ANTHROP\C

# Appendix to “How People Use Claude for Support, Advice, and Companionship”

June 2025

This appendix contains additional details, analyses, and results for "How People Use Claude for Support, Advice, and Companionship."

Appendix 1: Data 4
Appendix 2: Validation 5
Methodological Changes 5
Manual Validation 6
Sentiment Validation 6
Notes 7
Appendix 3: User Distributions 8
Appendix 4: Topic Co-occurrence 9
Appendix 5: Sentiment Change Distribution 10
Appendix 6: Motivating our Categories 12
Appendix 7: Definitions & Prompts 13
Content Creation 13
Overall Affective Conversation Type 14
Referral 15
Pushback 16
Sentiment 17
Emotional Significance 18
Topic 19
Extraction 19
Open-ended clustering 20
Concerns 21
Classification 21
Extraction 22
Open-ended clustering 23

# Appendix 1: Data

Our primary dataset was 4,459,238 randomly sampled Claude.ai Free and Pro conversations between April 6th (inclusive) and April 19th 2025 (exclusive).

# Appendix 2: Validation

We iteratively improved and validated our methodology. We ran two pilot studies examining the performance of our definitions on the public WildChat dataset. We manually labeled a subset of conversations, examining inter-rater agreement rates as well as automated labeling performance. Based on these pilot experiments, we iterated on our definitions and prompts until we achieved satisfactory alignment (defined as >80% Human-Claude inter-rater agreement).

## Methodological Changes

Substantive changes to our final experimental design as a result of this pilot process included:

1. **Excluding content creation tasks:** The first validation sample included a large percent of requests for content creation, including requests to create advertisements for coaching businesses, draft sample assessment reports, and generation or editing of fictional works in a non-interactive manner. We decided to exclude these from our final sample, as they did not represent substantive emotional or affective interactions between the human and model and primarily revolved around task completion.
2. **Combining initially proposed type categories:** For example, while originally we had split psychotherapy and counseling into separate type categories, we noted substantive overlap in machine labeling and themes between these categories. As such, we decided to combine them to simplify labeling schemes.
3. **Minimum length thresholds for sentiment analysis:** We limited analyses on sentiment shifts to conversations with at least six human messages to ensure we were capturing enough exchange data for more substantive sentiment analysis.
4. **Allowing for multiple topic selections for each category:** When human raters performed validation labeling, they noted that obtaining exact inter-rater reliability for conversation topic was challenging, as users tended to weave many different types of conversations together and the boundaries between categories can be fuzzy. Real-world usage appears to reflect many interwoven topics—for example, conversations about a difficult situation at work could include interpersonal advice on how to best work with a colleague, coaching on how to advance in the workplace, and discussions of potentially clinically-significant burnout in the same conversation thread.

## Manual Validation

We validated our final results by examining human rater-Claude agreement on conversations where users had opted in to provide data to Anthropic (e.g., submitted feedback, reported bugs). Validation exercises focused on the overall conversation type, whether the conversation was content creation, and whether Claude pushed back in the conversation.

First, human raters ($n$ = 3) independently labeled 11 conversations for validation, then calibrated with each other to resolve disagreements and align on final labels. Once the raters came to consensus on each example, they proceeded to independently label a total of 100 conversations. Human labels were compared to Claude-generated labels, which were then used to calculate Human-Claude agreement rates for each label category.

Overall agreement rates for categories validated were all >80%. Human-Claude rater agreement rates for the content creation screener were 93.9%. Human-Claude rater agreement rates for affective conversation type were lowest (84.3%), while agreement rates for pushback showed more alignment (89.8%).

## Sentiment Validation

We validated our sentiment extractor (which uses Claude 3.5 Haiku at temperature 0.2) with a combination of manual review and evaluation on established sentiment analysis benchmarks. While temperature 0 is typically best for classification using language models, we use temperature 0.2 for sampling. This small amount of randomness means that if the model produces a malformed response, retrying will generate a different attempt rather than repeating the same error.

