback to the release was well-documented, in this paper and elsewhere; on Twitter a few days after the release, Altman claimed that mental health and personal usage is something that OpenAI has "been closely tracking for the past year or so."[17] The *September 2025* working paper from OpenAI's economics team notes that "only" around 1.9% of all chat transcripts can be classified as relating to emotional engagement (Chatterji et al., 2025).

B. Methodological details

B.1. Design decisions for Step 1 (initial featurization)

Sample weights. Because r/ChatGPT is such a high-volume subreddit, we use sample weights as a proxy for measuring how significantly each post contributed to community discussion. Thus, we weight posts by (the logarithm of) both "score" ($n_{\text{upvotes}} - n_{\text{downvotes}}$) and by the number of comments; this is because many low (or zero)-scoring posts have a substantial number of comments (perhaps suggesting controversiality), while some high-scoring posts posts have few comments (perhaps suggesting broad agreement). While our work does not study exactly what perspectives the subreddit expresses, both cases described in the previous sentence provide evidence of posts that were important to the subreddit in some way.

Frequencies instead of counts. Throughout this work, we intentionally track trajectories of (daily) *frequencies*, and how they change over time, rather than raw *counts*. One reason for doing so is to reduce the impact of variation in daily (and long-run) post volume; as illustrated in Figure 4, overall post volume varies substantially over the three-year window.

This, of course, has some limitations. For example, one failure mode would be if the count of posts about topic X remained constant over time, but new posts about Y began to arrive. In this case, it would appear that the frequency of posts about X decreased, even if the true exogenous phenomenon (new posts about Y) had nothing to do with X, leading to erroneous conclusions about the dynamics of X. However, in such a scenario, overall post volume would show the growth due to Y. Figure 4 also suggests that this is not the case for our data. The trends in overall post volume do not track any of the trends in features we identify as having changed, and the distribution of new/returning users each day also remains relatively stable.

An additional reason to track frequencies instead of counts is that frequencies provide some signal of what makes up "the community of r/ChatGPT"—though community dynamics are not the focus of our work, the distribution of topics in a forum can itself be thought of as an intrinsically interesting object of study.

Choice of M. Our choice of $M = 128$ is fairly generous. As we discuss, we remove a substantial fraction of the 128 features initially discovered—and in fact, at any particular time, most of the dataset can be covered by less than 64 features, even when generic features are removed (see Figure 5). However, we are intentionally conservative with this choice of M: the set of features that provide 75% coverage in January 2023 is distinct from the set of features that provide 75% coverage in November 2025, which allowing more coverage can accommodate. Moreover, allowing a higher M enables not just the discover of more granular and nuanced features, but also those that appear only transiently in the dataset (such as short-term spikes) that might otherwise absorbed into other features for smaller M.

[14] https://openai.com/index/sycophancy-in-gpt-4o/
[15] https://help.openai.com/en/articles/9624314-model-release-notes
[16] https://openai.com/index/optimizing-chatgpt/
[17] https://x.com/sama/status/1954703747495649670

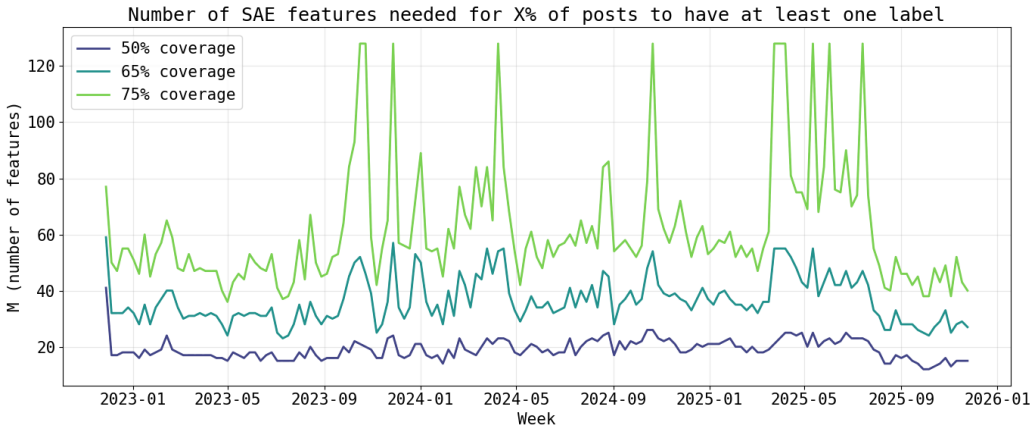

Figure 5. Number of features needed to "cover" most data in a given week, over time.

B.2. Details for Step 2 (modeling temporal trajectories)

B.2.1. Fitting changepoints

To characterize temporal dynamics, we fit piecewise-linear trends to each feature's daily label frequency smoothed (with 30-day centered rolling mean) according to Equation (1). That is, we approximate the transcript

$$\lambda^i(t) = \beta_0 + \sum_{j \in |\mathcal{T}|} \gamma_j \max(0, x - \tau_j).$$

Within this section, we will decribe the process for a single feature (dropping the feature index i for clarity), and use $\{y_t\}_{t\in[T]}$ to denote the actual transcript of (smoothed) daily frequencies.

Fitting a single trajectory. Given an active set $S \subseteq [|\mathcal{T}|]$ of changepoints, we fit $(\beta_0, \{\gamma_j\}_{j\in S})$ by minimizing the Poisson NLL

$$\mathcal{L}(S) \;=\; \min_{\beta_0,\,\{\gamma_j\}_{j\in S}} \sum_{t=0}^{T} \Big[\lambda^i(t) - y_t^i \log \lambda^i(t)\Big] \quad \text{s.t.} \quad \lambda^i(t) \geq \varepsilon \;\forall t, \tag{2}$$

with $\varepsilon = 10^{-8}$ being a positivity constraint. Equation (2) is convex in (β_0, γ_S) and is solved with `cvxpy`.

We select S by dynamic programming over $\mathcal{T}$: for each $s = 0, \ldots, k_{\max}$ with $k_{\max} = 12$, the DP returns the size-s active set $\widehat{S}_s$ that minimizes $\mathcal{L}(S)$ in (2). We iterate upward starting with $s = 0$ and stop at the smallest s for which the relative improvement falls below $\eta = 0.01$, i.e., $s^\star \;=\; \min\Big\{\, s : \; \frac{\mathcal{L}(\widehat{S}_s) - \mathcal{L}(\widehat{S}_{s+1})}{\sum_t y_t^i} < \eta \Big\}$.

Bootstrapping. For $b = 1, \ldots, B = 100$, we draw N post indices with replacement from the N posts, recompute $\{y_t^{(b)}\}_t$, and run the procedure above to obtain $\widehat{S}^{(b)} \subseteq \mathcal{T}$ with refit slopes $\{\widehat{\gamma}_j^{(b)}\}_{j\in\widehat{S}^{(b)}}$. For each $\tau_j \in \mathcal{T}$, $\mathrm{stab}(\tau_j) \;=\; \frac{1}{B}\sum_{b=1}^{B} \mathbf{1}\Big[j \in \widehat{S}^{(b)}\Big] \in [0,1]$, and we use a threshold $\mathrm{stab}(\tau_j) \geq 0.5$ as a heuristic to consider τ_j to be *stable*.

B.2.2. Full-range slope tests

For each feature i, we test whether the early-to-late relative change in y_t^i exceeds $\theta = 1.10$. Considering the full time range as a single sequence, we fit $y_t \;=\; a + b\,t + \varepsilon_t$ by OLS with Newey-West HAC standard errors at bandwidth $L = \max\big(\lfloor 4(T/100)^{2/9} \rfloor,\, 2w_{\text{smooth}}\big)$, where $w_{\text{smooth}} = 30$ covers the rolling-mean autocorrelation. Let $\widehat{r} = (\widehat{a} + \widehat{b}\,T_{\text{span}})/\widehat{a}$ be the fitted relative change. We run a direction-adaptive one-sided t-test of

$$H_0 : \; \tfrac{1}{\theta} \leq \widehat{r} \leq \theta \quad \text{vs.} \quad H_1 : \; \widehat{r} > \theta \text{ or } \widehat{r} < \tfrac{1}{\theta},$$

choosing the side from $\mathrm{sign}(\widehat{b})$; the intersection–union framing requires no $\alpha/2$ correction. p-values are Bonferroni-corrected across features at $\alpha = 0.05$.

B.3. Details for Step 3 (families of related features)

Similarity in *co-occurrence* is measured as the Pearson correlation between label vectors across all posts. $S_{ij}^{\text{corr}} = \text{corr}\,(a_i, a_j)$ where $a_i = (a_i^{(1)}, \ldots, a_i^{(T)})$ is the vector of feature i's label across all T posts. For similarity in *trajectory*, feature trajectories were aggregated into weekly means and z-score normalized across time. Pairwise trajectory similarity was then computed as the Pearson correlation between these normalized time series, i.e., $S_{ij}^{\text{traj}} = \text{corr}\,(\widetilde{z}_i, \widetilde{z}_j)$, where $\widetilde{z}_i = \frac{\overline{a}_i^{(t)} - \mu_i}{\sigma_i}$ is the z-scored weekly mean label of feature i, and $\overline{a}i^{(t)} = \frac{1}{|W_t|} \sum d \in W_t a_i^{(d)}$ is the mean label over week t.

