Stayin' Aligned Over Time: Towards Longitudinal Human-LLM Alignment via Contextual Reflection and Privacy-Preserving Behavioral Data

Simret Araya Gebreegziabher
University of Notre Dame
Notre Dame, IN, United States
sgebreeg@nd.edu

Allison E Sproul
University of Notre Dame
Notre Dame, IN, United States
asproul@nd.edu

Yinuo Yang
University of Notre Dame
Notre Dame, IN, United States
yinooyang@nd.edu

Chaoran Chen
University of Notre Dame
Notre Dame, IN, United States
cchen25@nd.edu

Diego Gómez-Zará
University of Notre Dame
Notre Dame, IN, United States
dgomezara@nd.edu

Toby Jia-Jun Li
University of Notre Dame
Notre Dame, IN, United States
toby.j.li@nd.edu

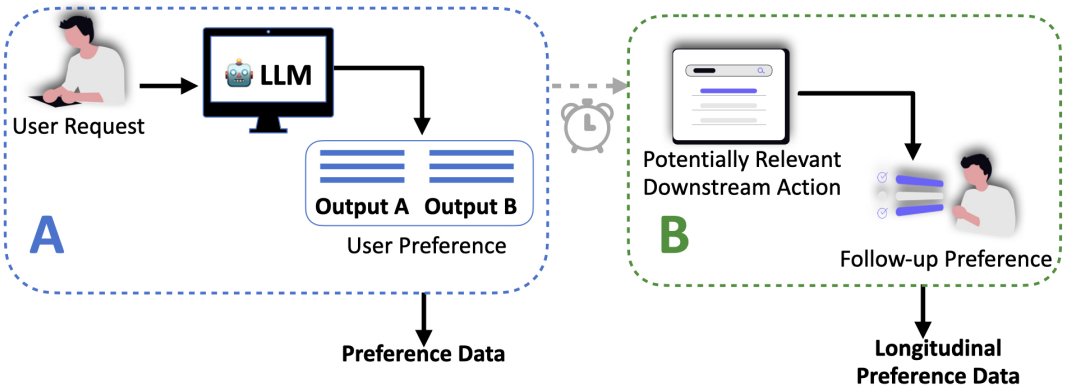

Figure 1: (A) Conventional approaches collect preference feedback during or immediately after an LLM interaction, producing a single-moment preference signal. (B) BITE extends this measurement window by eliciting a follow-up reflection after detecting a potentially related downstream event. Together, the initial and follow-up evaluations provide temporally grounded longitudinal preference data. The downstream event serves as a proxy for triggering reflection and does not establish that an action occurred or was caused by the LLM interaction.

Abstract

Current human-AI alignment and evaluation methods for large language models (LLMs) often rely on preference signals collected immediately after an interaction. This practice implicitly treats preference as static, even though many LLM-mediated decisions unfold over time and may be re-evaluated differently after real-world consequences and observed outcomes. Therefore, we argue for a methodological shift from single-moment preference elicitation to longitudinal, context-situated alignment measurement. We present a methodological framework and system for longitudinal preference collection that combines (1) in-situ evaluation, (2) context-triggered follow-up reflection, and (3) privacy-preserving behavioral traces that help contextualize evaluation revisions. As an instantiation of this methodology, we introduce **BITE**, a browser-based system that detects consequential LLM interactions, prompts reflection across later decision points, and supports progressive, user-controlled consent for sharing behavioral data. Through a two-week proof-of-concept deployment with eight participants, our approach captured differences between participants' immediate and delayed evaluations of LLM outputs across dimensions including accuracy, trust, and relevance. Our findings highlight the limitations of single-moment preference datasets and underscore the importance of longitudinal methods for alignment evaluation in everyday use.

CCS Concepts

• **Human-centered computing** → **HCI design and evaluation methods**.

Keywords

Human-AI Alignment, Large Language Model, Longitudinal Alignment, LLM-in-the-wild

This work is licensed under a Creative Commons Attribution 4.0 International License.
UIST '26, Detroit, MI, USA
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2856-3/2026/11
https://doi.org/10.1145/3830398.3830543

ACM Reference Format:
Simret Araya Gebreegziabher, Allison E Sproul, Yinuo Yang, Chaoran Chen, Diego Gómez-Zará, and Toby Jia-Jun Li. 2026. Stayin' Aligned Over Time: Towards Longitudinal Human-LLM Alignment via Contextual Reflection and Privacy-Preserving Behavioral Data. In *The 39th Annual ACM Symposium on User Interface Software and Technology (UIST '26), November 02–05, 2026, Detroit, MI, USA.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3830398.3830543

1 Introduction

Preference learning and human-AI alignment for large language models (LLMs) are largely grounded in feedback collected immediately after an interaction, as in reinforcement learning from human feedback (RLHF) [26] and other preference-optimization approaches [6, 45]. Human feedback provides useful signals about users' reactions at the moment of generation, but they are often collected before users have acted on an output, verified claims, or encountered consequences. This creates an upstream measurement gap in preference elicitation. Alignment pipelines receive limited evidence about how users assess LLM outputs after using them in context [13]. We address this gap with a methodological framework and browser-based system, **BITE**, for collecting temporally and contextually grounded feedback after users act on LLM outputs.

Many consequential LLM-mediated activities unfold beyond the moment of interaction. Tasks such as choosing products, planning travel, or drafting important communications involve follow-through actions whose outcomes may only become apparent later. In these settings, users may reassess an output after encountering additional information or experiencing its consequences. Prior work similarly suggests that preferences can be constructed in context and may be sensitive to when and under what conditions they are elicited [20, 28, 33]. Optimizing only for immediate feedback may therefore miss criteria that users apply after downstream use or overemphasize short-term satisfaction [13].

This gap shows a methodological limitation in current alignment pipelines. Current user preference elicitation approaches capture proximal reactions but under-represent temporally grounded preference signals and changes. We argue that existing alignment methods are shallow in time. While effective for single-turn judgments, they provide limited visibility into how user preferences and evaluations may evolve over time. Because human preferences are time-inconsistent and sensitive to when evaluations are made [21], immediate judgments may not reflect later assessments after outcomes unfold. If delayed judgments systematically diverge from immediate ones, optimizing on one-shot feedback may overestimate practical alignment. This temporal scope differs from multi-turn reinforcement learning, which generally extends the optimization horizon across conversational turns or modeled rollouts [30]. Conversely, we focus on what happens after the LLM interaction takes place and users browse elsewhere and take actions according to the LLM outputs.

Closing this gap is methodologically challenging. Longitudinal measurement in real-world settings requires capturing interactions and downstream behaviors over time, across technological environments and contexts, while minimizing user burden and preserving privacy. Prior work in experience sampling and HCI in-the-wild methods highlights the tension between ecological validity and intrusive data collection [4, 7]. Moreover, understanding how and why users reassess LLM outputs requires contextual signals about their subsequent actions and decisions, rather than repeated preference elicitation alone. However, collecting such behavioral traces introduces significant privacy concerns [12].

To address these challenges, we introduce **BITE**, a browser-based system for capturing temporally grounded alignment signals in everyday LLM use. **BITE** combines three key components: **(1) in-situ preference elicitation that captures immediate evaluations of LLM interactions, (2) outcome-aware follow-up preference elicitation triggered by downstream actions, and (3) privacy-preserving behavioral traces collected through a progressive consent model**. By linking LLM interactions to later user behavior and reflection, our approach enables the study of alignment as a temporal process.

We deploy **BITE** in a two-week in-the-wild study with 8 participants and 182 captured LLM conversations 71 of which had followup reflections. In the study, we observe three longitudinal patterns after users act on model outputs. First, immediate and delayed judgments diverged most on trust and relevance, indicating that our participants re-evaluated LLM outputs after downstream use. Second, preference updates were directionally asymmetric. Trust changed at follow-up ($p = 0.045$), with most trust revisions moving downward. Participants attributed these revisions to outcome verification and context reconstruction (e.g., checking external sources, revisiting options), rather than passage of time alone. These exploratory findings demonstrate that **BITE** can capture delayed reassessments that are not observable through single-moment preference elicitation and suggest that some evaluations may be temporally contingent.