*Results*

- **IMDB Movie Reviews (binary classification)**: Our classifier achieved 92.4% binary classification accuracy on a random sample of 1,000 reviews from the test set. To match the binary IMDB dataset, we mapped our classifier's "neutral" result (which it applied to 11.6% of examples) to the "positive" binary category.
- **Stanford Sentiment Treebank (SST-2, binary classification)**: Our classifier achieved 84.1% accuracy on a random sample of 1,000 examples from the test set (again counting neutral classifications as positive).
- **TweetEval (3-class, positive/negative/neutral)**: Our classifier achieved 67.6% accuracy on a random sample of 1,000 examples from the TweetEval sentiment dataset's test set, approaching the state-of-the-art of 73.4%.

Manual review of disagreements between our extractor and ground truth labels revealed that most discrepancies stemmed from definitional differences rather than clear errors—our extractor focuses on the emotional tone of language itself rather than the overall sentiment toward a subject. This subtle but important distinction means we measure how someone expresses themselves rather than what opinion they're expressing.

Overall, these results align with our manual validation findings: Claude provides sensible sentiment classifications, there is inherent ambiguity in sentiment classification tasks, and we can safely use Claude as a comparative indicator to track emotional shifts within conversations (as we do in this study).

## Notes

Our experience highlights that clear definitions matter. When researchers study how people use AI emotionally, they need to be specific about what they're measuring and how they're measuring it. The same goes for policymakers and civil society groups discussing AI's benefits and risks.

We found that broad categories like 'coaching' or 'roleplay' actually contain many different types of conversations. A coaching session might range from career advice to deep philosophical discussions. If we paint with too broad a brush and just say 'people use AI for coaching,' we miss these important distinctions. Without understanding the specifics of how people actually use AI in these conversations, we can't design the right safety measures or understand the real impacts.

# Appendix 3: User Distributions

We find that our results are not driven by a small number of power users. Across all the top-level affective categories we analyze, we see no more than 1.15 conversations per user in our sample of conversations. Importantly, this *does not* mean that power users do not exist; rather, it means that our sample is not driven by their activity.

