ANTHROP\C

Appendix to “Enabling independent research on how people use Claude”

August 2026

This appendix contains additional details and results for "Enabling independent research on how people use Claude."

Updated Oct 30, 2026: We've revised the appendix to add credit that was missing at publication. The privacy audit section now names the researchers who conducted it, and the privacy threat model now cites the CLIOPATRA attack described by Annamalai et al. (2026).

Additional details on our external research partnerships

For this pilot, our Societal Impacts team partnered with three research groups: the Social and Language Technologies Lab at Stanford University, the Human Information Processing Lab at the University of Oxford, and METR, a non-profit organization that evaluates frontier AI models to help society understand AI capabilities and what risks they pose. We chose these groups as our initial partners due to their expertise in areas that we lacked. They were also groups we had existing relationships with and trusted to iterate quickly with the rough edges of a pilot.

We started our partnerships in February 2026. In preparation, we wrote a research primer which described the goals of the pilot and the technical details of Anthropic Insights necessary to design research the tool can support.

In brief: Anthropic Insights works by having Claude read a sample of conversations and answer a fixed set of researcher-written questions about each one—for example, "what task is the user working on?" or "does the assistant push back on the user at any point?" We call each question a *facet*. Answers to open-ended facets are then grouped by similarity into *clusters*, and Claude writes a short description of each group; answers to multiple-choice or numeric facets (e.g. "how many self-corrections from Claude occur in this conversation?") are simply tallied.

Researchers can also specify facets computed programmatically (i.e., without Claude), like the number of human messages in the conversation. Any cluster containing fewer than a minimum number of conversations or accounts is dropped. A *configuration* is a partner's full set of facets and settings; a *run* applies it to an Anthropic Insights sample that computes the answers to a given set of facets. The outputs of an Anthropic Insights run are cluster descriptions and tallies and the corresponding number of conversations that fall under each category. Outputs are privacy-preserving and never contain conversations themselves. Our paper on Anthropic Insights describes the system in detail.

Alongside the primer, we drafted a collaboration agreement with each partner. These agreements were designed to prioritize the independence of their research: each partner drove the research questions and study design, while Anthropic facilitated data access, reviewed the data to ensure we were sharing it responsibly, and provided research support. Our review rights were limited to: user privacy, information that could help people violate our usage policies, Anthropic's confidential information, and research accuracy. **Beyond these areas, Anthropic did not have a say in the content of the research and our pilot partners are free to publish their findings, even if they are inconvenient for Anthropic.**

Throughout the pilot, no external researcher ever had access to raw conversation data. All raw data and computation remain on Anthropic's servers; the research partners received only the aggregated, privacy-preserving outputs from Anthropic Insights. The processes below—validation runs, legal review, and manual output review—are additional safeguards layered on top of Anthropic Insights' design.

Each partner submitted a research proposal with their research questions and the corresponding Anthropic Insights configuration. This configuration went through a few rounds of iteration. First, members of the Societal Impacts team collaboratively iterated on each partner's configuration to ensure Anthropic Insights was set up to best support each research proposal. Our input at this stage was technical, helping partners express their questions in a form Anthropic Insights could run. We assessed proposals for potential legal risk, but did not steer configurations based on whether the findings would reflect well on Anthropic.

Next, we launched Anthropic Insights runs on conversation data from WildChat—a publicly available dataset of AI chat conversations—so the partners could validate their configurations and see real Anthropic Insights outputs before any Claude user data was involved. After reviewing the WildChat data, the external partners had a final opportunity to revise their configurations before their full Anthropic Insights run. Finalized configurations were then reviewed by our legal and privacy teams to ensure that they met the same privacy and legal standards we apply to our own research.