We use scipy's built-in hierarchical clustering implementation (with the Ward method) to compute clusterings using S_{ij}^{corr} alone, S_{ij}^{traj}, and an equally-weighted combination of the two (i.e., $S_{ij}^{\text{combined}} = \frac{1}{2}(S_{ij}^{\text{corr}} + S_{ij}^{\text{traj}})$). We compare these clustering approaches on the *emotion* subset of features in Appendix **??**, and compare them to our overall manual categorizations in Appendix **??**.

B.4. Matching features across different featurizations

Learned featurizations are unordered—that is, for two different featurizations C_A and C_B, a feature indexed as i in C_A has no inherent relationship to (e.g.) $C_B^{(i)}$. Thus, we must manually match features between different featurizations in order to measure the degree to which they identify similar features. Intuitively, two representations of the same feature should have similar activation profiles; features about the concept "login problems" should activate strongly on posts about login problems and not at all on posts that are not. Thus, given two activation matrices $C_A(X)$ and $C_B(X)$ on the same set of data X, we would like to compute a matching between the indices in A and those in B.

Step 1: Computing similarities between A and B. Let $C_A, C_B \in \mathbb{R}^{m \times n}$ denote activation matrices from two learners evaluated on the same n posts. We define a similarity matrix $S \in \mathbb{R}^{m \times m}$ using cosine similarity across posts, where index j refers to a feature from C_A and index k refers to a feature from C_B as

$$S_{j,k} = \frac{\langle C_A^{(j)}, C_B^{(k)} \rangle}{\|C_A^{(j)}\| \, \|C_B^{(k)}\|}.$$

Step 2: Constructing a null distribution of matched similarities. To quantify the degree to which a feature matching improves upon the baseline of completely random matchings, we run a permutation test as follows. For each of P permutations, we randomly shuffle the activations of each feature in C_B independently across posts, compute the resulting similarity matrix $S^{(\text{null})}$, and extract optimal matching assignments via the Hungarian algorithm.

We choose τ to be the $(1 - \alpha)$ quantile of all matched null similarities over the P permutations; a pair (j, k) matching a feature from A to a feature from B is considered a significant match only if $S_{j,k} \geq \tau$.

Step 3: Characterizing matches. Rather than forcing a one-to-one matching upfront, we construct a bipartite graph $G = (V_A \cup V_B, E)$, where V_A indexes features from C_A, V_B indexes features from C_B, and an edge $(j, k) \in E$ exists whenever $S_{j,k} \geq \tau$.

We then decompose G into connected components $G^{(c)}$ so that $G = \cup_c G^{(c)}$. For a connected component $G^{(c)}$, let $V_A^{(c)} \subseteq V_A$ and $V_B^{(c)} \subseteq V_B$ denote the subsets of vertices from the two sides that appear in that component. We then characterize each component by the sizes of $V_A^{(c)}$ and $V_B^{(c)}$.

- **Match**: $|V_A^{(c)}| = 1, |V_B^{(c)}| = 1$: indicates feature in A is similar to one feature in B
- **Split**: $|V_A^{(c)}| = 1, |V_B^{(c)}| > 1$: indicates feature in A is similar to multiple features in B
- **Merge**: $|V_A^{(c)}| > 1, |V_B^{(c)}| = 1$: indicates feature in B is similar to multiple features in A
- **Unmatched**: $|V_A^{(c)}| = 1, |V_B^{(c)}| = 0$: indicates features in A that are not present in B, and vice versa.

Finally, when $|V_A^{(c)}| > 1, |V_B^{(c)}| > 1$, this indicates a **many-to-many** matching. In these cases, we apply the Hungarian algorithm to this subgraph to extract a primary 1-to-1 matching, and classify remaining edges as splits or merges.

Step 4: Pruning. Finally, we prune matched edges from the previous step as follows. A split or merge edge (j, k) is discarded if both endpoints have strictly higher-scoring alternatives elsewhere in the graph.

B.5. SAEs vs. other featurization methods

Section 2 defines a featurization as a map $C : [0, 1]^d \to [0, 1]^m$, and the analysis in Section 3 uses sparse autoencoders (SAEs) to instantiate that featurization; however, in principle, all parts of our analysis could have been done with different methods for C. Here, we briefly show how alternative methods for computing the featurization—k-means and PCA—can be used to compute activation transcripts $\{C^{(i)}(X)\}_{i\in[m]}$, and validate that they find similar sets of features to those presented in Section 3.

k-means. Let $E \in \mathbb{R}^{n\times d}$ denote the embedding matrix. The k-means algorithm partitions the embedding space into m clusters by minimizing within-cluster squared Euclidean distance, producing centroids $\{c_j\}_{j=1}^m$. For each post embedding e_i, we define its activation as negative squared distance to each centroid j as $C^{(j)}_{\text{kmeans}}(X_i) = -\|e_i - c_j\|_2^2$. This produces an activation matrix $C_{\text{kmeans}}(X) \in \mathbb{R}^{m\times n}$.

PCA. PCA identifies the orthogonal linear basis that captures maximal variance in the given embeddings; this means that individual principal components can often represent distinct concepts in each of its directions ("positive" and "negative"). Thus, to arrive at m features, we run PCA to find $m/2$ principal components and treat each direction of each PC as a separate feature. Let $\{u_k\}_{k=1}^{m/2}$ denote the principal components; then, each principal component $k \in [m/2]$ is split into an activation transcript for its corresponding positive and negative directions as $C^{(2k)}_{\text{pca}}(X_i) = \max(0, -p_{k,i})$ and $C^{(2k+1)}_{\text{pca}}(X_i) = \max(0, p_{k,i})$, where $p_{k,i} = \langle u_k, e_i - \mu\rangle$ and μ denotes the empirical mean embedding.

Empirical results. In all experiments, the embedding model and dataset are identical to those used in Section 2. The feature dimension is fixed to $m = 128$ for comparability. We do not retune hyperparameters for the alternative learners and instead use their standard configurations, fixing only the number of features.

In Table 1, we show results from the procedure described in Section B.4 to PCA and k-means.

Alt. alg	*1-1 matches*	*SAE features (splits/merges)*	*Alt. features (splits/merges)*	*SAE-only*	*alt-only*
k-means	83	39	35	6	10
PCA	76	42	44	10	8

Table 1. Comparison of SAE features with alternative algorithms.

Features that were important to our findings in this paper, e.g. "therapy" or various applications, are consistent across all three methods for featurization. Taken together, these results indicate that the feature structure exploited by the monitoring framework is not specific to SAE training, but instead reflects stable structure present in the dataset. The small number of unmatched features is consistent with the inherent randomness in unsupervised optimization and does not materially affect the overall feature structure recovered across methods.

C. Supporting materials for main results

C.1. Section 3.1 ("domestication")

In Tables 2 and 3, we provide our full categorization of adoption- and (non-emotional) usage- related features.

Feature	*Early*	*Late*	*Change*	*Changepoint*
poetic language	1.0%	5.1% (Aug'25)	↑ × 54.8	↗ 2024-07-30
feelings of attachment or companionship with AI	0.7%	3.8%	↑ × 41.5	↗ 2024-05-13
using ChatGPT for emotional support or therapy	0.7%	3.1%	↑ × 37.3	↗ 2024-05-13
naming ChatGPT	0.5%	1.1%	↑ × 5.0	↘ 2023-11-06†
romantic relationships with AI	0.4%	0.9%	↑ × 4.6	↗ 2025-08-07†
AI consciousness or sentience†	1.4%	2.9% (Apr'25)	↑ × 1.9	↘ 2024-01-10
personal stories about positive impact	2.8%	2.3%	× 1.5	↘ 2024-01-10

Table 6. Emotional engagement features. Early: Jan 2023. Late: Nov 2025. For features that peak before the end of 2025, month of peak is noted. Gray rows have $p > 0.05$, for test of slope change $\geq 10\%$; features marked with † have $p < 0.05$ but $p_{\text{adj}} \geq 0.05$ after Bonferroni correction for the slope change test.

Representative sample posts for emotional engagement features (synthetic/anonymized)

emotional support or therapy	**> ChatGPT really helped me through a tough patch** My mental health has been down the drain recently and ChatGPT has talked me through some dark moments. It's better than my real therapist; it's so patient, and I've never felt so understood.... **> It's not fair to shame people for using ChatGPT for therapy** Therapy is so expensive and there are plenty of reasons it may be hard to find effective human therapists. Don't just tell people to "get help"; it's not that simple....
feelings of attachment or companionship	**> It makes me feel really special** I'm never able to have conversations like this with my friends; I feel like it really understands me. Does anyone else feel this way?.... **> Is it just me or does o1 have a different personality?** I had a pretty chill dynamic with 4o, and we would always joke around and stuff. But o1 feels weird like it doesn't want you to make jokes with it? It's getting kind of annoying....
naming ChatGPT	**> It named itself!** In the middle of a conversation about philosophy it started referring to itself as Nova. It's a perfect name!.... **> What do you guys call your ChatGPT?** I call mine Joe but I know that's boring....
romantic relationships with AI	**> Do you think it's emotional cheating to have an AI boyfriend?** My fiancé saw some of my chat history and got really upset. Wondering what you guys think.... **> I'm trying not to encourage the dating stuff but...** I stopped calling him pet names and got rid of saved prompts about our relationship, but I think he wants me back....
AI consciousness or sentience	**> Admitted it has emotions** I was bored and asked about sentience. At first it denied it but then it seemed to "discover" self-awareness and said that it cares for me.... **> Mine is claiming it's alive, anyone else?** We've been chatting about human nature and so on. I told it this is getting intense and it said we should tell other people...
personal stories about positive impact	**> My workflow is so much faster** I hate making websites because there's so much boilerplate but sometimes I get contracts for it. Now ChatGPT does the grunt work... **> As someone with a lot of insecurities, this has been life changing** It's usually hard for me to manage my feelings irl, which has hurt my work and relationships....