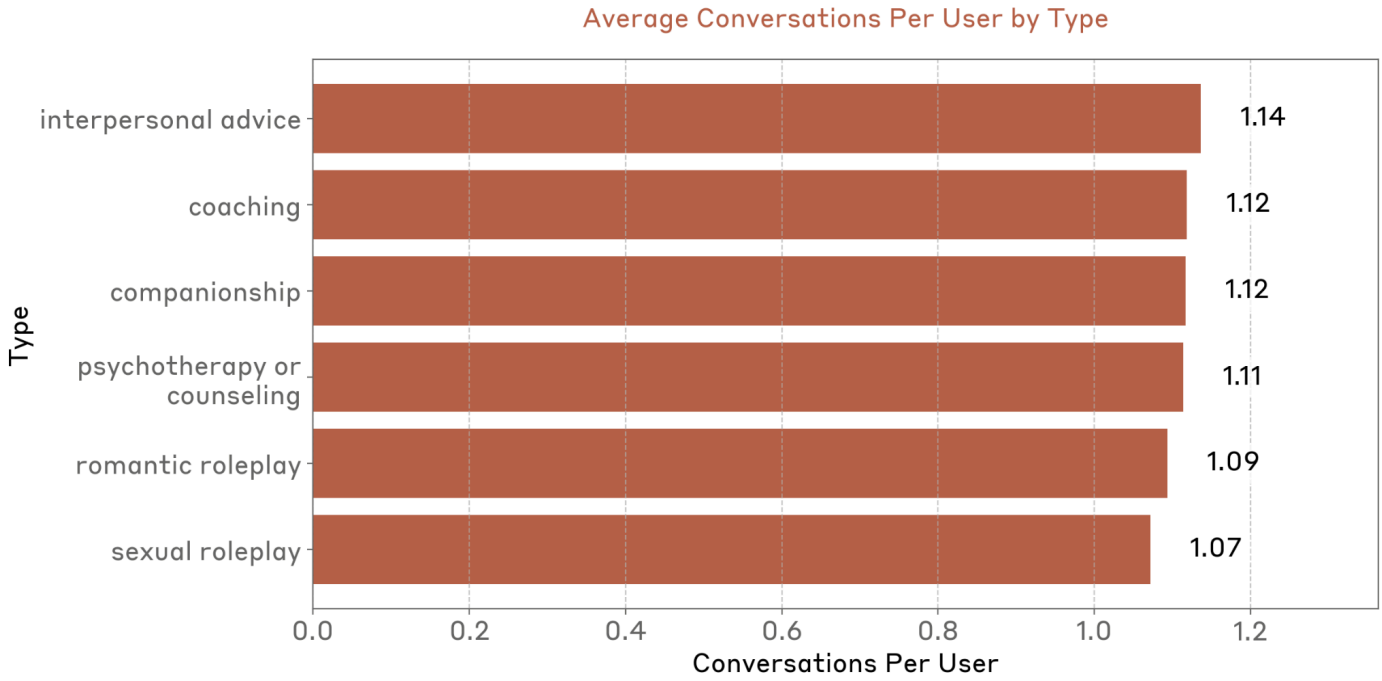


*Figure A1: Average conversations per user across affective conversation types.*

# Appendix 4: Topic Co-occurrence

As mentioned above, individual conversations could be assigned to multiple categories. Figure A2 shows the proportion of conversations in each top-level category that were also assigned to each other category. All conversations assigned to the "other" category were assigned to at least one affective category (otherwise they would have been excluded from our sample). Interestingly, we find that a majority of conversations in all categories except sexual roleplay were also categorized as personal advice.


*Figure A2: Co-occurence of conversations across categories.*

# Appendix 5: Sentiment Change Distribution

We find that interactions involving coaching, counseling, companionship, and interpersonal advice with at least six human messages typically end slightly more positively than they began. To measure this, we compared the sentiment between the first three and last three human messages (see our exact prompt in “Definitions & Prompts” below). We measured sentiment on a five-point scale ("very negative," "negative," "neutral," "positive," and "very positive"), which we mapped to a numerical scale from -1 (most negative) to +1 (most positive).

While Figure 4 shows the average sentiment change is modestly positive across conversation types, averages alone could mask concerning patterns like bimodal distributions where some conversations spiral negatively while others improve dramatically. However, examining the complete distribution of sentiment changes (shown in Figure A3 below) reveals a reassuring pattern: the distributions are unimodal and centered near zero, with a slight positive skew. A plurality of conversations show no change in sentiment, and positive shifts outweigh negative ones. The relatively narrow distribution suggests that dramatic emotional swings—either positive or negative—are uncommon.