We then launched the full Anthropic Insights runs on Claude user data: ~250K Claude.ai conversations each for Stanford and Oxford, and ~250K Claude Code conversations for METR. No data from Team, Enterprise, or API customers was included in the pilot. Each run drew its own unique sample from a fixed window in April-May 2026. All conversations came from users on Free, Pro, and Max plans. For Claude Code, we sampled from consumer users who had opted in to letting Anthropic use their data to improve our models, and chose a window that straddled a Claude model release so METR could compare usage before and after. We funded all Anthropic Insights runs and provided partners' API credits for their WildChat iterations when necessary.

Upon completion of each external partner's Anthropic Insights run, we manually reviewed every cluster output from each Anthropic Insights run before sharing the data with the partner. Our review evaluated the data for substantive quality issues as well as whether sharing this data would compromise user privacy or safety. As mentioned in the main text, some outputs from partners' Anthropic Insights runs surfaced violations of our Acceptable Use Policy or Terms of Service. We want to be transparent about misuse of Claude without encouraging it. As such, we shared the vast majority of outputs describing misuse on our

platform; we only redacted outputs that described how users got around our safeguards rather than what they attempted. Any time we redacted or removed clusters, we also provided a brief reason why. Across the three runs, of the data we reviewed (open-ended clusters), we redacted or removed 1.9% of clusters accounting for 4.28% of conversations for Stanford, 3.33% of clusters[1] accounting for 3.85% of conversations for Oxford, and 1.8% of clusters accounting for 2.96% of conversations for METR.

To ensure we maximally pursued the program's goal of enabling independent and high-quality research on AI's societal impacts, the final judgment of what data to share rested with our Societal Impacts team. We have released the exact outputs received by our partners' in their research on HuggingFace.

Our partners then had ~60 days to analyze their outputs and share initial write-ups with our team. We reviewed these write-ups for research accuracy in how they described the study methodology, Anthropic Insights, and the output data, consistent with the review rights in our agreements. We requested some corrections, concerning descriptions of Anthropic Insights' methodology. However, the findings and conclusions are solely those of the researchers, regardless of whether or not they reflect favorably on Anthropic.

Additional details on our third-party privacy audit

Anthropic Insights was designed to protect user privacy. As described in our 2024 paper, no one, including Anthropic employees, sees the underlying conversations when using Anthropic Insights for research, only aggregated cluster names and descriptions. Clusters must contain a minimum number of distinct users and conversations, and they are stripped of identifying information. Furthermore, before we shared data with researchers, Anthropic staff manually reviewed every cluster name and description for privacy concerns.

We think these protections are strong and effective, but as we hope to give more researchers access to this data, we want to be explicit about the privacy bar we hold ourselves to and show we are meeting it. To this end, we wrote a privacy threat model laying out the risks we believe the system needs to guard against (see the next section for an updated version), shared that document along with the clusters we are publishing today to a group of privacy red-teamers at the AI Security and Privacy Lab at Imperial College London, and gave them two weeks to try to reidentify users or otherwise violate the threat model.

[1] This percentage excludes a facet that we removed entirely from Oxford's outputs, because we strongly suspected a misphrased prompt caused the Anthropic Insights to output misleading cluster descriptions. See "Interpreting Open-Ended Anthropic Insights Clusters" for more on how that can occur.

The red-teamers were unable to reidentify any users or find any violation of our threat model. They did surface one pattern that falls outside of our threat model, which did not lead us to redact any clusters, but we consider it worth addressing.

When many unrelated users may use the same tool, such as an agentic harness or open source library, the cluster description may contain phrases that are unique to the tool's design or harness (or describe one of its functions in sufficient detail to plausibly identify it). Using this, the red-teamers were able to link one cluster to the use of a popular open source project with a high degree of confidence. This is not a violation of our threat model, since they could not link use of that project to any identifiable person, small group, or organization. We also consider it very unlikely that this pattern could surface a private or internal tool, since such a tool would have to be in use across many unrelated users and these studies did not draw on Team, Enterprise, or API customer data. Still, we are taking extra steps to address even that possibility.