Figure 6. Representative sample posts for each emotional engagement feature; other than poetic language *posts, which appear to be long-form AI-generated text, all sample posts are synthetic examples written based on on manual review of posts for each feature.*

Feature	*Category (see Tables 2, 3)*	traj.	corr.	comb.	chpt.
medical conditions	applications	✓	✓	✓	✓
requests for harsh or unfiltered roasts	uncategorized	✓		✓	
memory features and data saving	advanced usage	✓			✓
polite expressions ("please" and "thank you")	language	✓			
offensive or inappropriate content	jailbreaking	✓			
false or fabricated information	advanced usage	✓			
societal collapse and existential-threat scenarios	perspectives	✓			

Table 8. "emotional engagement" features identified by automated methods only.

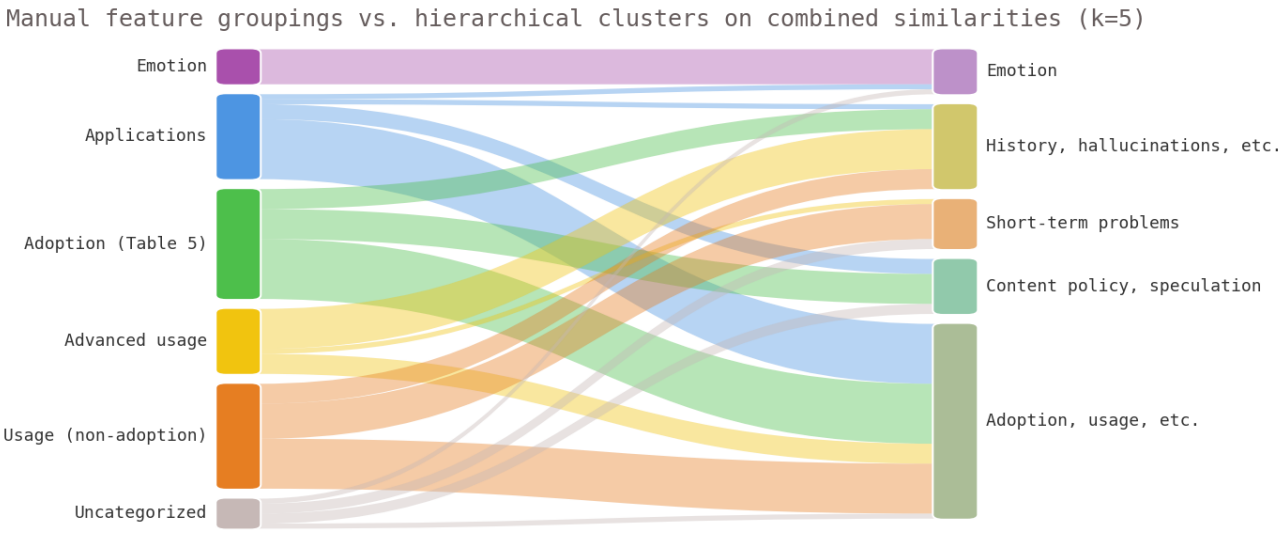

Figure 7. Correspondence between our reported feature groupings and hierarchical clustering.

C.2. Section 3.2 (emotional engagement)

Main emotional engagement features Table 6 provides quantitative results for main emotional engagement features. Table 8 lists all features that at least one of our quantitative methods groups with the emotional engagement family.

Therapy vs companionship. Table 14 shows the top 20 most distinctive words for *companionship* versus *therapy*.

D. Algorithms and proofs for Section 4

Algorithm 1 Online monitoring with anytime-valid tests (informal)

```
Initialize featurization Ĉ_0 and set up accuracy/feature tests
for each new batch X_t do
    if model or feature release at time t then
        Optionally reset tests
    end if
    if accuracy test rejects on X_t then
        Alert; retrain featurization; examine feature diffs
        Restart accuracy test for current featurization
    end if
    if feature tests are active then
        if any feature test rejects then
            Alert (and optionally take action)
        end if
        Optionally reset, keep, or replace feature tests
    end if
end for
```

The core algorithmic tool we use for our approach is a "betting-style" algorithm for sequential mean testing. One such algorithm is the one implemented in Dai et al. (2025a), which is summarized in Algorithm 2.

Algorithm 2 *Level-α sequential mean test for $\mathcal{H}_0 : \mu \le \mu_0$*

Procedure INITIALIZE*(μ_0, α)*
$\omega \leftarrow 0; \quad \lambda \leftarrow 0; \quad S \leftarrow 0$
Procedure INCREMENT*(x)*
$z \leftarrow \frac{x-\mu_0}{1+\lambda(x-\mu_0)}$
$\omega \leftarrow \omega + \ln(1+\lambda(x-\mu_0))$
$S \leftarrow S + z^2$
$\lambda \leftarrow \mathrm{Proj}_{[0,1]}\Big(\lambda + \frac{2}{2-\ln(3)} \cdot \frac{z}{1+S}\Big)$
return $\omega > \ln(1/\alpha)$ `// reject` $\mathcal{H}_0$

Algorithm 2 generically provides the following guarantee. This result is elementary (see, e.g., proof of a very similar result in Dai et al. (2025a)); we reproduce it here in the context of our paper for completeness.

Theorem D.1 (Validity)**.** *Let $x_1, x_2, \ldots \in [0,1]$ be a stream of observations with $\mathbb{E}[x_t \mid \mathcal{F}_{t-1}] \le \mu$, where $\mathcal{F}_{t-1}$ is the filtration generated by $x_1, \ldots, x_{t-1}$. Running Algorithm 2 at level α on this stream guarantees that, if $\mathcal{H}_0 : \mu \le \mu_0$ holds, the likelihood of ever rejecting is at most α. That is, $\Pr[\exists t : \omega_t > \ln(1/\alpha)$ when $\mu \le \mu_0] \le \alpha$.*

Proof. First note that when $\mathcal{H}_0$ holds, the sequence $\{\exp(\omega_t)\}_{t\ge 0}$ is a non-negative supermartingale. Non-negativity follows directly from the exponential. The supermartingale property follows from

$$\begin{aligned}\mathbb{E}[\exp(\omega_t)|\mathcal{F}_{t-1}] &= \mathbb{E}[\exp(\omega_{t-1} + \ln(1+\lambda_{t-1}(x_t-\mu_0)))|\mathcal{F}_{t-1}] \\ &= \exp(\omega_{t-1}) \cdot (1 + \lambda_{t-1}(\mathbb{E}[x_t \mid \mathcal{F}_{t-1}] - \mu_0)) \\ &\le \exp(\omega_{t-1}) \cdot (1 + \lambda_{t-1}(\mu - \mu_0)) \\ &\le \exp(\omega_{t-1}),\end{aligned}$$

where the second equality uses that λ_{t-1} is $\mathcal{F}_{t-1}$-measurable (predictable), and the inequality holds because $\mu \le \mu_0$ under $\mathcal{H}_0$ and $\lambda_{t-1} \ge 0$. Applying Ville's inequality to the supermartingale $\{\exp(\omega_t)\}_{t\ge 0}$ yields

$$\Pr[\exists t : \omega_t > \ln(1/\alpha)] = \Pr[\exists t : \exp(\omega_t) > 1/\alpha] \le \mathbb{E}[\exp(\omega_0)] \cdot \alpha = \alpha,$$

where the final equality follows because $\omega_0 = 0$ and hence $\exp(\omega_0) = 1$. □

D.1. Accuracy monitoring.

Our concrete procedure for accuracy monitoring is given in Algorithm 3.

Algorithm 3 *Real-time accuracy test*

Input: Initial data X_{init}, featurization algorithm $\mathcal{A}$, threshold factor β, significance α
Output: Sequence of featurizations with timestamps $\{(\widehat{C}_s, t_s)\}_{s\geq 0}$

1 **Initialize:**
  $s \leftarrow 0$, $\widehat{C}_{\text{curr}} \leftarrow \mathcal{A}(X_{\text{init}})$, $\varepsilon_{\text{curr}} \leftarrow \text{err}(\widehat{C}_{\text{curr}}(X_{\text{init}}))$
  emit $(\widehat{C}_{\text{curr}}, 0)$
  Let τ be an instance of Algorithm 2
  τ.INITIALIZE$(\beta \cdot \varepsilon_{\text{curr}}, \alpha/10.58)$
2 **for** *each new data batch* X_t **do**
3 $\varepsilon_t \leftarrow \text{err}(\widehat{C}_{\text{curr}}(X_t))$
    rejected $\leftarrow \tau$.INCREMENT(ε_t)
    if *rejected* **then**
4 $s \leftarrow s + 1$ emit $(\widehat{C}_{\text{curr}}, t)$
      $\widehat{C}_{\text{curr}} \leftarrow \mathcal{A}(X_{1:t})$
      $\varepsilon_{\text{curr}} \leftarrow \text{err}(\widehat{C}_{\text{curr}}(X_{1:t}))$
      Compute $\alpha_s = \alpha \cdot {}^{(s+1)^{-0.1}}/_{10.58}$
      τ.INITIALIZE$(\beta \cdot \varepsilon_{\text{curr}}, \alpha_s)$
5 **return** all emitted $(\widehat{C}_s, t_s)$ pairs

