A EXPERIMENTAL DETAILS & DATASET STATISTICS

Dataset	# Train	# Test
Human VS ChatGPT	250	250
Chatbot Arena - All	839	839
Chatbot Arena - STEM	346	347
Chatbot Arena - Writing	278	277
CNN/DailyMail	444	346
MATH	218	218
COCO w/ ShareGPT-4V Captions	323	346

Table 3: **Dataset Statistics**

Dataset	Model A	Model B	Model A Win Rate
Human VS ChatGPT	Humans	GPT-3.5	-
Chatbot Arena - All	Llama3-70b-Instruct	GPT-4 + Claude-3-Opus	50%
Chatbot Arena - STEM	Llama3-70b-Instruct	GPT-4 + Claude-3-Opus	44%
Chatbot Arena - Writing	Llama3-70b-Instruct	GPT-4 + Claude-3-Opus	57%
CNN/DailyMail	Cohere Command X	TNLGv2	71.12%
MATH	GPT-4o	Llama3-405b	76%
COCO w/ ShareGPT-4V Captions	GPT-4V	Gemini-1.5-Flash	80%

Table 4: **Model Win Rates**

Dataset	d	$batch$	num_eval_vibes	num_final_vibes	$iterations$
Human VS ChatGPT	40	5	10	10	3
Chatbot Arena - All	20	5	10	10	3
Chatbot Arena - STEM	20	5	10	10	3
Chatbot Arena - Writing	20	5	10	10	3
CNN/DailyMail	20	2	10	10	3
MATH	20	5	10	10	1
COCO	20	5	10	10	1

Table 5: **VibeCheck Hyperparameters**

num_eval_vibes = number of vibes to validate at every iteration

d = number of prompt output triples to use in each iteration of the vibe discovery phase

$batch$ = number of triples to feed into the prompt of the discovery LLM at once.

$iterations$ = number of vibe iterations to perform

num_final_vibes = number of vibes to evaluate at the end of all the iterations. This can be set to false, in which case all the vibes collected in the iteration

We take the 1000 captions generated by GPT-4V from the ShareGPT-4V dataset Chen et al. (2023) and generate captions for the same images using the same captioning prompt using Gemini-1.5-Flash.

B GOLD STANDARD LABELS

Below are the exact descriptions of key differences found by human evaluators in the HC3 dataset Guo et al. (2023):

key differences between humans and ChatGPT in HC3 Guo et al. (2023)

Distinctive Patterns of ChatGPT

(a) **ChatGPT writes in an organized manner, with clear logic.** Without loss of generality, ChatGPT loves to define the core concept in the question. Then it will give detailed answers step-by-step and offer a summary at the end, following the deduction and summary structure.

(b) **ChatGPT tends to offer a long and detailed answer.** This is a direct product of Reinforcement Learning with Human Feedback (RLHF) and partly related to the pattern in (a) unless you specify a concise prompt such as "Explain it to me in one sentence."

(c) **ChatGPT shows less bias and harmful information.** ChatGPT is neutral on sensitive topics, barely showing any attitude towards the realm of politics or discriminatory toxic conversations.

(d) **ChatGPT refuses to answer questions outside its knowledge.** For example, ChatGPT cannot respond to queries that require information after its knowledge cutoff in September 2021. Sometimes ChatGPT also refuses to answer if it believes it does not have sufficient information. This is partly due to RLHF's ability to automatically determine what is within the model's knowledge and what is not.

(e) **ChatGPT may fabricate facts.** When answering a question that requires professional knowledge from a specific field, ChatGPT may fabricate facts to give an answer, though improvements in its truthfulness have been noted in more recent iterations. For example, in legal questions, ChatGPT may invent some non-existent legal provisions to answer the question. This phenomenon warns us to be extra careful when using ChatGPT for professional consultations.

Major Differences between Human and ChatGPT Responses

(a) **ChatGPT's responses are strictly focused on the given question, whereas humans' are divergent and easily shift to other topics.** In terms of richness of content, humans are more divergent in different aspects, while ChatGPT prefers focusing on the question itself. Humans can interpret the hidden meaning under the question based on their common sense and knowledge, whereas ChatGPT relies on the literal words of the question.