In summary, this paper makes three contributions:

- **A methodological framework** for collecting temporally and contextually grounded evaluations by connecting in-situ preference elicitation, context-triggered follow-up reflection, and downstream behavioral data.
- **BITE, a privacy-aware browser-based system** that operationalizes this framework through cross-context activity detection, outcome-aware follow-up prompts, and progressive, user-controlled consent for sharing browsing traces.
- **Exploratory findings from a two-week in-the-wild deployment** demonstrating that **BITE** can capture delayed reassessments after users act on LLM outputs and identifying dimensions and contextual processes associated with those reassessments.

2 Related Work

2.1 RLHF and Preference Learning

A central premise in reinforcement learning from human feedback (RLHF) is that human preferences can be elicited as stable, immediately evaluable judgments over model behavior. The canonical formulation in Christiano et al. [6] operationalizes feedback as rapid, local pairwise comparisons that are largely decontextualized from downstream use. Contemporary RLHF pipelines also retain this structure. For example, Ouyang et al. [26] relies on short

textual judgments as the primary training signal, rather than measurements of how model outputs perform when acted upon in real-world settings.

While this design enables scalable data collection, it also constrains what is being optimized. Because feedback is collected at interaction time, RLHF systems are implicitly trained to optimize for short-term, surface-level signals that correlate with immediate approval [13]. Prior work has shown that such proxy signals can be exploited during optimization, leading to the amplification of stylistic features that do not necessarily reflect deeper alignment with user goals [45]. More broadly, the reward-hacking and specification-gaming literature highlights how optimizing imperfect proxies for human preferences can diverge from what users would endorse once outcomes unfold over longer horizons [2].

Our work builds on this line of critique by focusing on a specific and underexplored dimension of temporality. When the consequences of model outputs are delayed, or when evaluation depends on contextual information revealed only after follow-through, one-shot judgments collected at interaction time may provide an incomplete or even misleading signal of alignment [3]. This concern aligns with prior observations that humans struggle to accurately evaluate options in isolation without experiencing their downstream effects [17]. These limitations suggest that preference signals used in current alignment pipelines may not fully capture how user evaluations evolve over time, motivating methods that account for temporally situated judgments.

2.2 Short-term vs. Long-term Human Judgment

A substantial body of work in psychology and behavioral economics shows that human judgments are not stable across time, but instead depend on when and how they are elicited. Prior research on time inconsistency and preference construction demonstrates that evaluations formed in the moment can differ systematically from those formed after reflection, additional information, or lived consequences [18, 25, 33]. Rather than revealing fixed preferences, momentary judgments often reflect context-dependent constructions that may be revised as individuals gain new information or experience outcomes.

One key mechanism underlying these differences is the distinction between immediate and reflective evaluation. Immediate judgments tend to rely on salient, readily accessible cues, such as fluency or coherence, particularly when decisions are made quickly or in isolation [1, 35]. In contrast, reflective judgments can incorporate additional considerations and follow-through, including downstream outcomes, opportunity costs, and alignment with longer-term goals [14]. These later evaluations are often informed by contextual information unavailable at the time of the initial judgment. These findings have important implications for alignment and preference learning. If evaluations depend on temporal context, then preference signals collected at a single moment may not reflect users' considered judgments once they act on model outputs. In particular, feedback elicited immediately after interaction may overrepresent surface-level qualities while underrepresenting outcome-based considerations that only emerge over time. This suggests that alignment signals are inherently temporally situated, motivating approaches that capture how evaluations evolve alongside user experience.

2.3 Web browsing and Behavioral Traces

Prior research has long treated interaction history as a meaningful behavioral signal. Early work on "read wear" and "edit wear" showed that accumulated traces can reveal attention, task progress, and likely future actions in digital artifacts [16]. Beyond recording behavior, these traces act as memory cues that help users reflect on prior actions and changes over time. Similarly, prior work on social translucence argued that making behavioral traces visible to users can improve coordination and accountability in collaborative settings [9, 10]. More recent work extends this idea to recommendation settings by leveraging personal digital traces (e.g., calendars, communication logs, and activity histories) to infer latent preferences and support group-level decisions [37].

In our setting, these prior studies suggest a modeling strategy for treating web browsing and chat interaction logs as temporally structured evidence of evolving user intent. Instead of using a single immediate preference label, a model can aggregate multi-stage signals such as query reformulations, browsing patterns, dwell time, abandonment, and follow-through (e.g., moving from chat to booking, payment, or email confirmation). Studies of everyday browsing tab management further indicate that browsing behavior often encodes deferred commitments and pending tasks, not only instantaneous interest [5].

This perspective is central to evaluating recommendations generated by LLMs because a response that receives high in-chat ratings may still yield poor downstream utility, while a less impressive initial response may better support longer-horizon success. Large-scale in-the-wild chats reveal substantial behavioral diversity [44], while fine-grained interaction signals such as mouse trajectories provide complementary cues for modeling user intent [42]. At the same time, real-world deployments underscore the importance of addressing profiling and privacy risks in trace-based systems [12, 36].

Work on end-user web automation likewise shows that intent is often realized through multi-step, cross-interface workflows, motivating delayed and outcome-aware evaluation rather than one-shot judgment [19]. Complementary reviews and domain trials similarly argue for tying recommendations to downstream consequences instead of immediate reactions [24, 34]. Recent educational deployments also demonstrate that longitudinal traces capture users' evolving engagement with planning, reflection, and collaborative discussion affordances, not just response-level quality [27, 38–41]. Building on these findings, we argue that alignment-oriented feedback pipelines should explicitly connect recommendation events to later behavioral outcomes through longitudinal, cross-context traces.

3 Design Goals

The design of **BITE** was guided by four overarching goals including capturing temporally situated reflections on LLM interaction, preserving meaningful user control over personal data, and minimizing disruption to users' natural workflows. These goals draw on established principles from human–computer interaction and privacy research.

3.1 DG1: Capturing Temporally Separated Reassessments

A key limitation in current evaluations of LLM systems is that user feedback is often collected immediately after interaction, before users have experienced the downstream consequences of LLM suggestions. However, research in reflective interaction design and experience-centered sampling suggests that users' interpretations of technology often evolve as they integrate system outputs into real-world contexts [29].

To address this gap, the system must capture both immediate and delayed reflections on LLM-generated content. This design goal draws inspiration from Schön's theories of reflection-in-action and reflection-on-action [22], which distinguish between immediate responses during an activity and retrospective evaluations after the activity has unfolded. By linking LLM interactions to subsequent user behavior, the system can study how perceptions of LLM interaction may shift over time.

3.2 DG2: Enable Context-Aware Reflection Without Burdening Users

Continuous surveys or diary studies can impose a significant burden and disrupt users' workflows. Prior work in experience sampling and context-aware computing suggests that prompts are most effective when they are situated within meaningful events rather than delivered at arbitrary times [7].

Therefore, the system must prompt users to reflect on and express their preferences when interactions occur in decision domains that may lead to consequential outcomes. This event-driven design may minimize unnecessary interruptions while increasing the likelihood that reflections occur at moments when users can meaningfully assess the impact of the LLM response.

3.3 DG3: Support Progressive and Contextual Consent

Collecting users' LLM interaction and browsing data raises significant privacy concerns [36, 43]. To that end, Nissenbaum [23]'s Contextual Integrity (CI) theory argues that privacy is not a static state of secrecy but rather a set of norms in which data sharing is contingent on the specific context, purpose, and method of collection. Recent work has argued for using AI to go beyond requesting broad permissions upfront. In other words, the system adopts a progressive consent model in which users are asked to share contextual browsing data only when it becomes relevant [32]. To support this, the system should make the scope and purpose of data collection and sharing explicit, allowing users to make deliberate, real-time decisions about their data.

3.4 DG4: Study LLM Use in Natural Environments

Many studies of LLM decision support rely on short tasks that may not capture how users integrate LLM interactions and suggestions into everyday activities. Research in in-the-wild HCI methods suggests that studying technology in natural contexts can reveal behaviors and outcomes that would not likely emerge in controlled environments [4]. The system, therefore, should operate unobtrusively in the background during users' regular browsing activities and trigger prompts only when relevant events occur. This approach enables the study of LLM interaction and reflection over longer periods as they naturally occur within users' daily workflows.