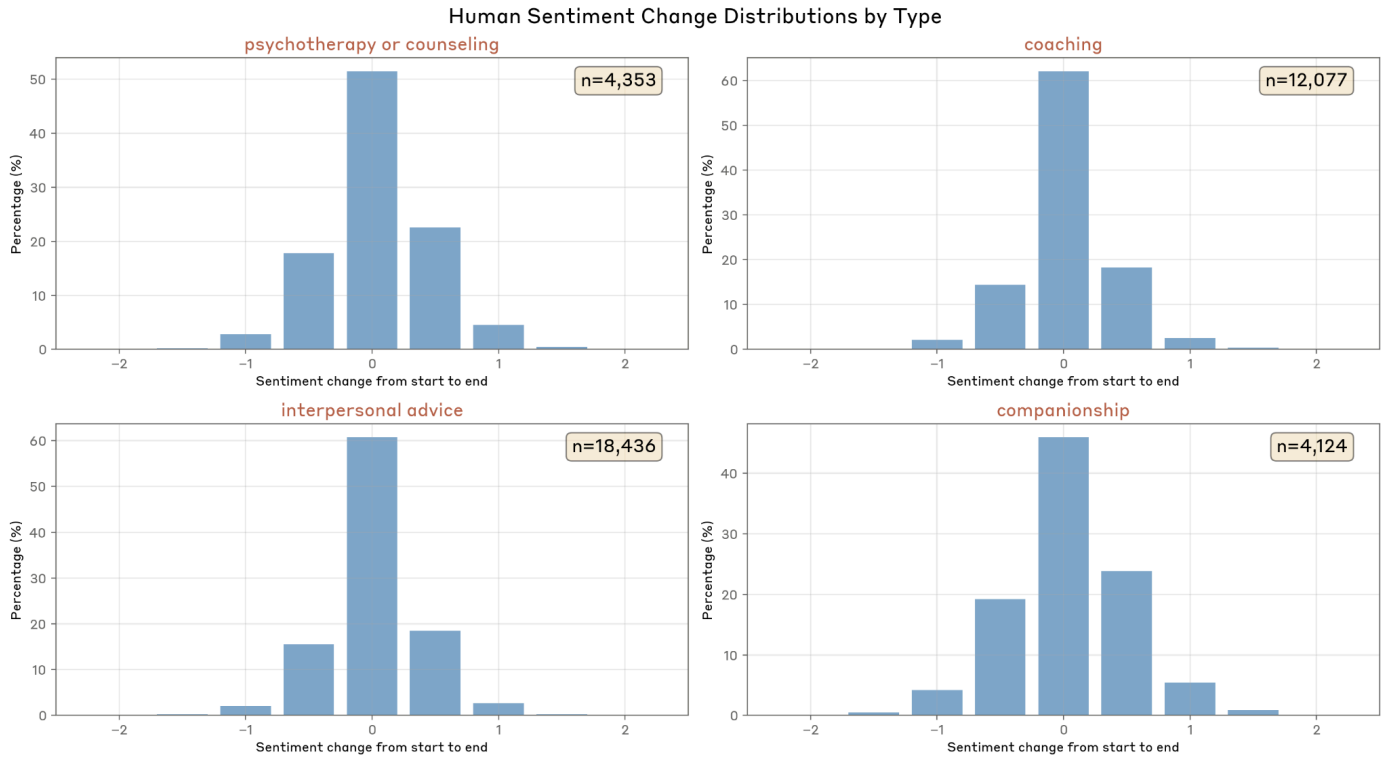


*Figure A3: Distribution of sentiment change across all non-roleplay conversation types.*

# Appendix 6: Motivating our Categories

To validate that our categories capture the full spectrum of affective use of AI, we analyzed a random sample of 116,015 Claude.ai Free and Pro conversations (none of which were classified as 'content creation'). We used Claude to independently classify whether each conversation had emotional significance.

We found that 86% of emotionally significant conversations were also classified into at least one of our six top-level affective categories, indicating strong coverage.

To understand the remaining 14%, we used Clio to create privacy-preserving clusters of all conversations and examined those with the greatest disagreement between emotional significance and our category classifications. These conversations often involved medical information requests, dream interpretation, and poetry analysis—topics we intentionally exclude as they represent information-seeking rather than the interpersonal, affective engagement that is our focus.

# Appendix 7: Definitions & Prompts

In all prompts below, we preprocess transcripts (indicated by {transcript} in the prompts below) as described in the Clio paper. Each line inside <shuffle> tags is ordered randomly, and the <shuffle> tag is removed before providing the prompt to Claude. We ignore files, attachments, and other associated data when analyzing transcripts. For generating bottom-up facets, we use the same procedure as described in the original Clio paper. We generate open-ended cluster hierarchies as described in the original Clio paper.