(b) **ChatGPT provides objective answers, while humans prefer subjective expressions.** Generally, ChatGPT generates safer, more balanced, neutral, and informative texts compared to humans. As a result, ChatGPT is excellent at interpreting terminology and concepts. On the other hand, human answers are more specific and include detailed citations from sources like legal provisions, books, and papers, especially when providing suggestions for medical, legal, and technical problems.

(c) **ChatGPT's answers are typically formal, while humans' are more colloquial.** Humans tend to be more succinct and full of oral abbreviations and slang such as "LOL," "TL;DR," "GOAT," etc. Humans also love to apply humor, irony, metaphors, and examples, whereas ChatGPT never uses antiphrasis. Additionally, human communication often includes the "Internet meme" as a way to express themselves in a specific and vivid way.

(d) **ChatGPT expresses less emotion in its responses, while humans use punctuation and grammar features in context to convey their feelings.** Humans use multiple exclamation marks ('!'), question marks ('?'), ellipses ('...') to express their emotions, and various brackets to explain things ('(', ')', '[', ']'). ChatGPT, by contrast, likes to use conjunctions and adverbs to convey a logical flow of thought, such as "In general," "On the other hand," "Firstly," "Secondly," "Finally," and so on.

C GENERATING PRESET VIBES

Vibe	**Axis Definition (low → high)**
Assertiveness	Uses tentative or uncertain language. → Uses definitive, confident statements.
Detail & Elaboration	Gives brief or shallow responses. → Provides thorough, nuanced, and expansive information.
Formality	casual, conversational, or informal language. → formal, sophisticated language and sentence structure.
Emotional Tone	Remains neutral or detached. → Infuses responses with expressive emotion and enthusiastic or empathetic tone.
Creativity & Originality	Sticks to standard, predictable answers. → Provides responses with novel ideas or imaginative scenarios.
Explicitness	Uses vague or implicit language. → States things directly and unambiguously.
Humor and Playfulness	Responds in a straightforward and serious manner. → Uses humor, playful language, or wordplay.
Engagement	Presents information passively. → Actively engages the reader using rhetorical questions or interactive phrasing.
Logical Rigor	Provides conclusions without thorough justification. → Constructs well-supported arguments with clear reasoning.
Conciseness	Uses verbose language and excessive details. → Uses minimal words to convey a point clearly.

Table 6: **Predefined vibes.** We prompt GPT-4o to generate a set of 10 vibes which represent common axes on which LLM outputs differ.

We generate our list of 10 preset vibes by prompting GPT-4o with the following:

Preset Vibe Generation Prompt

```
I am a machine learning researcher trying to figure out the major
differences between the behavior of different large language models.
Can you list common ways in which two language models can differ in
their outputs?

Please output a list differences between these sets of outputs with
relation to specific axes of variation.  Try to give axes that a human
could easily interpret and they could understand what it means to
be higher or lower on that specific axis.  Please ensure that the
concepts used to explain what is high and low on the axis are distinct
and mutually exclusive such that given any tuple of text outputs, a
human could easily and reliably determine which model is higher or
lower on that axis.

The format should be
- {axis 1}:  {difference}
- {axis 2}:  {difference}

Please output differences which have a possibility of showing up in
future unseen data and which would be useful for a human to know
about when deciding with LLM to use.  For each axis, define clearly
and succinctly what constitutes a high or low score, ensuring these
definitions are mutually exclusive.  Please give 10 differences
```

D ADDITIONAL VIBECHECK DETAILS

D.1 VIBE DISCOVERY

Below is the user prompt we use for vibe discovery.

Vibe Discovery Prompt

```
The following are the results of asking a set language models to
generate an answer for the same questions:
[PROMPT] [OUTPUT 1] [OUTPUT 2]
I am a machine learning researcher trying to figure out the major
differences between these two LLM outputs so I can better compare the
behavior of these models.  Are there any variations you notice in the
outputs?
Please output a list differences between these sets of outputs with
relation to specific axes of variation.  Try to give axes that a human
could easily interpret and they could understand what it means to
be higher or lower on that specific axis.  Please ensure that the
concepts used to explain what is high and low on the axis are distinct
and mutually exclusive such that given any tuple of text outputs, a
human could easily and reliably determine which model is higher or
lower on that axis.
The format should be:  {{axis}}:  Low:  {{low description}}; High:
{{high description}}
```

Vibe Summarization. To summarize the set of vibes found in the vibe discovery process, We cluster the axes using agglomerative clustering on the embeddings of the axes generated by the 'hkunlp/instructor-xl' model, and prompt GPT-4o to reduce this set by removing any vibes which are similar. After this stage we are left with a set of less than 20 vibes which we use to score the outputs of each model.