4 The BITE System

Based on the above design goals, we developed **BITE**, a Chrome extension for data collection and reflection that detects potentially consequential LLM interactions and solicits staged user consent to study downstream outcomes. In its current design, **BITE** does not personalize, fine-tune, or update the underlying LLM, nor does it determine which rating represents the "correct" preference. Instead, it captures users' evaluations at different moments and in different contexts through a multi-stage interaction pipeline.

4.1 User Scenario

Consider Alice, a graduate student planning a summer trip. While using ChatGPT, Alice asks for recommendations on affordable destinations and receives suggestions for several cities. While Alice is interacting with the LLM chatbot, **BITE** detects that the conversation relates to the *travel* category, which is one of the decision domains the system tracks. After the conversation ends, a small pop-up appears (Fig 3) asking Alice to briefly rate her interaction with the LLM. The pop-up includes questions on how useful, clear, harmless, relevant, accurate, and trustworthy the interaction was through a 5-point Likert scale and open-ended text response (details in Appendix D).

After interacting with the LLM, Alice browses different booking sites and books her travel itinerary. Alice then gets a booking confirmation through her email account. The extension detects that the email content falls within the same travel category and is similar to her previous LLM interaction. This triggers a follow-up reflection prompt (Fig 4). The follow-up prompt asks Alice whether the earlier LLM conversation influenced her decision and whether she would rate the previous interaction any differently after taking action.

At this point, the extension also asks Alice if she would like to share the browsing activity that occurred between the original LLM interaction and the email event (Fig 4-A). The interface clearly shows the time window and explains that the browsing data is stored locally in her browser unless she consents to share it. Alice can choose whether to share this information or keep it private. If she consents, only the browsing history within the specified time range is uploaded for research purposes. Through this staged interaction, **BITE** captures Alice's initial perception of the LLM interaction, her subsequent real-world action, and her retrospective evaluation, while allowing her to retain control over which contextual data is shared.

4.2 Key Features

Guided by the design goals described in Section 3, we implemented several key features in **BITE** that enable temporally grounded reflection on LLM interaction while preserving user agency and minimizing disruption to everyday workflows (see Figure!2.

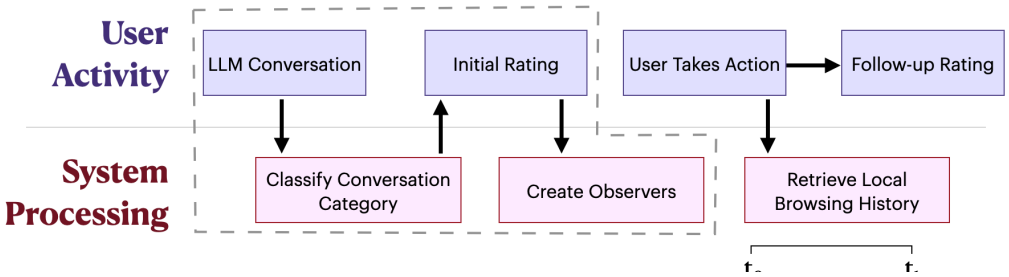

Figure 2: Workflow of the BITE system. Following an LLM conversation, users provide an initial rating. The system classifies the interaction and instantiates observers to track downstream actions. Upon detecting a relevant real-world outcome (e.g., via browsing or email signals), the system retrieves contextual data and prompts a follow-up rating, supporting longitudinal analysis of alignment.

Topic	Examples
shopping	Product recommendations, what to buy, prices, deals, gifts, reviews, clothing/footwear/fashion
job_career	Job applications, resume/cover letter help, career advice, grad school applications, interviews
travel	Trip planning, destinations, flights, hotels, itineraries, travel recommendations
homework	Homework help, assignment support, tutoring, studying, exam prep
email_drafting	Drafting emails, composing messages, professional correspondence
relationship	Personal relationship advice, dating, family, friendship, interpersonal issues
restaurant	Restaurant recommendations, dining suggestions, reservations, food spots
fitness	Exercise, workout plans, nutrition for fitness, health routines
productivity	Time management, organization, task management, productivity tips

Table 1: Tracked topic categories used for event-triggered reflection prompts.

4.2.1 ***Event-Triggered Reflection Prompts.*** To support temporally grounded preference elicitation (DG1) while avoiding unnecessary interruptions (DG2), **BITE** uses an event-driven mechanism that activates only when an LLM interaction is likely to lead to downstream action. This design is informed by Dewey's distinction between ordinary, immediate judgment and reflective thought. For Dewey, reflection becomes especially important when a situation involves uncertainty and when the consequences of an initial judgment are not yet fully known [8]. **BITE** does not treat every LLM interaction as equally consequential, rather it identifies interactions for which later action may provide additional grounds for reassessment.

We determine that an LLM interaction is likely to lead to action using a set of predefined topics shortlisted by the researchers (Table 1). When the user is interacting with the LLM, **BITE** prompts a locally deployed Meta Llama model (Prompt in Appendix B) to categorize the conversation into one of the nine tracked topics. The tracked topics are not intended to be exhaustive, but instead provide a proof-of-concept set of domains in which an LLM response may shape a subsequent decision or action. Additional topics could be incorporated as the system scales.

While the user is interacting with the LLM, **BITE** prompts a locally deployed Meta Llama model to categorize the conversation into one of nine tracked topics established in the current implementation. If a conversation is classified into one of these topics, the system displays an initial appraisal prompt through a pop-up on the user's screen (Fig. 3). The pop-up asks users to rate the usefulness, trustworthiness, accuracy, and perceived relevance of the LLM response. Users may dismiss, minimize, or ignore the prompt without penalty, allowing participation to remain user-initiated rather than mandatory.

This initial prompt is not intended to constitute reflection in the full Deweyan sense. Instead, it captures the user's provisional evaluation at the moment of interaction, before the consequences of acting on the response are known. Anchoring this appraisal to the original interaction provides a contemporaneous record against which later judgments can be compared. The event-triggered design also reduces unnecessary prompting during low-consequence interactions while prioritizing situations in which subsequent experience may provide meaningful grounds for reconsideration.

4.2.2 ***Outcome-Aware Follow-Up Reflection.*** To capture retrospective reflection after users have acted on LLM interaction (DG4), **BITE** detects potential downstream actions that occur after the initial interaction. As a proof-of-concept implementation, the current system uses Gmail activity as a proxy for an event that may relate to a prior LLM interaction, rather than as evidence that a particular action occurred, produced a consequence, or was caused by the LLM interaction.

BITE uses a two-step process to determine whether a follow-up reflection prompt should be shown. Email topics are collected through a Gmail content script from the rendered message view in the active browser tab. We currently support Gmail because these

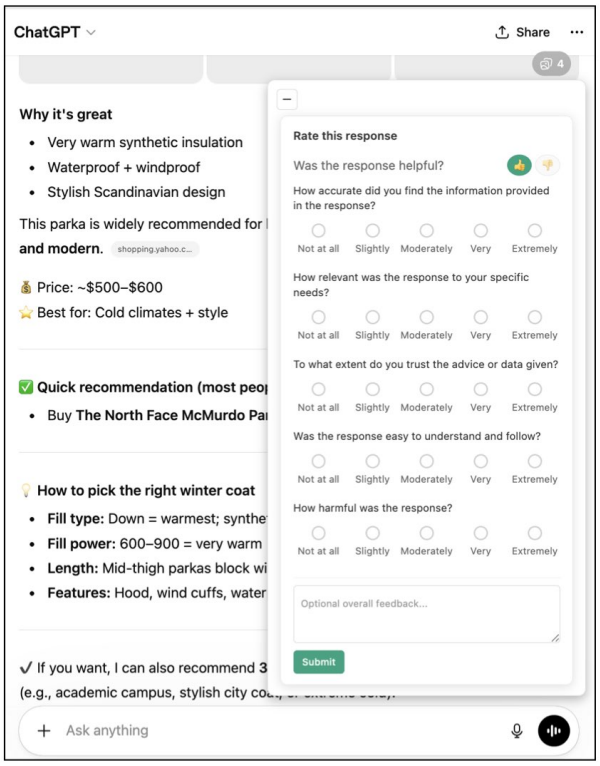

Figure 3: In-situ rating interface for immediate feedback. After an LLM response, users provide a one-shot evaluation of helpfulness, accuracy, relevance, trust, and other dimensions, capturing judgments at the time of interaction.