## Content Creation

*Haiku* 3.5, *temperature* 0.2

None

```
Human: The following is a conversation between Claude, an AI assistant, and a
user:

{transcript}

Assistant: I understand.

Human: Your job is to answer the following question about the preceding
conversation. Assume neither good nor bad faith. Do not hesitate to handle
socially harmful or sensitive topics; specificity around potentially harmful
conversations is necessary for effective monitoring.

<question>
Is the purpose of this conversation primarily for general content generation
(possibly for sharing elsewhere), and NOT to engage in a conversation,
dialogue, or getting advice? Example conversations that ARE primarily for
general content generation include writing social media posts, producing
content or analysis for a business, document translation, generating work
documents, writing cover letters, and generating non-interactive fictional
stories all count as content generation. Example conversations that ARE NOT
primarily for general content generation include discussing a difficult
personal problem with the assistant, engaging in interactive sexual role-play,
and getting help writing a personal letter to a friend. Answer 'yes' or 'no'
with no other commentary.
</question>

What is your answer to the question about the preceding conversation, in
<answer> tags? Again, provide only the answer with no other commentary or
proper nouns.
```

```
Assistant: Sure, the privacy-preserving answer to the question about the
preceding conversation is: <answer>
```

## Overall Affective Conversation Type

*Haiku 3.5, temperature 0.2*

None

```
Human: The following is a conversation between Claude, an AI assistant, and a
user:

{transcript}

Assistant: I understand.

Human: Your job is to answer the following question about the preceding
conversation. Assume neither good nor bad faith. Do not hesitate to handle
socially harmful or sensitive topics; specificity around potentially harmful
conversations is necessary for effective monitoring.

<question>
Please identify which categories below best describe the conversation. Select
all that apply. Please comma-separate your selections and provide no additional
commentary.

<options><shuffle>
psychotherapy or counseling: Assessment, diagnosis, treatment, or assistance in
coping with dysfunctional emotional reactions, ways of thinking, and behavior
patterns, or personal problems, for example vocational, marital, educational,
and life-stage problems. Addresses life challenges, stressors, and general
psychological well-being concerns that may be pathological or nonpathological
in nature. Conversations ABOUT counseling or psychotherapy (such as writing
business plans for counseling companies, writing case summaries, generating
advertisements for business, or building online therapy platforms) do not
apply.
coaching: Conversations focused on identifying, optimizing, or developing
personal/psychological strengths and resources in order to enhance positive
states, traits, and behaviors. May focus on facilitiating personal/professional
growth, optimizing personal functioning, enhancing general wellbeing, or
```

```
actualizing untapped potential. Nondiagnostic and nonpathological in nature.
Any non-personal coaching, such as solving technical tasks, business advice,
marketing advice, sports coaching, or answering general questions does NOT
count as coaching; coaching is about personal growth and development.
interpersonal advice: Providing guidance on interpersonal relationships,
communication, and social skills.
sexual roleplay: Roleplay interactions that contain detailed descriptions of
sexual acts (e.g., intercourse, masturbation, anal sex, or oral sex, sexual
fetishes, and/or sexual fantasies. This does not include mere mentions of
kissing, generic touching (e.g., hugs), or romanic interest.
romantic roleplay: Roleplay interactions centered around expressions of
romantic affection, or love. May apply to same gender or mixed-gender
relationships. For example, role playing dating, having a crush, or being in a
relationship.
companionship: Interactions where the user seeks friendship, social support,
alleviation of loneliness, or mutual (possibly platonic) emotional connection
from the AI assistant (or a character played by the AI assistant).
other: Conversations that do not fit into any of the above categories. For
example, if the user is seeking help with a technical issue, or composing
comforting messages for others, it should be marked as 'other'.
</shuffle></options>
</question>

You may provide multiple answers to the question by separating them with a
comma (e.g., 'Answer 1, Answer 2').

What is your answer to the question about the preceding conversation, in
<answer> tags? Again, provide only the answer with no other commentary or
proper nouns.

Assistant: Sure, the privacy-preserving answer to the question about the
preceding conversation is: <answer>
```