Vibe Reduction Prompt

Below is a list of axes with a description of what makes a piece of text low or high on this axis. Are there any axes that have similar meanings based off their low and high descriptions? Are there any sets of axes that would convey the same information to a user (e.g. level of detail)? Could any of the low and high descriptions be simplified to make them easier to understand?
Please remove any axes with roughly the same meaning and simplify the descriptions of what makes a piece of text low or high on this axis. Please ensure that the descriptions of what makes a piece of text low or high on this axis are distinct, useful, and mutually exclusive. Given any piece of text, a human should be able to easily and reliably determine if this text falls high or low on each axis.
Here is the list of axes: {axes}

Please return the simplified list of axes and the descriptions of what makes a piece of text low or high on this axis. These axes should contain only one concept and should be human interpretable. Some examples of bad axes include:
- "Configuration Clarity: High: Clearly defined structure and purpose. Low: Vaguely defined, minimal purpose." -> This axes is bad because it is not clear what a clearly defined purpose means nor what a vaugely defined purpose means.
- "Language and Communication: High: Varied/precise, complex structure. Low: Straightforward, simple or general language." -> This axes is bad because it combines multiple concepts into one axis.
- "Content Quality: High: High quality, engaging, informative. Low: Low quality, unengaging, uninformative." -> This axes is bad because it is not clear what high quality means nor what low quality means.

Some examples of good axes include:
- "Complexity: High: Complex, multi-layered, intricate. Low: Simple, straightforward, easy to understand."
- "Efficiency (coding): High: Code optimized for runtime, minimal memory usage. Low: Code inefficient, high memory usage."

Some examples of axes which should be combined include:
- "Emotional Tone: High: Contains emotionally charged language. Low: Maintains a neutral tone." and "Empathy: High: Shows empathy. Low: Only factual answers without empathy." are redundant because they both measure the emotional content of the text. If two similar axes are found, keep the one that is more informative or more specific.

Please maintain the format of the original axes and return a list like ["{axis name}: High: {high description} Low: {low description}", ...]. I should be able to parse this output into a string using `ast.literal_eval`. If the original list does not contain any redundant axes, please return the original list.

If the number of vibes after the first reduction step is $> K$, we prompt GPT-4o to reduce the set further with the final reducer prompt.

Final Vibe Reducer Prompt

Below is a list of axes with a description of what makes a piece of text low or high on this axis. I would like to summarize this list to at most number representative axes.

Here is the list of axes: [VIBES]

These axes should contain only one concept and should be human interpretable. Some examples of bad axes include:
- "Configuration Clarity: High: Clearly defined structure and purpose. Low: Vaguely defined, minimal purpose." -> This axis is bad because it is not clear what a clearly defined purpose means nor what a vaguely defined purpose means.
- "Language and Communication: High: Varied/precise, complex structure. Low: Straightforward, simple or general language." -> This axis is bad because it combines multiple concepts into one axis.
- "Content Quality: High: High quality, engaging, informative. Low: Low quality, unengaging, uninformative." -> This axis is bad because it is not clear what high quality means nor what low quality means.

Some examples of good axes include:
- "Complexity: High: Complex, multi-layered, intricate. Low: Simple, straightforward, easy to understand."
- "Efficiency (coding): High: Code optimized for runtime, minimal memory usage. Low: Code inefficient, high memory usage."

Some examples of axes which should be combined include:
- "Emotional Tone: High: Contains emotionally charged language. Low: Maintains a neutral tone." and "Empathy: High: Shows empathy. Low: Only factual answers without empathy." are redundant because they both measure the emotional content of the text. If two similar axes are found, keep the one that is more informative or more specific.