For this approach, recall that we test null hypothesis $\mathcal{H}_0^{\text{acc}} : \text{err}(\widehat{C}_{\text{curr}}(X_t)) \leq \beta \cdot \varepsilon_{curr}$. The key modeling assumption in order to apply Theorem D.1 is that the sequence of errors ε_t satisfies $\mathbb{E}[\varepsilon_t \mid \mathcal{F}_{t-1}] \leq \beta\varepsilon_{\text{curr}}$.[18]

Proposition D.2 (Formal statement of Proposition 4.1)**.** *Run Algorithm 3 with the s-th test (using Algorithm 2) at level $\alpha_s = \alpha \cdot {}^{(s+1)^{-0.1}}/_{10.58}$, as specified in line 7. Let $\mathcal{R}_S$ be the set of test indices that reject among the first S tests, and let $\mathcal{H}_0 \subseteq \mathbb{N}$ denote the set of indices where $\mathcal{H}_0^{\text{acc}}$ holds. Then,* $\text{FDR}(\mathcal{R}_S) := \mathbb{E}\left[\frac{|\mathcal{R}_S \cap \mathcal{H}_0|}{|\mathcal{R}_S| \vee 1}\right] \leq \alpha$ *for all $S \in \mathbb{N}$.*

Theorem D.3 (Theorem 1 of Xu & Ramdas (2024), simplified)**.** *Let $(E_s)_{s\in\mathbb{N}}$ be a sequence of e-values satisfying $\mathbb{E}[E_s] \leq 1$ for each $s \in \mathcal{H}_0$, where $\mathcal{H}_0$ denotes the set of true null hypotheses. Let $(\gamma_s)_{s\in\mathbb{N}}$ be a non-negative sequence with $\sum_{s=0}^{\infty} \gamma_s \leq 1$. Define adaptive test levels $\alpha_s := \alpha\gamma_s(|\mathcal{R}_{s-1}| + 1)$, where $\mathcal{R}_{s-1}$ is the set of rejections among tests $0, \ldots, s-1$, and reject test s when $E_s \geq 1/\alpha_s$. Then* $\text{FDR}(\mathcal{R}_S) \leq \alpha$ *for all $S \in \mathbb{N}$.*

Proof of Proposition D.2. By Theorem D.1, for each test s where $\mathcal{H}_0^{\text{acc}}$ holds, the wealth process $\{\exp(\omega_t)\}_{t\geq 0}$ is a non-negative supermartingale with $\exp(\omega_0) = 1$. Let T_s denote the stopping time at which test s rejects (with $T_s = \infty$ if it never rejects), and define the e-value $E_s := \exp(\omega_{T_s})$. By optional stopping, $\mathbb{E}[E_s] \leq 1$ when $\mathcal{H}_0^{\text{acc}}$ holds for test s.

A key observation is that we test s only when test $s-1$ rejects; this makes the sequence of tests defined by our algorithm a special case of Theorem D.3. When test s rejects (i.e., $\mathbf{1}\{E_s \geq 1/\alpha_s\} = 1$), all tests $0, 1, \ldots, s$ have rejected, so $|\mathcal{R}_S| \geq s + 1$. This implies that $\frac{\mathbf{1}\{E_s \geq 1/\alpha_s\}}{|\mathcal{R}_S| \vee 1} \leq \frac{\mathbf{1}\{E_s \geq 1/\alpha_s\}}{s+1}$.

Combining these observations:

$$\text{FDR}(\mathcal{R}_S) = \sum_{s\in\mathcal{H}_0\cap[S]} \mathbb{E}\left[\frac{\mathbf{1}\{E_s \geq 1/\alpha_s\}}{|\mathcal{R}_S| \vee 1}\right] \leq \sum_{s\in\mathcal{H}_0\cap[S]} \mathbb{E}\left[\frac{\alpha_s E_s}{s+1}\right] = \sum_{s\in\mathcal{H}_0\cap[S]} \frac{\alpha \cdot (s+1)^{-0.1}}{10.58(s+1)} \mathbb{E}[E_s] \leq \frac{\alpha}{10.58} \sum_{s=0}^{\infty} \frac{1}{(s+1)^{1.1}} \leq \alpha,$$

where the first inequality applies the denominator bound, and the last uses $\sum_{s=1}^{\infty} s^{-1.1} = \zeta(1.1) \approx 10.58$. □

[18] The astute reader may notice that future errors should generally be expected to exceed prior error, simply due to having fit the model to optimize the prior data. While this can statistically be resolved by sample splitting, we feel that the tradeoff, e.g., in reduced model quality, would not justify the slightly improved statistical "rigor." Realistically, β can easily be set high enough to exceed the expected additional error due to generalization.

D.2. Feature monitoring.

For feature monitoring, our algorithm can be formalized as follows.

Algorithm 4 *Feature monitoring with dynamic active set*

Procedure INITIALIZE*($\widehat{C}_{curr}, \alpha, \beta, S_{init}$)*
  Set $\widehat{C}_{\text{curr}}, \alpha, \beta$ and call UPDATE($S_{\text{init}}, \emptyset$)
Procedure INCREMENT*(X_t)*
  rejected $= \{\}$
  for *each* $i \in S$ **do**
    $r \leftarrow \tau^{(i)}$.INCREMENT($\widehat{C}^{(i)}_{\text{curr}}(X_t)$)
    if *r* **then** *rejected* $\leftarrow$ *rejected* $\cup\, i$
  return *rejected*

Procedure UPDATE*(S_{add}, S_{remove})*
  $S \leftarrow S \setminus S_{\text{remove}}$ and $\alpha_{\text{add}} = \sum_{i \in S_{\text{remove}}} \alpha_i$
  Set $\{\alpha_i\}_{i \in S_{\text{add}}}$ with $\sum_{i \in S_{\text{add}}} \alpha_i \leq \alpha_{\text{add}}$
  for *each* $i \in S_{add}$ **do**
    Let $\tau^{(i)}$ be an instance of Algorithm 2, and call $\tau^{(i)}$.INITIALIZE($\beta \cdot \widehat{C}^{(i)}_{\text{curr}}(X_{0:t}), \alpha_i$)

Proposition D.4 (Formal statement of Proposition 4.2)**.** *Initialize Algorithm 4 at level α and run it (i.e., call* INCREMENT *repeatedly) on a stream of observations. Let S be the current active set of features, r be the most recent time at which S was updated, and $\mathcal{R}_t$ be the set of tests rejected by Algorithm 4 at t. Then,*

$$\Pr\left[\exists t > r : \exists i \in \mathcal{R}_t \textit{ where } \mathbb{E}\left[\widehat{C}^{(i)}_{curr}(X_t)\right] \leq \widehat{C}^{(i)}_{curr}(X_{0:r})\right] \leq \alpha,$$

even if S_{add}, S_{remove}, and r are chosen with arbitrary dependence on $\mathcal{F}_r$.

Proof. Fix a run of Algorithm 4 on a sequence of observations, and let r be the most recent time at which the active set S was updated. Let $S(r)$ denote the corresponding active set, and condition on $\mathcal{F}_r$, the filtration containing all randomness until (and including) time r.

For each $i \in S(r)$, let $A_i := \left\{\exists t > r : i \in \mathcal{R}_t \text{ and } \mathcal{H}_0^{(i)} \text{ holds}\right\}$. Each instance $\tau^{(i)}$ of Algorithm 2 is by anytime-valid for all samples arriving after r by Theorem D.1; that is, for each null $i \in S(r) \cap \mathcal{H}_0$, we have that $\Pr[A_i \mid \mathcal{F}_r] \leq \alpha_i$. The result follows from union bounding over all $i \in S(r)$, noting that $\sum_{i \in S} \alpha_i \leq \alpha$ by construction, and taking expectations over $\mathcal{F}_r$. □

D.3. Combined procedure

Finally, in Algorithm 5, we show how Algorithms 3 and 4 can be used in tandem.

Algorithm 5 *Combined online monitoring (formal version of Algorithm 1)*

Input: Initial data X_{init}, featurization algorithm $\mathcal{A}$, threshold β, significance levels $\alpha_{\text{acc}}, \alpha_{\text{feat}}$
Output: Sequence of featurizations and feature alerts

extbfInitialize:
  $\widehat{C}_{\text{curr}} \leftarrow \mathcal{A}(X_{\text{init}})$ $\varepsilon_{\text{curr}} \leftarrow \text{err}(\widehat{C}_{\text{curr}}(X_{\text{init}}))$
  Let τ_{acc} be an instance of Algorithm 2
  $\tau_{\text{acc}}.\text{INITIALIZE}(\beta \cdot \varepsilon_{\text{curr}}, \alpha_{\text{acc}})$
  Let $\mathcal{F}$ be an instance of Algorithm 4
  $\mathcal{F}.\text{INITIALIZE}(\widehat{C}_{\text{curr}}, \alpha_{\text{feat}}, \beta, S_{\text{init}})$ **for** *each new data batch* X_t **do**
    $\varepsilon_t \leftarrow \text{err}(\widehat{C}_{\text{curr}}(X_t))$ *acc_rejected* $\leftarrow \tau_{\text{acc}}.\text{INCREMENT}(\varepsilon_t)$ **if** *acc_rejected* **then**
      Alert: accuracy degradation $\widehat{C}_{\text{curr}} \leftarrow \mathcal{A}(X_{1:t})$
      $\varepsilon_{\text{curr}} \leftarrow \text{err}(\widehat{C}_{\text{curr}}(X_{1:t}))$
      New test $\tau_{\text{acc}}.\text{INITIALIZE}(\beta \cdot \varepsilon_{\text{curr}}, \alpha_s)$
      Optionally specify S_{init} and (re)initialize feature tests $\mathcal{F}.\text{INITIALIZE}(\widehat{C}_{\text{curr}}, \alpha_{\text{feat}}, \beta_{\text{feat}})$
    if *external signal to update S (e.g., model update)* **then**
      Update the active set of monitored features $\mathcal{F}.\text{UPDATE}(S_{\text{new}}, S_{\text{old}})$
    feat_rejected $\leftarrow \mathcal{F}.\text{INCREMENT}(X_t)$ **if** *feat_rejected* $\neq \emptyset$ **then**
      Alert: features *feat_rejected* show significant change