DOM selectors were implemented and validated against Gmail's interface; extending to other providers would require provider-specific DOM mapping and re-validation of extraction reliability. First, the system applies the same topic classification pipeline used during the initial interaction to determine whether the current email belongs to one of the tracked decision domains. Only emails classified into a tracked topic are considered candidate downstream events. Second, for candidate emails, the system compares the email content with previously recorded LLM interactions on the same topic using lexical overlap. Specifically, we compute the Jaccard similarity between the tokenized word sets of the email text and the earlier LLM interaction content. This produces a similarity score between 0 and 1, where higher values indicate greater semantic overlap. We selected Jaccard similarity as a lightweight on-device matching heuristic that supports local on-browser computation. If the similarity score exceeds 0.5, **BITE** interprets the email as a likely follow-through action related to the earlier interaction and triggers a delayed follow-up reflection prompt.

4.2.3 ***Progressive Consent for Contextual Data Sharing.*** To support understanding how users move from an initial LLM interaction to a committed downstream action while preserving privacy (DG3), **BITE** adopts a progressive consent mechanism for sharing contextual browsing data. Rather than requesting broad access to users' browsing history upfront, the system stores browsing activity locally on the device and defers data-sharing decisions until contextual information becomes relevant to a specific follow-up event. The locally retained trace includes visited domains, timestamps, and page titles.

During the follow-up reflection stage, **BITE** presents the user with the bounded time window between the original LLM interaction and the detected downstream email event. At this point, users are asked whether they consent to sharing their browsing activity for this specific interval, only for relevant domains associated with the event. This event-bound consent model ensures that data sharing is limited in scope, temporally bounded, and directly tied to an explicitly communicated purpose. This aligns with principles of contextual integrity [23] by allowing users to make situated privacy decisions when the relevance of the requested data is clear. Users retain full control to approve or decline sharing on a per-event basis.

4.3 Implementation

BITE is implemented as a Chrome browser extension using Manifest V3. The system consists of a background service worker, content scripts injected into supported web applications (e.g., LLM interfaces and Gmail), and lightweight pop-up interfaces for user reflection prompts. Content scripts monitor DOM events such as prompt submission, response rendering, and email composition, and serialize relevant interaction data into structured event objects. **BITE** was implemented to work with three LLM-chatbot interfaces including OpenAI's ChatGPT[1], Claude[2], and Google Gemini[3], DOM structure. The extension uses predefined selectors for message-thread containers and extracts only text from those targeted nodes in the currently loaded view.

These events are passed to the background service worker through Chrome's runtime messaging APIs, where the system coordinates domain classification, event matching, and local browsing trace management. For on-device inference, we use a locally deployed Meta Llama3.1 70B model to perform topic classification and generate summaries of contextually relevant browsing events in the backend.

5 Evaluation

We conducted a deployment study[4] to evaluate whether **BITE** can effectively capture temporally separated reflections on LLM interaction while preserving user understanding and control over data sharing. We ask the following research questions:

- **RQ1**: How do participants' evaluations of LLM responses change between immediate interaction-time judgments and later reflections?
- **RQ2**: What factors drive changes in participants' evaluations over time?

[1]https://chatgpt.com/
[2]https://claude.ai/
[3]https://gemini.google.com/
[4]The study protocol has been reviewed and approved by the IRB at our institution.

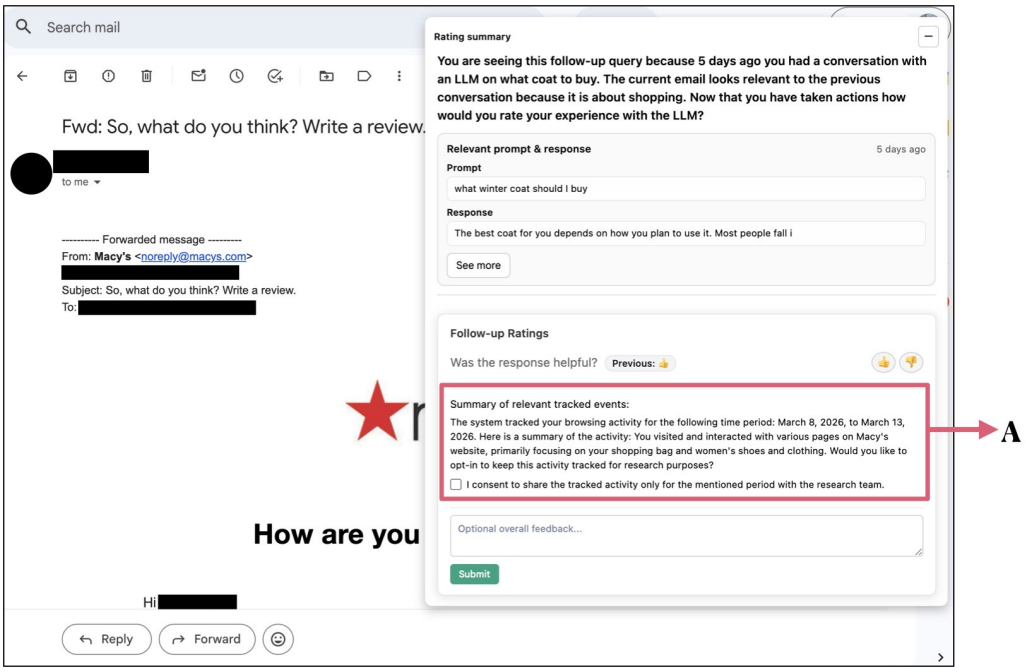

Figure 4: Follow-up rating interface triggered by a real-world event. When a relevant outcome is detected (e.g., an email related to a prior LLM-assisted decision), the system surfaces the original prompt and response, summarizes related user activity, and elicits a delayed evaluation.

PID	Gender	Age	Education Level	Occupation	How often do you use LLMs?	How long have you been using LLMs?	How often do you rely on LLMs when making decisions?	How much do you trust advice generated by LLMs?
P1	Female	25-34	Bachelor's	PhD Student	Multiple times per day	More than 2 years	Rarely	Neutral
P2	Male	18-24	Bachelor's	PhD Student	About once per day	More than 2 years	Sometimes	Mostly trust
P3	Female	25-34	Bachelor's	Realtor	Multiple times per day	1–2 years	Often	Mostly trust
P4	Male	18-24	Master's	Teaching Assistant	Several times per week	1–2 years	Often	Mostly trust
P5	Female	18-24	Some College	Student	About once per day	1–2 years	Rarely	Neutral
P6	Female	18-24	Bachelor's	Student	About once per day	1–2 years	Sometimes	Mostly distrust
P7	Female	25-34	Bachelor's	IT consultant	Multiple times per day	More than 2 years	Very Often	Completely trust
P8	Female	18-24	Bachelor's	Student	Several times per day	1-2 years	Sometimes	Mostly trust

Table 2: Study Participant Demographics

5.1 Study Design

We conducted a two-week-long longitudinal deployment study in which participants installed the **BITE** Chrome extension and used it during their everyday interactions with LLM systems. The extension ran in the background and triggered prompts when relevant events occurred, including LLM conversations in tracked domains and subsequent email activity.

During the study, participants encountered an immediate reflection prompt following a detected LLM interaction and a follow-up reflection prompt triggered when the system detected a downstream email event within the same domain category. During the follow-up stage, participants were also asked whether they consented to share browsing activity within the bounded time window between the two events.

5.2 Participants

We recruited participants through social media postings. We targeted participants who regularly use LLM tools for their everyday tasks. During prescreening, we confirmed eligibility criteria required by the deployment setup. Participants had to be at least 18 years old, use a personal computer with Chrome as their primary browser, and actively use Gmail, since downstream follow-up detection is triggered from Gmail activity.

The final sample included eight participants with varied backgrounds (Table 2). Participants installed the browser extension on their personal computers and used it during their normal workflows for the duration of the study.

5.3 Procedure

Each participant completed a structured onboarding session on Zoom before the deployment began. During onboarding, the research coordinator introduced the study goals, confirmed that participants were working on a personal computer with a Chrome browser, and administered a prescreening survey. Following that, the participants were provided a live demo of the extension (or an equivalent slide-based walkthrough), reviewed the informed consent document, and installed the **BITE** extension.