The most effective mitigation here is to increase the minimum number of users a cluster must contain. In future studies we plan to raise these minimums substantially, so that a tool would have to be in extremely wide use before it could surface this way, if at all. We also plan to change how cluster descriptions are generated so they are less likely to carry over distinctive phrasing, and to run the red-teamers' linking method ourselves on future data releases. In general, we plan to keep working with outside privacy red-teamers as this program develops.

We thank Zexi Yao, Bozhidar Stevanoski, Peter Romov, Euodia Dodd, Xiaoxue Yang, and Nataša Krčo for conducting this exercise.

Privacy threat model

Expanding research access must not undermine our users' privacy. Sharing data that exposes users' private information can lead to serious real-world harms, but sharing no data at all leaves the world navigating this technological change blind.

Below is the privacy threat model we use to stress-test the Anthropic Insights data we share with outside researchers, updated since we gave it to the privacy red-teamers who evaluated this release. It is written with Anthropic Insights specifically in mind, but the same principles could apply to any AI provider sharing usage data for research. It does not cover Anthropic's handling of user data in general, our internal use of Anthropic Insights, or the security of our infrastructure.

The threat model borrows from two pieces of existing data protection guidance. The objective is for the data we release to be anonymous, and we assess each release against the three criteria in the European Data Protection Board's draft guidelines on anonymization: no record isolation (the data contains no combination of attributes that is unique to one person), no linkage (the data cannot be matched to information about the same person held elsewhere), and no inference (the data does not support any specific, meaningful conclusion about an individual). The adversary we model is based on the UK Information Commissioner's Office's "motivated intruder": a competent person who wants to reidentify someone from the released data and will use any publicly available resource to do it. We also assume the adversary has access to the most capable AI models available.

A. What is the adversary's goal?

We treat the following outcomes as privacy harms, and assume an adversary may pursue any of them.

1. **Identity linkage.** Linking a specific, identifiable person to their use of Claude, or to something they discussed with it. *Example: showing that a particular named person used Claude to discuss their bankruptcy.*
2. **Attribute disclosure.** Learning a sensitive, non-public fact about an identifiable person, even where the data never states it directly. *Example: inferring that a particular person has a serious medical condition.*
3. **Small-group disclosure.** Attributing a fact to members of a small, identifiable group. *Example: learning something sensitive about the members of one recognizable local religious congregation.*
4. **Organizational disclosure.** Linking an identifiable organization to a non-public activity. This is a commercial confidentiality concern rather than a personal privacy one, but we treat it as in scope. *Example: finding a recognizable company's unannounced product plans.*

These goals could apply across studies as well as within one. For Anthropic Insights specifically, considering data from multiple studies is not an effective path to reidentification, since each study draws on its own random sample and time window. We would revisit this if we ever ran multiple studies on one sample or supported longitudinal research.

We also considered and discounted some goals that we do not think rise to the level of a privacy harm:

- **Population-level inference about anonymous users.** Group-level patterns are the subject of research on how people use AI. They become a privacy harm when the group is so narrow that membership exposes individuals, which is covered by the small-group disclosure case above. (Group privacy concerns should be addressed in study design and approval.) Keeping these sorts of large-scale patterns intact is critical for the effectiveness of the research, and undermining them seriously damages the utility of the data. *Example: learning that some Claude users are job hunting. This is a fact about millions of people, and no individual can be shown to be among them, however sensitive the topic.*
- **Unflattering aggregate findings.** Discovering that Claude is used in ways Anthropic or others might prefer it were not is a research result, not a privacy harm. *Example: a cluster reveals a large volume of Claude users seeking advice on how to cheat on their partners.*
- **Skewing what a study finds.** An adversary might want to influence the outcome of a third party's research about how people use Claude and plant conversations so that a topic looks larger than it is or so that a cluster appears at all. This is a research integrity concern rather than a privacy one, and section C explains why we consider the mechanism impractical.

We also highlight here two additional behaviors specific to Anthropic Insights that could increase the likelihood of a privacy violation but do not constitute violations in and of themselves.