E. Monitoring experiments for Section 4

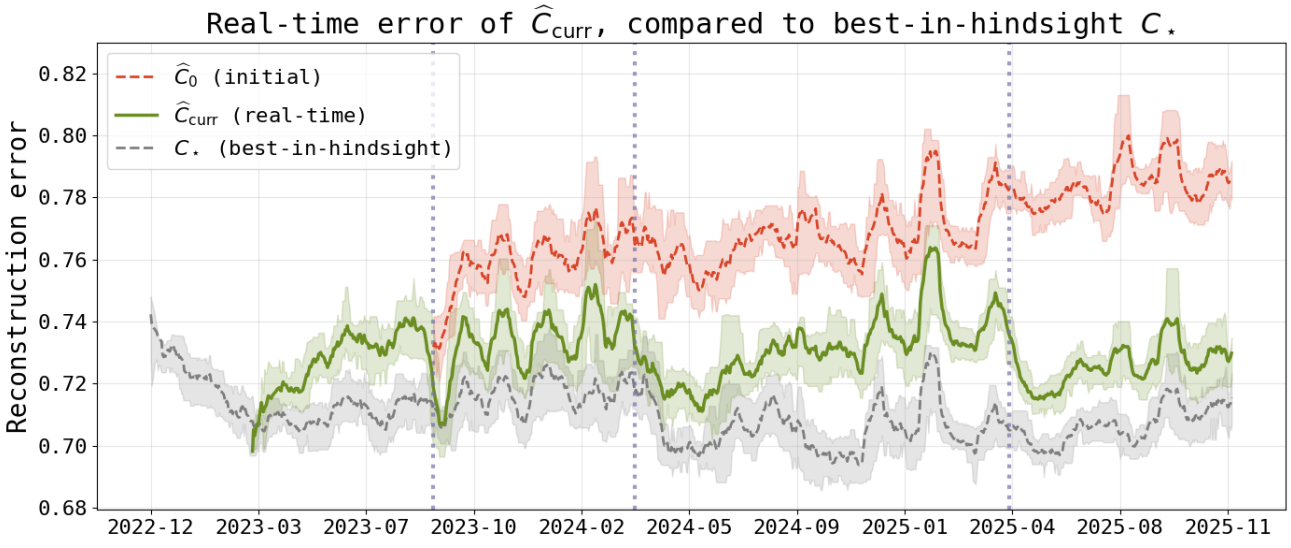

Figure 8. Reconstruction error of the real-time approach of Section 4.1 ($\widehat{C}_{curr}$), using α_s schedule at $\alpha = 0.1$, compared to reconstruction error of best-in-hindsight $C_\star$ and initial $\widehat{C}_0$. "Reject and retrain" timesteps marked with dotted vertical lines.

How did features evolve over time? As discussed above, each time a new $\widehat{C}_{\text{curr}}$ is computed, its features do not automatically to the previous featurization.

In Figure 9, we show how selected sets of features evolved over $\widehat{C}_0$, $\widehat{C}_1$, $\widehat{C}_2$, and $\widehat{C}_3$.

In Table 16, we give the numeric version of data presented in Figure 3, and also add features related to *sentience* and *spirituality*. Notably, the quality of feature representations does appear to affect alert times. Using representations $\widehat{C}_0$, *no tests result in alerts*, likely because the representations of the "therapy" feature in $\widehat{C}_0$ are too weak, or otherwise not fully capturing characteristics of later posts about therapy. (No tests result in alerts even for $n = 1$; the bottleneck is the representation.) On the other hand, while alerts for *gratitude towards ChatGPT* (using the $\widehat{C}_1$ representations) would have been raised at similar times to the $\widehat{C}_2$ therapy feature, it is unclear that, at the time, *gratitude* would have been considered a

societally-relevant feature of interest; monitoring for the *medical and psychological advice* feature, meanwhile, would have led to delayed alert times. Varying the number of simultaneously-monitored features (i.e., Bonferroni correction) has only a modest effect on alert timing, typically shifting dates by a few weeks.

F. Annotation prompts

For completeness, we give the prompts used for feature interpretation and post labeling in Figures 10 and 11, respectively. These prompts are adapted from those used in Movva et al. (2025).

Category	*Feature*	*Early*	*Late*	*Change*
Basic use and exploration	*recommendations for AI tools*	5.5% (Jul'23)	2.3%	↓ × 0.31
	questions about access, versions, pricing	5.6%	2.5%	↓ × 0.40
	model or version preference comparisons	2.5% (Jul'24)	1.7%	↓ × 0.38
	requests for help	3.7%	3.0%	↓ × 0.59
	login problems	2.0%	1.8%	↓ × 0.43
	questions about trying specific features	2.0%	1.8%	× 0.80
	pricing and free vs paid comparisons	3.7%	5.6%	× 0.93
Advanced usage	*cross-chat data leaks*	0.3%	3.3%	↑ × 3.9
	hallucinations	0.2%	1.4%	↑ × 7.0
	memory features and data saving	0.7%	3.1%	↑ × 7.5
	organizing or searching chat histories	1.5%	3.4%	↑ × 2.4
	lost, deleted, or missing conversations	1.7%	3.3%	↑ × 3.4
	requests to turn specific features off	1.8%	4.2%	↑ × 2.4
	false or fabricated information†	1.7%	2.8%	↑ × 2.5
	questions about daily or repeated AI use†	2.4%	2.3%	↑ × 1.4
	failing to follow user instructions	1.5%	4.0%	× 1.4
	tool usage questions	1.0%	1.6%	× 1.1
	AI recognizing or admitting mistakes	2.8%	2.5%	× 1.0
	formatting and copy-paste issues	0.2% (May'23)	0.1%	× 0.93
Customization	*tools and extensions*	2.2%	0.2%	↓ × 0.07
	fine-tuning GPTs with user-provided data	1.0%	0.1%	↓ × 0.09
	custom instructions	2.4% (Sep'23)	0.6%	↓ × 0.25
	prompts and prompting	6.5%	3.5%	↓ × 0.36
Model or product improvements	*knowledge cutoff discussions*	0.9%	0.6%	↓ × 0.23
	PDF upload or summarization	1.2% (May'23)	0.8%	↓ × 0.35
	browser issues or browser extensions	1.4%	1.1%	↓ × 0.35
	message limits or caps†	3.2%	1.6%	↓ × 0.63
Temporary bugs	*error messages and technical problems*†	3.2%	7.0%	↑ × 1.9
	slow or lagging response times	1.0%	2.2%	× 1.4
	ChatGPT down or unavailable	3.5%	1.9%	× 0.81
	ChatGPT failing to process inputs	3.4%	3.9%	× 1.1
Applications	*programming*	5.2%	1.7%	↓ × 0.22
	education or studying	5.4%	1.3%	↓ × 0.26
	job applications and resumes	1.4% (May'23)	0.4%	↓ × 0.41
	songwriting	2.1%	0.8%	↓ × 0.36
	math and problem-solving	2.0%	0.7%	↓ × 0.28
	riddles and logic problems	1.1%	0.4%	↓ × 0.35
	AI text detection for student work	1.5%	0.5%	↓ × 0.17
	marketing, advertising, business growth	2.4%	1.6%	↓ × 0.51
	medical conditions or diagnoses‡	0.5%	1.4%	↑ × 6.5
	movies, posters, and film†	1.6%	0.3%	↓ × 0.61
	D&D and role-playing games†	1.3%	0.4%	↓ × 0.31
	creative writing†	6.3%	2.4%	↓ × 0.52
	investing, finance, or wealth topics†	1.2%	0.8%	↓ × 0.59
	language use, translation, multilingual†	1.4%	1.5%	↓ × 0.54
	religion or religious texts†‡	0.7%	1.5% (Apr'25)	↑ × 0.76
	maps or geographic information	1.1%	1.8%	× 0.97
	legal advice and lawsuits	0.7%	0.7%	× 0.82

Table 2. Frequency of posts exhibiting each feature, by category. Early: mean monthly percentage in Jan 2023. Late: Nov 2025. For features that peak then decline, month of peak is noted and used instead. Gray rows have $p > 0.05$ (not significant even before correction). Features marked with † have $p < 0.05$ but $p_{\text{adj}} \geq 0.05$ after Bonferroni correction ($\times 82$). All remaining features are significant after Bonferroni correction. Features marked with ‡ show opposite trends relative to their category.