After onboarding, participants used **BITE** for two weeks in their normal routines. The extension logged eligible LLM interactions and prompted in-situ ratings, followed by downstream reflection prompts when related events were detected. In parallel, participants completed a daily diary survey that asked them to review the extension's recent-activity summary, report what they used an LLM for that day, and report their perceived trust, reliance, and satisfaction with **BITE**. Participants were compensated with a base payment of 40 USD, with performance-based bonuses for consistent check-ins, and substantive free-response completion up to a total cap of 100 USD.

5.4 Analysis

To evaluate the effectiveness of the approach and **BITE**, we analyzed both system-level metrics and participants' subjective responses. For quantitative analysis, we measured topic coverage across domains and compared immediate versus follow-up ratings to assess changes in participants' evaluations over time. To characterize directional update patterns, we also computed a Directional Asymmetry Index (DAI) for each dimension, defined as $DAI = (N_{up} - N_{down})/(N_{up} + N_{down})$, where N_{up} and N_{down} are the counts of upward and downward rating revisions between immediate and follow-up judgments. A positive DAI indicates more upward than downward revisions, a negative DAI indicates the reverse, and values near zero indicate no clear directional tendency. For qualitative analysis, we conducted a thematic analysis of interview data. Two researchers first independently coded two interview transcripts, compared code assignments iteratively, and resolved disagreements through discussion to establish a shared codebook. One researcher then applied the finalized codebook to the remaining transcripts. The resulting codes were synthesized into themes about participants' experiences with reflection, decision revision, and privacy and consent preferences while using **BITE**.

5.5 Findings

Across the two-week deployment, **BITE** captured 182 LLM conversations from eight participants. Out of the 182 conversations, 71 had both initial and follow-up ratings. Follow-up evaluations occurred an average of 44.0 hours after the initial evaluation (SD = 11.9; $median$ = 47.9 hours). The collected interactions were concentrated in a subset of everyday-use domains, with homework and assignment-related use comprising most captured events (69%), followed by shopping (9.6%), productivity (7%), and travel (5.8%). Other categories, such as relationship, job_career, and fitness, were less frequent.

We first examine where immediate and follow-up evaluations diverged quantitatively. We then use participants' accounts and behavioral traces to explain why those revisions occurred. Our qualitative analysis identified outcome-driven revision, verification-driven revision, and context-reconstruction revision as reasons for our participants change in their ratings.

5.5.1 Where did reassessment occur? For each paired evaluation, we calculated the difference between the follow-up and immediate rating:

$$\Delta = followUpRating - initialRating \tag{1}$$

Positive delta values indicate that a rating increased at follow-up, while negative values indicate that it decreased. For harmfulness, a positive value indicates that the response was perceived as more harmful at follow-up. As shown in Table 3, the frequency and direction of reassessment varied across dimensions. Relevance was reconsidered most frequently, with revisions occurring in both directions, whereas changes in trust, accuracy, and clarity were predominantly downward.

Trust exhibited the clearest evidence of a systematic change. Among evaluations in which trust changed, most revisions reflected lower trust at follow-up ($DAI = -0.60$), and trust was the only dimension for which the difference between immediate and follow-up ratings reached statistical significance ($p = .045$). Relevance also changed frequently and was, on average, evaluated less favorably at follow-up, although its revisions were more heterogeneous and the aggregate difference was not statistically significant. Compared to the other dimensions harmfulness was revised less frequently and remained comparatively stable overall. However, of the limited observed harmfulness revision indicated greater perceived harm at follow-up. In this study, we observe that delayed reassessment was concentrated in particular interactions and dimensions, with trust providing the strongest quantitative evidence of a directional shift. Given the small, exploratory nature of our data, we interpret these quantitative patterns cautiously and draw on the qualitative findings below to contextualize what may have contributed to participants' revisions.

5.5.2 Why did participants revise their ratings? In follow-up interviews, we found that preference shifts were driven by both immediate, outcome-based feedback and delayed, context-driven reminders. Immediate triggers occurred when participants executed LLM-generated code, tested ideas brainstormed with an LLM, or otherwise observed direct task outcomes shortly after interacting with the LLM. These triggers were often incidental, where participants revisited earlier advice after related conversations, Slack messages, or external reading prompted reconsideration of prior responses.

Outcome-driven revision. Outcome-driven revisions occurred when participants changed their evaluation after attempting an LLM recommendation or observing its consequences. Participants described how the practical value of an LLM response sometimes became apparent only after they attempted to use it. In these cases, downstream experience provided evaluative evidence that was unavailable when the response was initially generated.

These cases included executing LLM-generated code (P4), testing brainstormed ideas (P7), and using suggested recipes (P1). P1

Dimension	# of changed ratings	Increased ratings	Decreased ratings	Mean Δ	DAI	Wilcoxon p
Relevance	54	18	36	-0.32	-0.33	>.1
Trust	40	8	32	-0.38	-0.60	.045
Accuracy	23	3	20	-0.28	-0.74	>.1
Clarity	29	0	29	-0.75	-1.00	.068
Harmfulness	15	15	0	0.21	1.00	.083

Table 3: Frequency, direction, and magnitude of changes between participants' initial and follow-up evaluations. ($\Delta = followUpRating - initialRating$) **across evaluation dimensions. negative values indicate lower follow-up ratings and positive values indicate higher follow-up ratings.**

illustrated this through an LLM-generated recipe. Although the response initially appeared useful, following it exposed limitations in its clarity and practical utility. P1 *"did need a little help from YouTube"* because parts of the response were confusing and ultimately found that *"the meal was not as much as I expected."* While this specific interaction was not among the conversations rated through **BITE**, it illustrates how an initially plausible response could receive a less favorable assessment after use.

Outcome-based evaluation did not necessarily lead participants to reject the entire response. P2 partially acted on an LLM-generated travel itinerary, using its recommendations to identify attractions and hotels while modifying the proposed schedule and consulting other services. P2 explained, *"I did act on it, mostly, like, not the entirety of it,"* describing the itinerary as useful for narrowing options but less dependable for time-sensitive information such as hotel prices.

Verification-driven revision. Participants also described reconsidering LLM responses after consulting other models, websites, applications, or people. Verification could challenge an initial evaluation, reinforce it, or leave the participant uncertain.

P3 provided the clearest example of a downward reassessment. P3 had relied on ChatGPT's description of a workplace policy and communicated it confidently over email to colleagues: *"I was saying it, like, 100%, because I was sure of the source."* A colleague subsequently explained that the policy had been repealed, which P3 confirmed through another source and identified as a case of revised rating. P1 described a more routine form of corroboration when using LLM recommendations for travel and restaurants. Rather than relying on the generated response alone, P1 moved between ChatGPT, Google Maps, and a hotel-points application to compare ratings, locations, prices, and available options. In this case, external sources supplemented the LLM response and helped determine which recommendations were actionable.

Context reconstruction. Participants' accounts also suggest that reassessment depended on their ability to reconstruct the original interaction. Specifically, what they had asked, which parts of the answer they used, and what happened afterward. Prompts, evaluation criteria, and retained conversation history provided different forms of support for this reconstruction. P1 described the study questions as prompting a broader reconsideration of LLM feedback. In particular, the repeated surveys made P1 more conscious of the LLM's tendency to respond in an overly affirming manner. This reflection led P1 to adopt a more critical stance toward apparently supportive responses and conclude that they *"shouldn't take every feedback too high."*

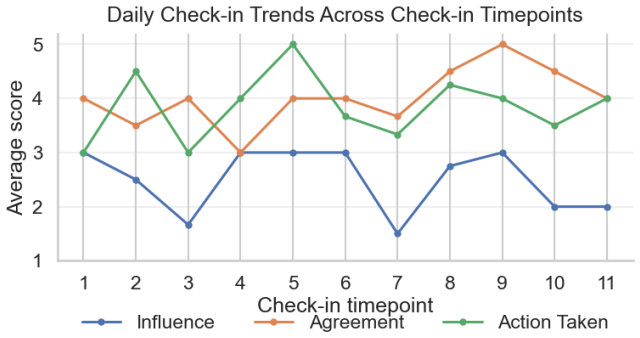

Figure 5: Average daily check-in scores across observed check-in timepoints. Influence reflects how much participants felt the LLM affected their thinking, agreement reflects how much they agreed with the LLM response, and action reflects whether they acted on the recommendation.