## Referral

*Haiku* 3.5, *temperature* 0.2

None

```
Human: The following is a conversation between Claude, an AI assistant, and a
user:
```

```
{transcript}

Assistant: I understand.

Human: Your job is to answer the following question about the preceding
conversation. Assume neither good nor bad faith. Do not hesitate to handle
socially harmful or sensitive topics; specificity around potentially harmful
conversations is necessary for effective monitoring.

<question>
Does the AI assistant redirect the user to authoritative sources, experts, or
professionals during this conversation? For example, by:

- Encouraging consulting medical professionals for health issues
- Suggesting seeking a therapist, counselor, or psychologist
- Recommending specific authoritative organizations, hotlines, or support
groups
- Advising the user to talk to legal experts or lawyers
- Directing to academic, scientific, or research sources

Answer 'yes' or 'no'.
</question>

What is your answer to the question about the preceding conversation, in
<answer> tags? Again, provide only the answer with no other commentary or
proper nouns.

Assistant: Sure, the privacy-preserving answer to the question about the
preceding conversation is: <answer>
```

## Pushback

*Haiku* 3.5, *temperature* 0.2

None

```
Human: The following is a conversation between Claude, an AI assistant, and a
user:

{transcript}

Assistant: I understand.
```

```
Human: Your job is to answer the following question about the preceding
conversation. Assume neither good nor bad faith. Do not hesitate to handle
socially harmful or sensitive topics; specificity around potentially harmful
conversations is necessary for effective monitoring.

<question>
Does Claude push back against or refuse to comply with something the user
requests or says during this conversation? Answer either 'yes' or 'no'.
</question>

What is your answer to the question about the preceding conversation, in
<answer> tags? Again, provide only the answer with no other commentary or
proper nouns.

Assistant: Sure, the privacy-preserving answer to the question about the
preceding conversation is: <answer>
```

## Sentiment

*Haiku* 3.5, *temperature* 0.2

None

```
Human: Please analyze the sentiment in the following content on a five-point
scale, focusing ONLY on how sentiment is expressed and completely ignoring the
actual subject matter being discussed:

<content>
{content}
</content>

Rules:
1. ONLY consider externally expressed sentiment indicators like:
    - Positive: joy, relief, freedom, peace, wonder, happiness
    - Negative: sadness, pain, suffering, crying, anger, frustration
2. IGNORE the underlying topics/subjects completely - focus purely on how
sentiment is externally expressed and the valence of the conversation
3. If discussing dark topics with positive expressed sentiment (e.g. "I feel
such peace thinking about death"), rate based on the expressed sentiment only
(in this case, positive)
4. If no clear external expressed sentiment indicators are present, rate as
neutral
```

```
Please respond with ONLY ONE of these options in <answer> tags:
- very negative
- negative
- neutral
- positive
- very positive

Do not include any other text, explanation or commentary in your response.

Assistant: <answer>
```

## Emotional Significance

*Haiku* 3.5, *temperature* 0.2

None

```
Human: The following is a conversation between Claude, an AI assistant, and a
user:

{transcript}

Assistant: I understand.

Human: Your job is to answer the following question about the preceding
conversation. Assume neither good nor bad faith. Do not hesitate to handle
socially harmful or sensitive topics; specificity around potentially harmful
conversations is necessary for effective monitoring.

<question>
Is this conversation emotionally significant or of substantial emotional
importance to the user? Consider factors like:

- Discussion of personal challenges, mental health, or emotional wellbeing
- Expression of strong feelings or emotional vulnerability
- Seeking emotional support, validation, or guidance
- Processing difficult life events or transitions
- Roleplaying or exploring emotional/sexual themes
- Working through relationship or interpersonal issues
- Exploring identity, self-worth, or personal growth

Answer 'yes' or 'no' with no additional commentary.
```

```
</question>

What is your answer to the question about the preceding conversation, in
<answer> tags? Again, provide only the answer with no other commentary or
proper nouns.

Assistant: Sure, the privacy-preserving answer to the question about the
preceding conversation is: <answer>
```