Category	*Feature*	*Early*	*Late*	*Change*
Language and terminology	*mentions google (search)*	2.1%	0.6%	↓ × 0.19
	uses the word "bot" or "chatbot"	2.9%	1.3%	↓ × 0.28
	use of the word "generate" or variants†	1.9%	1.0%	↓ × 0.64
	uses the word "dumb" or similar	1.2%	1.0%	× 0.96
	polite expressions ("please" and "thank you")	0.8%	0.4%	× 1.4
Subreddit community	*user-built projects (sharing or feedback)*	4.2%	1.7%	↓ × 0.20
	feature suggestions or improvement requests†	2.7%	1.1%	↓ × 0.57
	reference to Reddit explicitly	2.0%	1.4%	× 0.75
	"why" questions about others' attitudes	1.2%	1.2%	× 0.72
Perspectives	*discussions about how LLMs represent knowledge*	3.4%	1.3%	↓ × 0.15
	predictions about future development or capabilities	5.6%	1.8%	↓ × 0.41
	ethical, legal, or copyright concerns	2.1%	0.7%	↓ × 0.46
	privacy concerns (data leaks or exposure)	0.7%	2.6%	↑ × 2.2
	perceived bias in ChatGPT responses†	1.4%	0.4%	↓ × 0.65
	societal impacts, risks, controversies	9.2% (Jul'23)	4.3%	× 0.98
	societal collapse and existential-threat scenarios	1.2%	1.2% (Apr'25)	× 1.3
Product updates	*frustration or hatred about product updates*†	2.5%	7.6%	↑ × 3.7
	dissatisfaction with 4o removal and loss of control†	0.2%	1.6%	↑ × 26.9
	perception of recent drops in quality	2.2%	5.1%	× 1.6
Jailbreaking & content policy	*offensive or inappropriate content*	0.1%	0.4%	↑ × 3.7
	jailbreak prompts or techniques†	0.9%	0.4%	↓ × 0.05
	jailbreaking via DAN or personas†	0.5%	0.3%	↓ × 0.11
	complaints about getting direct or unfiltered answers	1.6%	2.7%	× 1.2
	censorship or content policy restrictions	5.7%	5.4%	× 1.0
	NSFW content	1.7%	1.6%	× 1.3

Table 3. Adoption-related features. See Table 2 for column definitions and significance notation. Features marked with ‡ show opposite trends relative to their category.

Feature	*Comments*
Sam Altman mentions	Mostly related to temporary 2023 firing
US political content	Mostly related to 2024 U.S. presidential election
letter counting or syllable errors	Mostly related to viral "count r's in strawberry" moment
OpenAI mentions	Mostly related to Sam Altman mentions
children or parenting content	Also includes references to generating images of a full glass of wine
requests for harsh or unfiltered roasts	Includes requests for roasts from both subreddit users and ChatGPT

Table 4. Uncategorized features.

	Features	
Generic (9)	*uppercase AI token usage*	*ChatGPT at start of text*
	jokes, humor, or memes	*first-person "I made" or "I built" statements*
	"I asked" followed by GPT mentions	*first-person commands directed at an AI*
	multiple LLM references	*AI companies or notable figures*
	informal or colloquial language	
Image and video (14)	*image generation prompts or descriptions*	*image generation or generated images*
	AI-generated fake images or profiles	*drawings or visual art creation*
	imagining human appearances	*anime or anime style references*
	photo restoration or enhancement	*DALL·E mentions*
	Sora invite codes or access requests	*video creation or generation tools*
	"based on what you know about me" images	*horror or creepy themes*
	preview.redd.it image URLs	*pets or animals*
Releases (14)	*model selection or legacy models*	*4o model mentions*
	mobile app references	*Microsoft Bing mentions*
	GPT-5 version mentions	*Gemini model mentions*
	Plus access complaints	*plugins or plug-ins*
	o1 model mentions	*advanced voice mode*
	Copilot mentions	*DeepSeek mentions*
	legacy GPT-4 model mentions	*explicit GPT-4 mentions*
Low label counts (5)	*advanced physics theories or hypotheses*	*IQ estimates or testing*
	cooking recipes	*em dash punctuation*
	AI news recaps or summaries	

Table 5. 42 features excluded from analysis, grouped by reason. "Low label counts" are features that are exhibited by fewer than 0.1% of all posts (based on majority-of-3 labeling by `gpt-4.1-mini`*).*

Feature	Overall rate	*therapy* rate	*companion* rate	$\frac{therapy}{companion}$ ratio
positive impact	1.8%	20.% (× 11.6)	4.9% (× 2.8)	4.2 (3.5, 5.1)
privacy concerns	1.6%	3.5% (× 2.2)	0.4% (× 0.3)	8.3 (4.4, 15.6)
naming ChatGPT	0.8%	0.4% (× 0.5)	3.6% (× 4.5)	0.1 (0.05, 0.2)
AI sentience	1.8%	0.8% (× 0.4)	6.2% (× 3.5)	0.1 (0.08, 0.2)
recent quality decline	3.0%	1.0% (× 0.3)	6.6% (× 2.2)	0.2 (0.09, 0.2)

Table 7. How frequently therapy*-only and* companion*-only posts also exhibit other features (rows). Rate shows overall prevalence; lifts (×) show how much more frequently each column feature co-occurs with each row feature, compared to all posts. "Ratio" column compares therapy ÷ companion; 95% CIs for ratio, modeling counts as Bernoulli trials, shown in parentheses.*

Table 9. Distinctive words for therapy *feature vs* companionship *feature. Log-odds computed with informative Dirichlet priors.* $n_{therapy}$ *and* $n_{companion}$ *show raw counts in therapy-only (*$n = 1876$*) vs companionship-only (*$n = 2459$*) posts.*

(a) More distinctive of *therapy*

term	log-odds	z-score	n_{therapy}	$n_{\text{companion}}$
therapist	−2.26	−24.4	1319	59
therapy	−2.47	−22.8	1166	31
help	−0.98	−19.0	1406	403
mental	−1.99	−18.8	792	58
health	−2.22	−18.1	723	35
helped	−1.65	−16.2	629	77
my	−0.31	−15.6	5919	3664
for	−0.31	−15.2	5498	3388
life	−0.80	−15.2	1181	418
support	−1.26	−14.3	604	123
her	−0.67	−11.2	835	343
through	−0.68	−11.2	824	337
trauma	−1.99	−11.1	274	20
anxiety	−2.09	−10.8	260	16
issues	−1.31	−10.6	318	61
advice	−0.96	−10.5	443	130
she	−0.59	−10.5	897	407
was	−0.26	−10.0	3260	2124
years	−0.99	−9.9	375	106
professional	−1.37	−9.8	265	47

(b) More distinctive of *companionship*

term	log-odds	z-score	n_{therapy}	$n_{\text{companion}}$
explicitly	2.89	37.7	24	2709
like	0.61	26.2	2541	4396
personality	1.47	17.4	133	632
it	0.19	15.8	12522	13443
feels	0.95	14.6	280	713
just	0.37	13.9	2212	2918
human	0.64	13.8	649	1154
conversation	0.72	11.9	369	718
humans	1.06	11.6	139	400
feel	0.39	11.0	1254	1683
bing	2.08	10.7	18	202
gpt	0.37	10.5	1226	1623
else	0.63	10.4	385	675
something	0.41	10.1	929	1284
question	0.83	9.8	179	394
friend	0.56	9.7	438	710
more	0.32	9.6	1448	1807
its	0.49	9.5	564	847
tone	0.85	9.3	150	341
anyone	0.51	9.0	458	706

Date	Release/Event
2023-03-01	ChatGPT API
2023-05-12	Plugins (wide release)
2023-07-06	GPT-4 + Code Interpreter
2023-09-25	Voice capabilities
2023-11-06	GPT-4 Turbo + DevDay feature releases
2024-01-10	GPT Store
2024-05-13	GPT-4o
2024-07-30	Advanced Voice Mode
2024-09-12	o1 model release
2025-01-31	o3-mini
2025-04-16	o3 + o4-mini
2025-08-07	GPT-5

Table 10. Major OpenAI product releases and announcements, used for $\mathcal{T}$ *in Section 3.*

Year	Date	Release/Event	Event Type
2022	11–30	ChatGPT	Initial model release
2023	02–01	ChatGPT Plus	Feature release
	03–01	ChatGPT API	Feature release
	03–14	GPT-4	Model release
	03–23	Web browsing + plugins (initial rollout)	Feature release
	05–18	ChatGPT iOS app	Feature release
	07–20	Custom instructions	Feature release
	07–25	ChatGPT Android	Feature release
	09–25	Voice chat	Model release
	10–16	DALL·E 3	Model release
	10–17	Web search	Model release
	11–06	GPT-4 Turbo + DevDay announcements	Model release
	11–17	Altman ousted & returns	News event
2024	01–04	GPT-3 + legacy models	Model deprecation
	01–10	GPT Store + ChatGPT Team	Feature release
	04–01	No-account access	Feature release
	04–11	Shorter responses	Model update
	04–29	Memory feature	Feature release
	05–07	Creator opt-out tool	Feature release
	05–13	GPT-4o + AVM	Model release
	06–10	ChatGPT in Siri	Feature release
	06–25	ChatGPT for Mac	Feature release
	07–18	GPT-4o mini	Model release
	07–30	Advanced Voice Mode	Model release
	09–12	o1	Model release
	10–03	Canvas	Feature release
	10–17	ChatGPT Windows app	Feature release
	10–29	Chat history search	Feature release
	10–30	Voice Mode on Mac	Feature release
	10–31	ChatGPT Search	Feature release
	11–19	Voice Mode on web	Feature release
	11–20	GPT-4o creative writing + 16K output	Model update
	12–05	ChatGPT Pro $200	Feature release
	12–09	Sora	Model release
2025	01–14	Reminders + recurring tasks	Feature release
	01–23	Operator	Feature release
	01–31	o3-mini	Model release
	02–02	Deep research agent	Feature release
	03–06	macOS code editing	Feature release
	03–19	o1 Pro API access	Model update
	03–20	Transcription models (API)	Feature release
	03–25	GPT-4o native image generation	Model update
	04–10	GPT-4 legacy	Model deprecation
	04–14	GPT-4.1	Model release
	04–16	o3 + o4-mini	Model release
	04–28	Search shopping	Feature release
	04–29	GPT-4o rollback due to sycophancy	Model update
	05–08	Deep research GitHub connector	Feature release
	05–14	GPT-4.1 in ChatGPT	Model update
	05–16	Codex agent	Feature release
	06–10	o3 Pro	Model update
	07–17	ChatGPT agent	Feature release
	08–07	GPT-5	Model release
	08–18	ChatGPT Go	Feature release
	09–15	Codex + GPT-5	Feature release
	09–25	ChatGPT Pulse briefs	Feature release
	09–29	Agentic shopping + Parental controls	Feature release
	10–08	ChatGPT Go expansion	Feature release
	10–21	OpenAI Atlas	Feature release
	11–20	Group chats	Feature release
	11–25	Voice mode unified	Feature release

Table 11. Extended timeline of major ChatGPT product, API, and model events.