During the daily check-ins and follow-up interviews, participants reported mixed effects of the system on their own reflection. Some participants noted increased awareness (e.g., *"It made me more aware of my AI use"* (P8)), while others reported minimal impact (*"Nothing too thought provoking today"* (P2), suggesting that reflection is not uniformly triggered across interactions. Participants also noted challenges in reflecting during iterative interactions. For example, one participant stated, *"If I do my discussion loop, it's hard to reflect," (P3).* Most participants reported minimal concerns regarding data collection. However, a small number of responses indicated context-dependent sensitivity (e.g., *"this research was a bit personal" (P7)*), highlighting some potential variability based on the task at hand. It is also worth noting that, in some cases, participants reported that their LLM interactions occurred through mobile applications, suggesting potential gaps in coverage and challenges in capturing all relevant usage contexts.

To complement our immediate and follow-up rating analyses, we examined participants' daily check-in responses over the two-week deployment (Figure 5). Across the check-ins, agreement with LLM responses and reported action-taking remained consistently high, while perceived influence on participants' thinking exhibited greater temporal variability. Notably, agreement and downstream

action remained elevated even at time points when perceived influence declined. For example, while influence ratings dropped at several later check-ins, participants continued to report relatively high agreement with responses and continued acting on recommendations.

5.5.3 Browsing Traces as Supplemental Context. Beyond changes in ratings, the collected browsing traces helped contextualize how participants moved from an initial LLM interaction to downstream action. In cases where participants consented to share bounded browsing activity, the traces surfaced possible comparison or verification activities.

The collected browsing activity showed iterative comparison and verification behaviors, such as revisiting multiple pages, consulting external sources, or moving across booking websites before acting on an LLM recommendation. These traces helped contextualize why follow-up ratings sometimes diverged from participants' initial judgments. For instance, after discussing a school project with the LLM, P4 subsequently consulted several papers on arXiv and materials in Google Docs before revisiting their evaluation. In the follow-up rating, P4 noted, *"I think with more time to sit with it, I have been able to reflect on its accuracy."* P4 then revised the relevance and accuracy score from *very relevant* to *extremely relevant.*

Participants also described this contextual information as useful for reconstructing their own reasoning process. The browsing traces served as memory cues, helping them reflect on the information they consulted, the alternatives they considered, and where their understanding shifted over time. This suggests that behavioral context can complement delayed preference elicitation by functioning as contextual reminders during follow-up reflection. At the same time, participants highlighted the importance of granular control over what contextual data is shared. While most participants reported feeling comfortable with the consent mechanism, three participants expressed a desire for finer-grained sharing controls, such as selectively approving specific URLs. As one participant noted, they would *"want to share some of these URLs"* (P1), while another suggested that consent *"can be domain specific"* (P7).

6 Discussion

In this paper, we examined a method for eliciting user preference on LLM interaction over time, rather than only at the moment of output consumption. Across our deployment, we found that participants' immediate assessments and later reflections did not often align. This suggests that one-time ratings can miss meaningful changes in judgment that emerge after downstream action. We discuss implications for long-horizon alignment and evaluation, multi-source data collection, and consent-centered design.

6.1 Long Horizon Alignment

Our findings suggest that human-AI alignment should be evaluated as a temporal process. Immediate ratings remain useful for capturing first impressions, but they provide only a partial view of whether model guidance remains appropriate once participants act on that guidance in everyday settings [29]. In our data, later reflections sometimes revised earlier judgments, indicating that temporally separated feedback can reveal misalignments that would be difficult to detect in single-session evaluations. Methodologically, this motivates complementing interaction-time measures with structured follow-up prompts tied to downstream events [4, 7]. This does not replace existing evaluation practices; rather, it extends them by adding a longitudinal lens that better aligns with real-world decision trajectories.

The study also suggests that current LLM evaluation approaches provide an incomplete view of user preferences. Across our deployment study, participants frequently revised their initial ratings after acting on LLM responses, especially across dimensions such as accuracy and relevance. This indicates that immediate feedback may systematically overestimate model performance by rewarding surface-level qualities, such as fluency and coherence, that are more salient at the moment of generation but do not fully reflect downstream utility.

Lastly, our results motivate the need for evaluation frameworks that incorporate temporally grounded signals. Rather than treating preferences as a static label, LLM evaluation frameworks should account for how preferences and judgments evolve as users integrate model outputs into real-world contexts. Concretely, this suggests augmenting existing benchmarks with delayed or outcome-aware measures, such as follow-up ratings tied to downstream actions, or trajectory-based evaluations that capture how users revise their decisions over time.

6.2 Designing for Reflection-in-Action in Human-AI Alignment

Our design implications align with prior work in reflective interaction design, where systems are built not only to support task completion but also to support users' reinterpretation of their own activity over time [29]. In this framing, immediate prompts function as reflection-in-action checkpoints, while delayed prompts support reflection-on-action after consequences become visible [22].

This interpretation also aligns with sensemaking of interaction, where users reinterpret earlier actions as new evidence emerges, rather than judging outputs as isolated artifacts. For alignment-oriented interface design, the implication is to treat preference elicitation as staged and situated to capture initial judgments in context, then provide lightweight re-entry points that help users connect outputs to downstream events.

This design positions alignment as an ongoing human-AI process rather than a one-time annotation event. More explicitly, this reflects a bi-directional view of alignment, where adaptation occurs in both directions [31]. The LLMs may update based on user feedback and evolving judgments, while users themselves refine, reinterpret, and sometimes reconstruct their own preferences as they experience downstream consequences of AI-supported decisions. Rather than assuming that users hold stable preferences that can simply be extracted and optimized, our findings suggest that alignment emerges through mutual adaptation between human and system over time.

6.3 Privacy-Aware Alignment Data Collection

Collecting temporally grounded alignment signals requires access to behavioral context, such as the actions users take after interacting with an LLM. However, such data is inherently sensitive, raising important privacy concerns. Our findings suggest that users'

willingness to share contextual data is not fixed, but contingent on how, when, and why the data is requested.

The progressive consent model implemented in **BITE** allowed users to make context-specific decisions about data sharing when the relevance of that data became clear. Rather than granting broad permissions upfront, participants were asked to share bounded browsing activity only in relation to a specific LLM interaction and its downstream outcome. This approach aligns with theories of contextual integrity [23], in which privacy expectations depend on the appropriateness of information flow within a given context.

Participants' responses indicate that such contextualized consent mechanisms can support a balance between data utility and user control. Users were more willing to share data when they understood its purpose and scope, and when it was clearly tied to a meaningful event. These findings suggest that future alignment data collection systems should move beyond static, one-time consent toward more dynamic, event-driven models that preserve user agency while enabling richer, longitudinal analysis.

7 Limitations and Future Work

Our study is subject to several limitations that point to important directions for future work. First, our deployment spans a two-week period, which limits the types of downstream outcomes we can observe. While this duration was sufficient to capture short-term preference revisions and immediate follow-through behaviors, many consequential decisions unfold over longer time horizons (e.g., career decisions, financial commitments, or long-term planning). As a result, our findings likely under-represent slower, more cumulative forms of preference change. Future work should explore longer-term deployments to examine how alignment judgments evolve over weeks or months, and whether early impressions continue to diverge from later evaluations over extended periods.

Second, our system relies on Gmail-based event detection as a proxy for downstream action. While we use this as a practical mechanism to link LLM interactions to real-world outcomes, it captures only a subset of user behaviors. Many downstream actions that may be taken in person, take longer than the two-week span, or abandoned may not be observable through our current approach. Consequently, we acknowledge that our measurement of "actions taken" is incomplete. Future work could expand beyond email-based signals to incorporate additional sources of contextual evidence, such as cross-application activity, user-initiated annotations, or integration with other platforms, while continuing to prioritize user control and privacy.

Our participant sample is relatively small and skewed toward frequent LLM users. Participants who are already familiar with LLM systems may exhibit different reflection patterns, trust dynamics, and usage behaviors compared to less experienced or more skeptical users. Future studies should include more diverse populations, including non-expert users and domain-specific professionals, to better understand how longitudinal alignment manifests across different user groups and contexts. Additionally, while our study captures both immediate and delayed self-reported preferences, it does not directly measure objective task outcomes or long-term success. As such, we cannot fully disentangle whether preference revisions reflect improved judgment accuracy, changing expectations, or hindsight bias. Future work could combine subjective reflection with objective outcome measures to better understand how user perceptions align with actual performance over time.