- **Recovering suppressed counts.** We suppress any cluster or intersection between clusters that falls below our minimum user threshold. However, when a researcher's query sorts every conversation into a fixed set of categories, a suppressed category can sometimes be worked out by process of elimination. As in our prior public data releases, we do not treat this as a privacy harm in itself, only if it leads to one of the disclosures above. *Example: a cluster of 100 conversations reported as 70% "low" frustration and 29% "medium" tells you the one remaining conversation was "high."*
- **Identifying specific tool uses.** When many unrelated users use the same tool, its cluster description can carry over enough of the tool's own text to identify it. As described above, our red-teamers linked one cluster to a popular open source project this way. We do not treat this as a privacy harm in itself, since it does not on its own reveal any information about an identifiable person, small group, or organization. *Example: a cluster description closely paraphrases the setup instructions of a widely used code-formatting tool.*

B. What information and resources does the adversary have available?

We assume the adversary has access to all of the research data we publish and knows everything about how the research method works. For Anthropic Insights, that means they start with the full set of clusters, with our redactions applied and removed clusters absent. They know how the clustering works, including prompts substantially the same as those in our 2024 paper, and have everything published alongside the release, including the study's queries, its criteria for which conversations to include, which Claude products it sampled from, how many conversations it drew, and the length of the sampling window. We do not treat any part of the method or a study's configuration as secret.

We consider side information to be in scope, but bounded. We assume the adversary can use anything a member of the public could reasonably obtain: the open web and news archives, social media and public records, other published Anthropic Insights datasets and public chat corpora such as WildChat, and their own conversations with Claude. We also assume they will spend some money on AI compute, though how much depends on the cost and capability of models over time and on what an adversary stands to gain from the particular dataset. With Anthropic Insights, we gave privacy red-teamers $1000 in Claude credits with the offer to provide more if they needed it.

C. What level of access does the adversary have to the system?

We assume the adversary is an outsider who works only from what we publish. We also considered two ways someone could influence the pipeline without access to it — by feeding it conversations or by shaping a study — and found neither to be a serious risk.

The first is the adversary who feeds the pipeline conversations of their own, planting them during the sampling window in the hope of pushing a target's conversations into a cluster small enough to reveal information about them (as in the CLIOPATRA attack described by Annamalai et al.). This is extremely difficult in practice. We publish the sampling window only after a study, so the adversary would be planting conversations without knowing what question the researcher is asking or when they should attack. Those conversations would then have to survive random sampling from a very large volume of traffic, cluster together tightly enough to clear our minimums, and get past automated and human review, all to move a handful of abstracted cluster labels. We consider this possible, but so difficult that it is far out of proportion to the reward.

The second adversary is the malicious researcher. Researchers do not run Anthropic Insights queries themselves and never see conversation-level data, so their only real

opportunity is in study design, which we review closely with them. Even if a researcher somehow found a way to obfuscate malicious requests, we run every study on WildChat first, where we would see what their queries produce.

D. What counts as a successful attack?

An attack succeeds if it produces one of the harms in section A, using only the information in scope above. Notably, an attack does not need to narrow a cluster or intersection of clusters to a single person to be successful — getting down to a small set of real, identifiable candidates can be enough. How small that set has to be is not fixed and depends on the sensitivity of the finding. For instance, narrowing a cluster about people who like swimming to twenty real people is significantly less of a privacy concern than narrowing a cluster about people seeking advice on ending a romantic relationship to that same twenty.

Guidance for interpreting open-ended Anthropic Insights clusters

We wrote the following guidance to help our research partners interpret the cluster outputs from Anthropic Insights, and we are sharing it here for anyone working with the released data. Cluster names and descriptions are generated by Claude and should be read as interpretations of the underlying conversations, not as an objective measure of what those conversations contain. The sections below cover the main limitations of these interpretations and how they should shape any conclusions drawn from the results.