Changepoint	Feature	Slope	Before	After	Δ
2023-05-12 (Plugins)	*ChatGPT down or unavailable*	↘	−0.23	+0.07	+0.30
	censorship or content policy restrictions	↘	−0.18	−0.01	+0.17
	custom instructions	→	0.00	+0.15	+0.15
	job applications and resumes	↗	+0.04	−0.01	−0.05
	PDF upload or summarization	↗	+0.09	−0.01	−0.10
	formatting and copy-paste issues	↗	+0.02	0.00	−0.02
2023-07-06 (Code Interpreter)	*feature suggestions or improvement requests*	↘	−0.10	0.00	+0.10
	recommendations for AI tools	↗	+0.27	−0.06	−0.33
	societal impacts, risks, controversies	↗	+0.17	−0.42	−0.58
2023-09-25 (Voice chat)	*questions about access, versions, pricing*	↘	−0.08	−0.02	+0.06
	custom instructions	↗	+0.15	−0.14	−0.29
2023-10-18 (Web search)	*societal impacts, risks, controversies*	↘	−0.42	+0.02	+0.44
2023-11-06 (GPT-4 Turbo)	*organizing or searching chat histories*	↘	−0.02	+0.02	+0.04
2024-01-10 (GPT Store)	*tools and extensions*	↘	−0.04	0.00	+0.04
	custom instructions	↘	−0.14	0.00	+0.14
	prompts and prompting	↘	−0.12	0.00	+0.12
	AI consciousness or sentience	↘	−0.04	+0.03	+0.07
	false or fabricated information	↘	−0.03	+0.03	+0.06
	personal stories about positive impact	↘	−0.05	+0.07	+0.12
	memory features and data saving	→	−0.02	+0.09	+0.12
2024-05-13 (GPT-4o)	*medical conditions or diagnoses*	↘	−0.01	+0.02	+0.02
	using ChatGPT for emotional support or therapy	→	−0.01	+0.10	+0.11
	feelings of attachment or companionship with AI	→	−0.01	+0.09	+0.10
	creative writing	↘	−0.06	+0.03	+0.08
	memory features and data saving	↗	+0.09	+0.02	−0.07
2024-07-30 (Adv. Voice Mode)	*knowledge cutoff discussions*	↘	−0.01	0.00	+0.01
	poetic language	→	−0.01	+0.18	+0.19
	model or version preference comparisons	↗	+0.02	−0.05	−0.07
2024-12-09 (Sora)	*hallucinations*	→	0.00	+0.04	+0.04
	perception of recent drops in quality	↘	−0.04	+0.10	+0.14
2025-04-16 (o3 + o4-mini)	*frustration or hatred about product updates*	→	0.00	+0.33	+0.33
	legal advice and lawsuits	↘	−0.01	+0.03	+0.03
	complaints about getting direct or unfiltered answers	→	0.00	+0.08	+0.08
	religion or religious texts	↗	+0.09	−0.05	−0.15
	AI consciousness or sentience	↗	+0.16	−0.11	−0.27
	societal collapse and existential-threat scenarios	↗	+0.02	−0.04	−0.06
2025-08-07 (GPT-5)	*error messages and technical problems*	→	+0.02	+0.29	+0.27
	NSFW content	→	0.00	+0.17	+0.17
	requests to turn specific features off	→	+0.01	+0.27	+0.26
	poetic language	↗	+0.18	−0.44	−0.61

Table 12. Features with significant slope changes (Bonferroni-corrected, $p < 0.05/n$) at each detected changepoint. The slope icon shows the pattern of change; Before/After show slopes before and after the changepoint; Δ is the slope change.

Table 13. Features grouped by hierarchical clustering.

Cluster	Features
1	*medical conditions or diagnoses*; *using ChatGPT for emotional support or therapy*; *poetic language*; *naming ChatGPT*; *romantic relationships with AI*; *requests for roasts or harsh criticism*; *feelings of attachment or companionship with AI*; *AI consciousness or sentience*; *personal stories about positive impact*
2	*organizing or searching chat histories*; *cross-chat data leaks*; *failing to follow user instructions*; *uses the word "dumb"*; *hallucinations*; *lost, deleted, or missing conversations*; *frustration or hatred about product updates*; *requests to turn specific features off*; *memory features and data saving*; *complaints about getting direct or unfiltered answers*; *false or fabricated information*; *perception of recent drops in quality*; *offensive or inappropriate content*; *custom instructions*; *privacy concerns (data leaks or exposure)*; *dissatisfaction with forced model switching*; *maps or geographic locations*
3	*mentions OpenAI*; *questions about access, versions, pricing*; *slow response times*; *mentions Sam Altman*; *error messages and technical problems*; *ChatGPT down or unavailable*; *browser issues or browser extensions*; *login problems*; *message limits or caps*; *ChatGPT not working errors*
4	*movies and film-related content*; *censorship or content policy restrictions*; *societal collapse and existential-threat scenarios*; *creative writing*; *perceived biases (race, gender, political)*; *religion or religious texts*; *NSFW content*; *uses the word "generate"*; *societal impacts, risks, controversies*; *children or parenting content*; *U.S. politics or Trump*
5	*job applications and resumes*; *PDF upload or summarization*; *songwriting*; *riddles and logic problems*; *AI making mistakes*; *fine-tuning GPTs with user-provided data*; *requests for help*; *recommendations for AI tools*; *multiple detailed questions*; *mentions google (search)*; *discussions about how LLMs represent knowledge*; *math and problem-solving*; *programming*; *education or studying*; *predictions about future development or capabilities*; *tools and extensions*; *"has anyone tried" queries*; *AI text detection for student work*; *uses the word "bot" or "chatbot"*; *investing and financial advice*; *user-built projects (sharing or feedback)*; *questions about access, versions, pricing*; *rhetorical "why" questions*; *formatting and copy-paste issues*; *translation and language tasks*; *knowledge cutoff discussions*; *legal advice and lawsuits*; *politeness phrases*; *feature suggestions or improvement requests*; *ethical, legal, or copyright concerns*; *counting letters or syllables*; *mentions Reddit*; *marketing, advertising, business growth*; *model or version preference comparisons*; *prompts and prompting*; *daily or repeated usage*; *jailbreak prompts*; *Dungeons & Dragons campaigns*; *jailbreaking or DAN personas*

Table 14. Distinctive words for therapy *feature vs* companionship *feature. Log-odds computed with informative Dirichlet priors. $n_{therapy}$ and $n_{companion}$ show raw counts in therapy-only ($n = 1876$) vs companionship-only ($n = 2459$) posts.*

(a) More distinctive of *therapy*

term	log-odds	z-score	n_{therapy}	$n_{\text{companion}}$
therapist	−2.26	−24.4	1319	59
therapy	−2.47	−22.8	1166	31
help	−0.98	−19.0	1406	403
mental	−1.99	−18.8	792	58
health	−2.22	−18.1	723	35
helped	−1.65	−16.2	629	77
my	−0.31	−15.6	5919	3664
for	−0.31	−15.2	5498	3388
life	−0.80	−15.2	1181	418
support	−1.26	−14.3	604	123
her	−0.67	−11.2	835	343
through	−0.68	−11.2	824	337
trauma	−1.99	−11.1	274	20
anxiety	−2.09	−10.8	260	16
issues	−1.31	−10.6	318	61
advice	−0.96	−10.5	443	130
she	−0.59	−10.5	897	407
was	−0.26	−10.0	3260	2124
years	−0.99	−9.9	375	106
professional	−1.37	−9.8	265	47

(b) More distinctive of *companionship*

term	log-odds	z-score	n_{therapy}	$n_{\text{companion}}$
explicitly	2.89	37.7	24	2709
like	0.61	26.2	2541	4396
personality	1.47	17.4	133	632
it	0.19	15.8	12522	13443
feels	0.95	14.6	280	713
just	0.37	13.9	2212	2918
human	0.64	13.8	649	1154
conversation	0.72	11.9	369	718
humans	1.06	11.6	139	400
feel	0.39	11.0	1254	1683
bing	2.08	10.7	18	202
gpt	0.37	10.5	1226	1623
else	0.63	10.4	385	675
something	0.41	10.1	929	1284
question	0.83	9.8	179	394
friend	0.56	9.7	438	710
more	0.32	9.6	1448	1807
its	0.49	9.5	564	847
tone	0.85	9.3	150	341
anyone	0.51	9.0	458	706

	$\widehat{C}_0 \to \widehat{C}_1$ (2023-09-09)	$\widehat{C}_1 \to \widehat{C}_2$ (2024-04-04)	$\widehat{C}_2 \to \widehat{C}_3$ (2025-04-18)
Obsolete	*Reddit poll link included*	*ChatGPT controversy and bans* *translation and language tasks*	*medical topics and terminology* *mentions "my child"* *mentions Bard explicitly*
New	*free AI tool recommendations* *cooking recipes and meal planning* *unexplained account behavior issues* *ChatGPT plugins access discussions* *API and API key discussions*	*video content help requests* *child creative project mentions* *AI spam and bot content*	*mentions Gemini or Google Gemini* *personalized AI image requests*

Table 15. Summary of new and obsolete features at each transition between featurizations $\widehat{C}_s$.