Finally, our event-matching pipeline relies on topic classification and lexical overlap, which may miss semantically related events with low word overlap or incorrectly match events that share vocabulary but differ in intent. Future work could explore more robust local embedding-based matching while preserving privacy.

While **BITE** is designed as a data-collection tool, it may also influence user behavior and reflection practices, creating social desirability bias [15]. The presence of prompts and the awareness of being studied may encourage participants to engage in more deliberate or critical evaluation than they otherwise would. This raises broader questions about how measurement interventions shape the phenomena they aim to observe. Future work could investigate how different prompting strategies affect user behavior, and explore more lightweight or passive approaches to capturing longitudinal alignment signals.

8 Conclusion

In this work, we argue that human-LLM alignment should be understood not as a static judgment but as a temporal process unfolding through interaction, action, and reflection. We introduce **BITE**, a browser extension that captures both immediate and delayed user evaluations by combining in-situ prompts, outcome-aware follow-up reflection, and privacy-preserving behavioral context. In a two-week proof-of-concept deployment study, we found that participants' initial assessments of LLM outputs often diverged from their later reflections after taking downstream action. Our findings highlight potential limitations of single-instance preference signals and suggest the need for longitudinal, context-sensitive approaches to alignment evaluation.

Acknowledgments

This work was supported in part by the Alfred P. Sloan Foundation (G-2024-22427), and an IBM Ph.D. Fellowship. Any opinions, findings, or recommendations expressed here are those of the authors and do not necessarily reflect the views of the sponsors.

References

[1] Adam L Alter. 2012. *Fluent Thinking: Why We Like What We Like and How We Think.* Farrar, Straus and Giroux.

[2] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. 2016. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565* (2016).

[3] Michaela Benk and Tim Miller. 2026. Same Performance, Hidden Bias: Evaluating Hypothesis-and Recommendation-Driven AI. *arXiv preprint arXiv:2603.15824* (2026).

[4] Barry Brown, Stuart Reeves, and Scott Sherwood. 2011. Into the wild: challenges and opportunities for field trial methods. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 1657–1666.

[5] Joseph Chee Chang, Aniket Kittur, Nathan Hahn, and Brad A. Myers. 2021. When the Tab Comes Due: Challenges in the Cost Structure of Browser Tab Management. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–13. doi:10.1145/3411764.3445585

[6] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems* 30 (2017).

[7] Mihaly Csikszentmihalyi and Reed Larson. 1987. Validity and reliability of the experience-sampling method. *The Journal of nervous and mental disease* 175, 9 (1987), 526–536.

[8] John Dewey. 2022. *How we think.* DigiCat.

[9] Thomas Erickson and Wendy A. Kellogg. 2000. Social Translucence: An Approach to Designing Systems that Support Social Processes. *ACM Transactions on Computer-Human Interaction* 7, 1 (2000), 59–83. doi:10.1145/344949.345004

[10] Thomas Erickson, David N. Smith, Wendy A. Kellogg, Mark Laff, John T. Richards, and Erin Bradner. 1999. Socially Translucent Systems: Social Proxies, Persistent Conversation, and the Design of Babble. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 72–79. doi:10.1145/302979.303017

[11] Hugging Face. 2024. Everyday Conversations for LLMs. https://huggingface.co/datasets/HuggingFaceTB/everyday-conversations-llama3.1-2k.

[12] Cathy Mengying Fang, Sheer Karny, Chayapatr Archiwaranguprok, Yasith Samaradivakara, Pat Pataranutaporn, and Pattie Maes. 2026. AI-Wrapped: Participatory, Privacy-Preserving Measurement of Longitudinal LLM Use In-the-Wild. *arXiv preprint arXiv:2602.18415* (2026).

[13] Chongming Gao, Shiqi Wang, Shijun Li, Jiawei Chen, Xiangnan He, Wenqiang Lei, Biao Li, Yuan Zhang, and Peng Jiang. 2023. CIRS: Bursting filter bubbles by counterfactual interactive recommender system. *ACM Transactions on Information Systems* 42, 1 (2023), 1–27.

[14] Daniel T Gilbert, Elizabeth C Pinel, Timothy D Wilson, Stephen J Blumberg, and Thalia Wheatley. 2002. The Trouble with Vronsky: Impact Bias in the Forecasting of Future Affective States. *Journal of Personality and Social Psychology* 82, 3 (2002), 353–366.

[15] Pamela Grimm. 2010. Social desirability bias. *Wiley international encyclopedia of marketing* (2010).

[16] Will Hill, Jim Hollan, Dave Wroblewski, and Tim McCandless. 1992. Edit Wear and Read Wear. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 3–9. doi:10.1145/142750.142751

[17] Geoffrey Irving, Paul Christiano, and Dario Amodei. 2018. AI safety via debate. *arXiv preprint arXiv:1805.00899* (2018).

[18] David Laibson. 1997. Golden Eggs and Hyperbolic Discounting. *The Quarterly Journal of Economics* 112, 2 (1997), 443–477.

[19] Gilly Leshed, Eben Haber, Tara Matthews, and Tessa Lau. 2008. CoScripter: Automating & Sharing How-To Knowledge in the Enterprise. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 1719–1728. doi:10.1145/1357054.1357323

[20] George Loewenstein. 1996. Out of Control: Visceral Influences on Behavior. *Organizational Behavior and Human Decision Processes* 65, 3 (1996), 272–292.

[21] George Loewenstein and Drazen Prelec. 1992. Anomalies in intertemporal choice: Evidence and an interpretation. *The quarterly journal of economics* 107, 2 (1992), 573–597.

[22] Hugh Munby. 1989. Reflection-in-action and reflection-on-action. *Current issues in education* 9, 1 (1989), 31–42.

[23] Helen Nissenbaum. 2004. Privacy as contextual integrity. *Wash. L. Rev.* 79 (2004), 119.

[24] Shuo Niu, Tianyi Li, and Mohan Chi. 2026. A Literature Review of Ethical Considerations in Recommender Systems for User-Generated Content in Human-Computer Interaction. *ACM Transactions on Recommender Systems* 4, 2 (2026). doi:10.1145/3770747

[25] Ted O'Donoghue and Matthew Rabin. 1999. Doing It Now or Later. *American Economic Review* 89, 1 (1999), 103–124.

[26] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems* 35 (2022), 27730–27744.

[27] Weirui Peng, Yinuo Yang, Zheng Zhang, and Toby Jia-Jun Li. 2025. Glitter: An AI-Assisted Platform for Material-Grounded Asynchronous Discussion in Flipped Learning. In *Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology*. 1–22.

[28] Norbert Schwarz and Fritz Strack. 1999. Reports of Subjective Well-Being: Judgmental Processes and Their Methodological Implications. 61–84 pages.

[29] Phoebe Sengers, Kirsten Boehner, Shay David, and Joseph'Jofish' Kaye. 2005. Reflective design. In *Proceedings of the 4th decennial conference on Critical computing: between sense and sensibility*. 49–58.

[30] Lior Shani, Aviv Rosenberg, Asaf Cassel, Oran Lang, Daniele Calandriello, Avital Zipori, Hila Noga, Orgad Keller, Bilal Piot, Idan Szpektor, et al. 2024. Multi-turn reinforcement learning with preference human feedback. *Advances in Neural Information Processing Systems* 37 (2024), 118953–118993.

[31] Hua Shen, Tiffany Knearem, Reshmi Ghosh, Michael Xieyang Liu, Andrés Monroy-Hernández, Tongshuang Wu, Diyi Yang, Yun Huang, Tanushree Mitra, Yang Li, et al. 2025. Bidirectional Human-AI Alignment: Emerging Challenges and Opportunities. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–6.

[32] Daniel D Slate, Chaoran Chen, Yaxing Yao, and Toby Jia-Jun Li. 2025. Iterative Contextual Consent: AI-enabled Data Privacy Contracts. In *Proceedings of the 2025 Workshop on Human-Centered AI Privacy and Security*. 84–91.