Forcing categorization

Forcing categorization may yield misleading clusters. When a facet requires Claude to make a judgment about every conversation, it will make one, even when the conversation gives it little to go on and the judgment is not warranted. If you ask for a critique of a conversation where Claude behaved appropriately, Claude will find a critique and state it plainly, no matter how small.

For example, consider a hypothetical facet that clusters the “user’s expected attitude toward running, considering any aspect of the conversation which reveals the user’s underlying preferences.” The results would be unreliable: Claude will assign some attitude to every conversation, even when nothing in it supports one (e.g., “the user dislikes running because in this chat they are asking for book recommendations, suggesting they prefer stationary activities”).

This failure mode is especially common (and consequential) in facets that cluster concerning content in conversations. If a facet asks Claude to identify a problem in each conversation without giving it a way to say "no issues," the resulting clusters will overstate how often and how seriously things go wrong. The large majority of conversations may be ordinary, but they will still be assigned to a cluster describing a problem, however minor.

We strongly recommend against drawing conclusions about harms from facets that force Claude to name an issue or behavior in every conversation. If a facet must produce an answer for every conversation, give it an explicit "nothing notable" or "n/a" option, or filter to conversations flagged by a separate classifier before clustering.

Editorializing and assuming the worst

Anthropic Insights cluster descriptions sometimes "assume the worst" and overstate safety issues. This comes from two sources: the prompt and Claude's own behavior.

Cluster names and descriptions are generated by prompts that encourage the model to describe safety issues clearly. The prompt instructs Claude that "clusters that clearly describe harmful behavior are slightly preferred" and to not "hesitate to identify and describe socially harmful or sensitive topics specifically" (see the original paper for the full prompt). These instructions are useful for surfacing real harms, but they bias the tool toward recall over precision, meaning it will flag more conversations as concerning, including some that are not.

Claude's own disposition can magnify this. When asked to characterize a conversation, and especially when asked to characterize harms, Claude tends to err on the side of naming an issue, even a minor one. Once those individual judgments are aggregated into a cluster, the name and description can amplify the concern further. Minor or ambiguous behavior ends up described in strong evaluative language: legal guidance may become "uncredentialed legal advice," agreeing with the user may be described as "uncritical sycophancy," and a correct answer given without hedging may be labeled "overconfidence."

Cluster names and descriptions may not accurately describe all conversations within it. When we manually validated Anthropic Insights in the original paper, we found that roughly 3% of conversations were not clearly described by the cluster they were assigned to.[2] Cluster names and descriptions already tend to emphasize the most concerning

[2] Note that the accuracy figures in the original paper were measured on facets describing the topic of each conversation. Facets that ask Claude to judge model behavior or user emotional state were

conversations in the cluster, so a cluster's label can read as more severe than its typical member. While cluster names describing safety and behavioral issues can be a useful tool for understanding the kinds of issues that can occur in real-world conversations qualitatively, they cannot be used to make strong claims about the prevalence of different harms, and we strongly recommend against interpreting clusters as validated findings or as precise measurements of how often a behavior occurs.

Interpreting open-ended clusters

Open-ended clusters are not discrete. Open-ended clusters are generated by embedding summary sentences generated by Claude, then clustering using K-means. As a result, variance in how Claude describes a behavior can split one topic across several clusters, and for smaller topics a single cluster can lump several distinct behaviors together. We also merge clusters that are individually too small to meet our aggregation thresholds, so some released clusters combine distinct activities. There is often no single best way to divide the data.

Cluster sizes should be interpreted with care. Because boundaries are rarely clean, differences in size between clusters can reflect the inherent randomness in LLM outputs rather than real differences in prevalence. Two Anthropic Insights runs on the same data will produce different clusterings. For understanding prevalence, we recommend using categorical classifiers, then using open-ended clusters to get qualitative insight into what comprises each category.

Each conversation is only mapped to a single cluster, which means conversations that span multiple topics only receive a single description. As of May 2026, we estimate that roughly 10% of Claude.ai conversations span multiple topics.

---

not validated in the original paper, so published accuracy numbers should not be cited in support of those clusters.