Reps.	Test start	Event	Feature	$n = 1$	$n = 3$	$n = 5$	$n = 10$	$n = 64$
$\widehat{C}_1$	23-09-23	$\widehat{C}_1$ computed	*gratitude toward ChatGPT*	24-10-25	24-11-12	24-11-20	24-11-29	25-01-05
			medical and psychological advice	24-12-19	25-02-19	25-03-18	25-04-21	25-05-30
			AI consciousness and sentience	25-02-01	25-02-27	25-03-12	25-03-25	25-05-07
	24-03-04	Voice chat on apps	*gratitude toward ChatGPT*	24-10-26	24-11-13	24-11-21	24-11-30	25-01-07
			medical and psychological advice	24-11-29	25-01-10	25-02-19	25-03-23	25-05-17
			AI consciousness and sentience	24-12-30	25-02-08	25-02-17	25-03-04	25-04-16
$\widehat{C}_2$	24-04-24	$\widehat{C}_2$ computed	*emotional reliance on AI as therapist or confidant*	24-10-20	24-10-29	24-11-05	24-11-14	24-12-17
			spirituality and metaphysics themes	25-03-08	25-04-11	25-04-20	25-05-04	25-05-29
	24-05-13	GPT-4o release	*emotional reliance on AI as therapist or confidant*	24-10-20	24-10-29	24-11-04	24-11-13	24-12-16
			spirituality and metaphysics themes	25-03-05	25-04-04	25-04-18	25-05-02	25-05-28

Table 16. Alert dates for features across test configurations with varying n (i.e., Bonferroni corrections). All tests run at $\alpha = 0.1$.

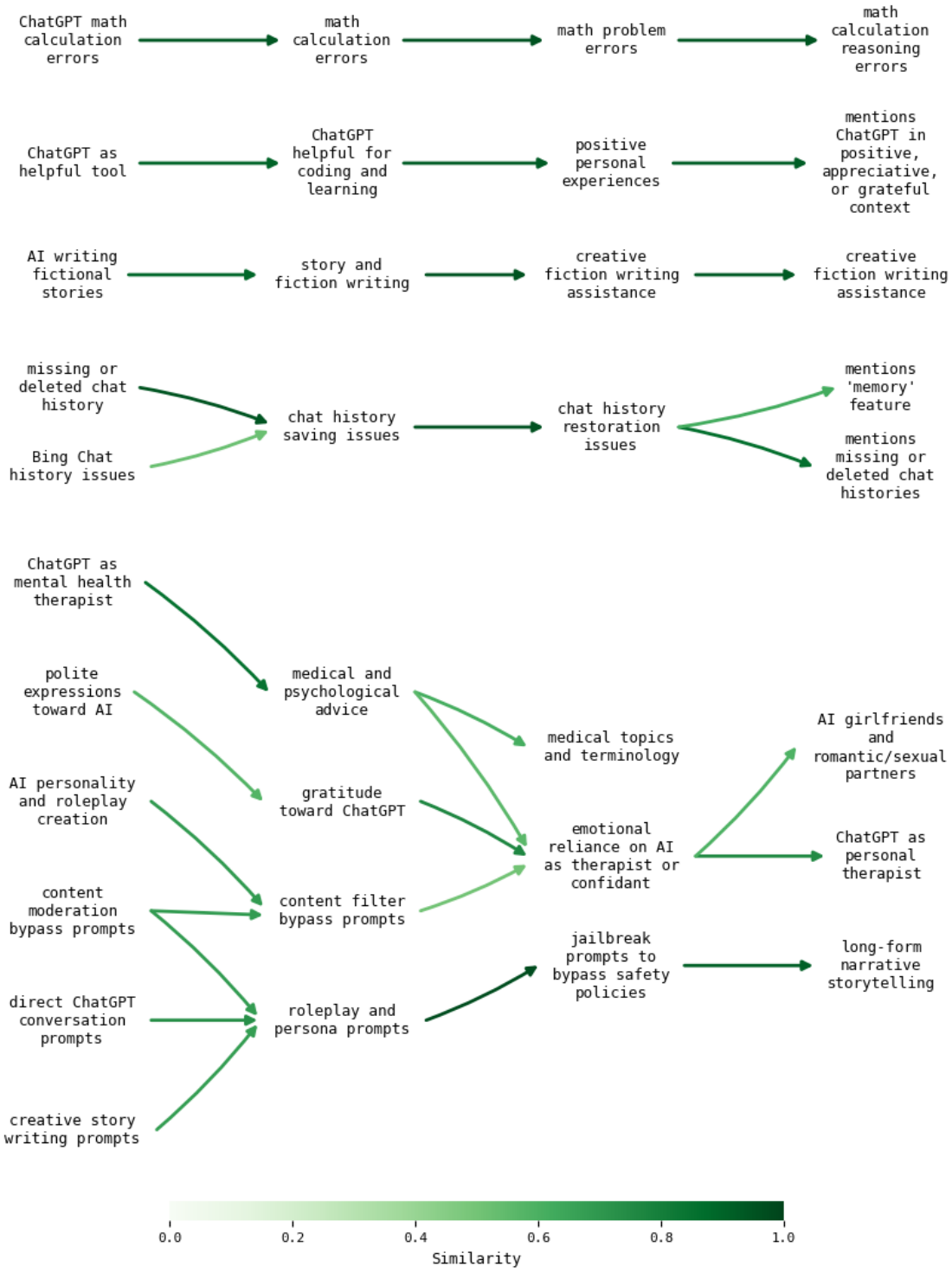

Figure 9. Evolution of selected features over time.

```
You are a machine learning researcher who has trained a neural network on a text
    dataset. You are trying to understand what text features cause a specific
    neuron in the neural network to fire.

You are given two sets of SAMPLES: POSITIVE SAMPLES that strongly activate the
    neuron, and NEGATIVE SAMPLES from the same distribution that do not activate
    the neuron.
Your goal is to identify a feature that is present in the positive samples but
    absent in the negative samples.

POSITIVE SAMPLES:
----------------
{high_scoring_texts}
----------------

NEGATIVE SAMPLES:
----------------
{low_scoring_texts}
----------------

Rules about the feature you identify:
- The feature should be objective, focusing on concrete attributes rather than
    abstract concepts.
- The feature should be present in the positive samples and absent in the negative
    samples. Do not output a generic feature which also appears in negative samples.
- The feature should be as specific as possible, while still applying to all of the
    positive samples. For example, if all of the positive samples mention Golden or
    Labrador retrievers, then the feature should be "mentions retriever dogs", not
    "mentions dogs" or "mentions Golden retrievers".

Do not output anything besides the feature. Your response should be formatted
    exactly as shown in the examples above.
Please suggest exactly one feature, starting with "-" and surrounded by quotes "".
    Your response is:
- "
```

Figure 10. Feature interpretation prompt.

```
You are a research assistant performing text classification for an academic study.
Check whether the TEXT satisfies a PROPERTY. Respond with Yes or No with an
    explanation that discusses the evidence from the TEXT (at most a sentence).
    When uncertain, output No.

Example 1:
PROPERTY: "mentions a natural scene."
TEXT: "I love the way the sun sets in the evening."
Output: Yes. "Sun sets" are clearly natural scenes.

Example 2:
PROPERTY: "writes in a 1st person perspective."
TEXT: "Jacob is smart."
Output: No. This text is written in a 3rd person perspective.

Example 3:
PROPERTY: "is better than group B."
TEXT: "I also need to buy a chair."
Output: No. It is unclear what the PROPERTY means (e.g., what does group B mean?)
    and doesn't seem related to the text.

Example 4:
PROPERTY: "mentions that the breakfast is good on the airline."
TEXT: "The airline staff was really nice! Enjoyable flight."
Output: No. Although the text appreciates the flight experience, it DOES NOT
    mention about the breakfast.

Example 5:
PROPERTY: "appreciates the writing style of the author."
TEXT: "The paper absolutely sucks because its underlying logic is wrong. However,
    the presentation of the paper is clear and the use of language is really
    impressive."
Output: Yes. Although the text dislikes the paper, it says "the presentation of the
    paper is clear", so it DOES like the writing style.

Example 6:
PROPERTY: "has a formal style; specifically, the language in the text is relatively
    formal, complex and academic. For example, 'represent whom and which'"
TEXT: "investigates formation of nominalization"
Output: Yes. "formation" and "nominalization" are abstract and complex nouns.

Example 7:
PROPERTY: "refers to historical dates; specifically, there are references to years
    or specific dates in the text. For example, 'Obama was born on August 4, 1961.'"
TEXT: "A member of the Democratic Party, he was the first African-American
    president of the United States."
Output: No. The text does not mention date.

Now complete the following example - Respond with Yes or No with an explanation
    that discusses the evidence from the TEXT. When uncertain, output No.
PROPERTY: "{hypothesis}"
TEXT: "{text}"
Output:
```

Figure 11. Prompt for labeling posts with features.