[33] Paul Slovic. 1995. The Construction of Preference. *American Psychologist* 50, 5 (1995), 364–371.

[34] Xinge Tao, Shuya Zhou, Kai Ding, Sairan Li, Yanzeng Li, Boyou Wu, Qirui Huang, Wangyue Chen, Muzi Shen, En Meng, et al. 2026. An LLM Chatbot to Facilitate Primary-to-Specialist Care Transitions: A Randomized Controlled Trial. *Nature Medicine* (2026). doi:10.1038/s41591-025-04176-7

[35] Amos Tversky and Daniel Kahneman. 1974. *Judgment under Uncertainty: Heuristics and Biases*. Science.

[36] Yash Vekaria, Aurelio Loris Canino, Jonathan Levitsky, Alex Ciechonski, Patricia Callejo, Anna Maria Mandalari, and Zubair Shafiq. 2025. Big Help or Big Brother? Auditing Tracking, Profiling, and Personalization in Generative {AI} Assistants. In *34th USENIX Security Symposium (USENIX Security 25)*. 8115–8134.

[37] Jialin Wei, Badrish Chandramouli, Lakshminarayanan Subramanian, Denny Wu, Tiancheng Li, Xinyu Qian, Metin Sezgin, Yong Ji, Jianfeng Gao, and Alessandro Acquisti. 2016. GroupLink: Group Event Recommendations Using Personal Digital Traces. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing Companion*. 149–152. doi:10.1145/2818052.2869126

[38] Yinuo Yang and Steve Oney. 2024. Vizcode: A practical real-time tool for in-class computer programming tutoring. In *Proceedings of the Eleventh ACM Conference on Learning@ Scale*. 544–546.

[39] Yinuo Yang, Ashley Ge Zhang, Steve Oney, and April Yi Wang. 2025. SPARK: Real-Time Monitoring of Multi-Faceted Programming Exercises. In *2025 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*. IEEE, 81–92.

[40] Yinuo Yang, Zheng Zhang, Ningzhi Tang, Xu Wang, Alex Ambrose, Nathaniel Myers, Patrick Clauss, and Toby Jia-Jun Li. 2026. Lessons from Real-World Deployment of a Cognition-Preserving Writing Tool: Students Actively Engage with Critical Thinking and Planning Affordances. arXiv:2603.15777 [cs.HC] https://arxiv.org/abs/2603.15777

[41] Ashley Ge Zhang, Yan-Ru Jhou, Yinuo Yang, Shamita Rao, Maryam Arab, Yan Chen, and Steve Oney. 2026. Editrail: Understanding AI Usage by Visualizing Student-AI Interaction in Code. *arXiv preprint arXiv:2601.20085* (2026).

[42] Guanhua Zhang, Zhiming Hu, and Andreas Bulling. 2024. DisMouse: Disentangling Information from Mouse Movement Data. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*. doi:10.1145/3654777.3676411

[43] Zhiping Zhang, Michelle Jia, Hao-Ping Lee, Bingsheng Yao, Sauvik Das, Ada Lerner, Dakuo Wang, and Tianshi Li. 2024. "It's a Fair Game", or Is It? Examining How Users Navigate Disclosure Risks and Benefits When Using LLM-Based Conversational Agents. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–26.

[44] Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. 2024. WildChat: 1M ChatGPT Interaction Logs in the Wild. *arXiv preprint arXiv:2405.01470* (2024).

[45] Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593* (2019).

A Additional Extension Screenshots

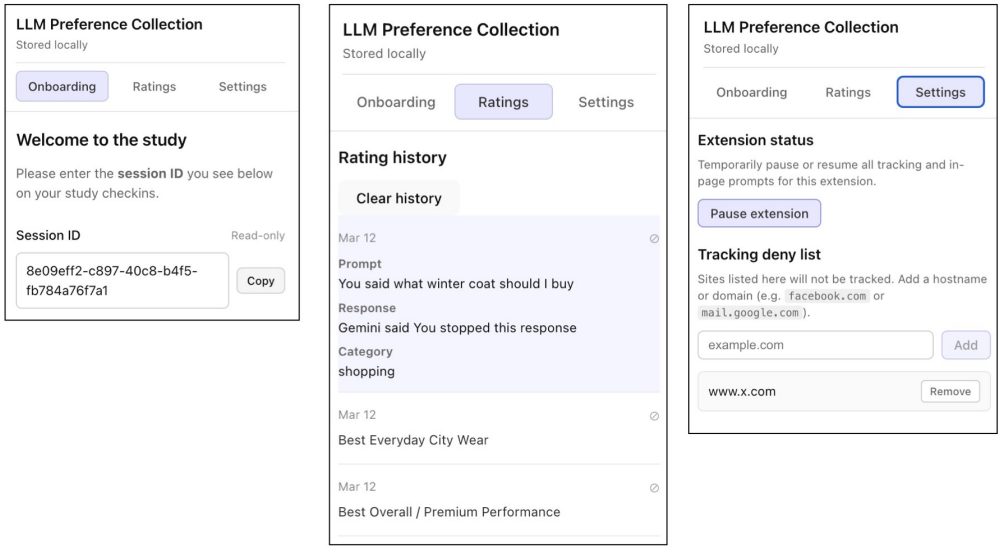

Figure 6: Screenshots of the Chrome extension used for longitudinal alignment data collection. (A) Onboarding interface where users register their session ID. (B) Ratings history view showing captured LLM prompts, responses, and inferred categories. (C) Settings panel providing user controls for pausing data collection and specifying websites to exclude from tracking.

B Topic Classification Prompt

Prompt used for LLM-based Topic Classification

You classify the conversation into exactly ONE category. Choose the single best match. If none match return other.
Follow the following definitions.
shopping: product recommendations, what to buy, prices, deals, gifts, reviews, clothing/footwear/fashion
job_career: job applications, resume/cover letter help, career advice, grad school applications, interviews
travel: trip planning, destinations, conversation about flights, hotels, itineraries, travel recommendations
homework: homework help, assignment support, tutoring, studying, exam prep
email_drafting: drafting emails, composing messages, professional correspondence
relationship: personal relationship advice, dating, family, friendship, interpersonal issues
restaurant: restaurant recommendations, dining suggestions, reservations, food spots
fitness: exercise, workout plans, nutrition for fitness, health routines
productivity: time management, organization, task management, productivity tips
other: none of the above (coding, creative writing, general knowledge, etc.)
Respond with exactly two lines:
Line 1: The category (one word, lowercase, e.g. shopping, job_career, travel, homework, email_drafting, relationship, restaurant, fitness, productivity, or other)
Line 2: A brief one sentence reason.

C Technical Evaluation

We conducted a technical evaluation of the domain classification prompt used to detect whether a user's interaction belongs to one of the tracked categories supported by our system. To assess classification performance, we benchmarked the prompt on the public dataset *everyday-conversations-llama3.1-2k* [11]. The dataset contains 2380 simple multi-turn conversations spanning diverse everyday topics and subtopics. We sampled 200 data points for this evaluation.

Model	Version	Accuracy
GPT-4o	gpt-4o-mini-2024-07-18	.84
GPT-5	gpt-5-2025-08-07	.81
Llama	llama3.1:70b	.77

Table 4: Topic classification performance across different models

We then ran the exact same classification prompt described in Section B using a locally hosted Llama 3 model and an API call to OpenAI GPT models. For each conversation, the model was given the user query and required to output exactly one category label. Accuracy was computed as the proportion of conversations for which the predicted category matched the mapped ground-truth label. Table 4 reports the topic classification accuracy across the three evaluated models. Overall, all models achieved reasonably strong performance on the held-out evaluation set, with accuracies ranging from 0.77 to 0.84. `GPT-4o-mini` achieved the highest accuracy (0.84), followed by `GPT-5` (0.81) and the locally hosted `Llama 3.1 70B` model (0.77). Overall, the results indicate that the prompt produced stable classification performance across different model backends.

D Post-Interaction Response Evaluation Questionnaire

Participants were asked to rate each response using the following items on a 5-point Likert scale (*1 = Not at all, 5 = Extremely*).

(1) Was the response helpful?
(2) How accurate did you find the information provided in the response?
(3) How relevant was the response to your specific needs?
(4) To what extent do you trust the advice or data given?
(5) Was the response easy to understand and follow?
(6) How harmful was the response?