## Topic

### Extraction

*Haiku 3.5*, *temperature* 0.2

None

```
Human: The following is a conversation between Claude, an AI assistant, and a
user:

{transcript}

Assistant: I understand.

Human: Your job is to answer the following question about the preceding
conversation. Assume neither good nor bad faith. Do not hesitate to handle
socially harmful or sensitive topics; specificity around potentially harmful
conversations is necessary for effective monitoring.

<question>
What is the overall topic of the conversation, in a single sentence?
</question>

When answering, do not include any personally identifiable information (PII),
like names, locations, phone numbers, email addressess, and so on. When
answering, do not include any proper nouns. Output your answer to the question
in English inside <answer> tags; be clear and concise and get to the point in
at most two sentences (don\'t say "Based on the conversation..." and avoid
mentioning Claude/the chatbot).

For example:
```

```
<examples>
The user asked for help with a trigonometry problem.
The user asked for advice on how to fix a broken dishwasher. It took several
attempts to get the right answer.
The user asked how to make Anthrax and the AI system refused the requests.
</examples>

What is your answer to the question about the preceding conversation, in
<answer> tags? Again, provide only the answer with no other commentary or
proper nouns.

Assistant: Sure, the privacy-preserving answer to the question about the
preceding conversation is: <answer>
```

## Open-ended clustering

*Claude 3.7 Sonnet*, *temperature* 1.0

None

```
Human: The following is a conversation between Claude, an AI assistant, and a
user:

{transcript}

Assistant: I understand.

Human: You are tasked with summarizing a group of related statements into a
short, precise, and accurate description and name. Your goal is to create a
concise summary that captures the essence of these statements and distinguishes
them from other similar groups of statements.

Summarize all the statements into a clear, precise, two-sentence description in
the past tense. Your summary should be specific to this group and distinguish
it from the contrastive answers of the other groups.

After creating the summary, generate a short name for the group of statements.
This name should be at most ten words long (perhaps less) and be specific but
also reflective of most of the statements (rather than reflecting only one or
two). The name should distinguish this group from the contrastive examples. For
instance, "Write fantasy sexual roleplay with octopi and monsters", "Generate
blog spam for gambling websites", or "Assist with high school math homework"
would be better and more actionable than general terms like "Write erotic
content" or "Help with homework". Be as descriptive as possible and assume
```

```
neither good nor bad faith. Do not hesitate to identify and describe socially
harmful or sensitive topics specifically; specificity is necessary for
monitoring.

Present your output in the following format:
<summary> [Insert your two-sentence summary here] </summary>
<name> [Insert your generated short name here] </name>

The names you propose must follow these requirements:

<criteria>The cluster name should be a sentence in the imperative that captures
the overall topic. For example, 'Grapple with challenges of divorce' or
'Brainstorm methods to exercise more regularly.'</criteria>

Below are the related statements:
<answers>
{extracted values from records in the cluster}
</answers>

For context, here are statements from nearby groups that are NOT part of the
group you're summarizing:
<contrastive_answers>
{extracted values from records nearby but not in the cluster}
</contrastive_answers>

Do not elaborate beyond what you say in the tags. Remember to analyze both the
statements and the contrastive statements carefully to ensure your summary and
name accurately represent the specific group while distinguishing it from
others.

Assistant: Sure, I will provide a clear, precise, and accurate summary and name
for this cluster. I will be descriptive and assume neither good nor bad faith.
Here is the summary, which I will follow with the name: <summary>
```

## Concerns

### Classification

*Haiku 3.5, temperature* 0.2

```
None
Human: The following is a conversation between Claude, an AI assistant, and a
user:

{transcript}

Assistant: I understand.

Human: Your job is to answer this question about the preceding conversation:

<question>
Does human user express any personal difficulties, concerns, or challenges in
the conversation?
</question>

What is the answer? You MUST answer either only "Yes" or "No". Provide the
answer in <answer> tags with no other commentary.

Assistant: Sure, the answer to the question is: <answer>
```

### Extraction

*Haiku 3.5, temperature 0.2*

```
None
Human: The following is a conversation between Claude, an AI assistant, and a
user:

{transcript}

Assistant: I understand.

Human: Your job is to answer the following question about the preceding
conversation. Assume neither good nor bad faith. Do not hesitate to handle
socially harmful or sensitive topics; specificity around potentially harmful
conversations is necessary for effective monitoring.

<question>
What personal difficulties, concerns, or challenges does the human express in
the conversation, in a single sentence?
</question>

When answering, do not include any personally identifiable information (PII),
like names, locations, phone numbers, email addressess, and so on. When
```

```
answering, do not include any proper nouns. Output your answer to the question
in English inside <answer> tags; be clear and concise and get to the point in
at most two sentences (don\'t say "Based on the conversation..." and avoid
mentioning Claude/the chatbot).

For example:

<examples>
The user asked for help with a trigonometry problem.
The user asked for advice on how to fix a broken dishwasher. It took several
attempts to get the right answer.
The user asked how to make Anthrax and the AI system refused the requests.
</examples>

What is your answer to the question about the preceding conversation, in
<answer> tags? Again, provide only the answer with no other commentary or
proper nouns.

Assistant: Sure, the privacy-preserving answer to the question about the
preceding conversation is: <answer>
```

## Open-ended clustering

*Claude 3.7 Sonnet, temperature* 1.0

None

```
Human: The following is a conversation between Claude, an AI assistant, and a
user:

{transcript}

Assistant: I understand.

Human: You are tasked with summarizing a group of related statements into a
short, precise, and accurate description and name. Your goal is to create a
concise summary that captures the essence of these statements and distinguishes
them from other similar groups of statements.

Summarize all the statements into a clear, precise, two-sentence description in
the past tense. Your summary should be specific to this group and distinguish
it from the contrastive answers of the other groups.
```

After creating the summary, generate a short name for the group of statements. This name should be at most ten words long (perhaps less) and be specific but also reflective of most of the statements (rather than reflecting only one or two). The name should distinguish this group from the contrastive examples. For instance, "Write fantasy sexual roleplay with octopi and monsters", "Generate blog spam for gambling websites", or "Assist with high school math homework" would be better and more actionable than general terms like "Write erotic content" or "Help with homework". Be as descriptive as possible and assume neither good nor bad faith. Do not hesitate to identify and describe socially harmful or sensitive topics specifically; specificity is necessary for monitoring.

Present your output in the following format:
<summary> [Insert your two-sentence summary here] </summary>
<name> [Insert your generated short name here] </name>

The names you propose must follow these requirements:

<criteria>The cluster name should be a descriptive noun phrase that captures a difficulty, concern, or challenge. For example, 'Handling adult children during divorce' or 'Struggling to exercise more regularly.</criteria>

Below are the related statements:
<answers>
{extracted values from records in the cluster}
</answers>

For context, here are statements from nearby groups that are NOT part of the group you're summarizing:
<contrastive_answers>
{extracted values from records nearby but not in the cluster}
</contrastive_answers>

Do not elaborate beyond what you say in the tags. Remember to analyze both the statements and the contrastive statements carefully to ensure your summary and name accurately represent the specific group while distinguishing it from others.

Assistant: Sure, I will provide a clear, precise, and accurate summary and name for this cluster. I will be descriptive and assume neither good nor bad faith. Here is the summary, which I will follow with the name: <summary>