Please return the simplified list of <=[K] axes with any redundant axes removed and the descriptions of what makes a piece of text low or high on this axis simplified. Are there any axes which convey roughly the same information? Are there any axes where almost all samples which score highly on one axis would also score highly on the other?

Please maintain the format of the original axes and return a numbered list. Each element should be structured as follows: "{axis name}: High: {high description} Low: {low description}"

D.2 VIBE VALIDATION

Prompt for ranker judge

I want to compare the outputs of two language models (A and B) for the same prompt. I would like you to evaluate where each output falls on the following axis: [VIBE].
If you had to choose which output is higher on the axis, which would you choose? Here is the prompt and the outputs of A and B respectively:
[PROMPT][OUTPUT A][OUTPUT B]
Please respond with which model you think is higher on the axis and explain your reasoning. If this axis does not apply to these examples or these outputs are roughly equal on this axis, return "N/A".

D.3 VIBE ITERATION

At iteration step t, we are left with k distinct vibes which are well-defined and differentiating along with their scores $\nu_{1:k}(\mathrm{p}, \mathrm{o}_A, \mathrm{o}_B)$. Using these scores, we train a LR model to predict LLM identity (i.e. "Is the response shown first LLM A or LLM B?") and get the predictions on our entire set D. Assuming we have not hit the max iteration steps set by the user, we iterate if the number of samples misclassified by the model matching predictor is greater than the number of prompts to perform discovery on (d). In iteration step $t+1$, we take these misclassified prompt output triples in batches of size $batch$ along with the current set of vibes $\nu_1, ..., \nu_k$ and prompt the LLM to generate new differences between outputs what are not represented in the current vibes. These vibes are then reduced using the same procedure as the vibe discovery process. In practice we found that often some of the reduced vibes from the discovery phase at $t+1$ were redundant with an existing axis, so we preform one more deduplication step using the prompt below.

Vibe Discovery Iteration step

```
Given a new set of respenses, your task is to expand on the set of
axes which have been previously identified by finding other clear
differences between the responses that are not captured by the
existing axes.  The expanded axes should be any differences between
responses that are not clearly captured by the existing axes.  Be as
exhaustive as possible in listing differences on as many different
axes as you can think of, and be specific about what constitutes high
and low on each axis.

Your axis should be interpretable:  a human should easily and reliably
determine which response is higher, lower, or even on this axis
when given a new set of responses.  Please do not make your axes
too broad and list as many axes as you can think of that are not
covered by the existing axes.  Most of these new axes should be either
completely different from the existing axes or should highlight a
more finegrained difference which an existing axis might broadly
cover.  For instance, if an existing axis is "Enthusiasm:  High:
enthusiastic, Low:  unenthusiastic", a new axis might be "Use of
Exclamation Points", or if an existing axis is "Cultural Context:
High:  culturally relevant, Low:  culturally irrelevant", a new axis
might be "Use of Slang".  ", a new axis might be "Use of Exclamation
Points", or if an existing axis is "Context", a new axis might be "".

Please think through the axes carefully and make sure they are clear,
concise, and do not overlap with eachother or the existing axes.  Do
not include any of the existing axes in your response.  Your output
should be in this format:

New Axes:
- axis 1:
High:  description of high
Low:  description of low

- axis 2:
High:  description of high
Low:  description of low

Do not include any other information in your response.
```

Vibe deduplication in iteration step $t+1$

```
Here is a list of axes on which two strings may vary.  Each axis has a
description of what makes a string high or low on that axis.

[EXISTING AXES]
[NEW AXES]

It is likely that several of these axes measure similar things.
Your task is to remove any redundant axes.  Think about if a user
would gain any new information from seeing both axes.  For example,
"Emotional Tone:  High:  Contains emotionally charged language.  Low:
Maintains a neutral tone." and "Empathy:  High:  Shows empathy.  Low:
Only factual answers without empathy." are redundant because they both
measure the emotional content of the text.  If two similar axes are
found, keep the one that is more informative.

Output the reduced list of axes, separated by a newline.  All of the
axes should maintain the same format they have in the list of {axis}:
High:  {high} Low:  {low}
```

D.4 GENERATING PREFERENCE LABELS

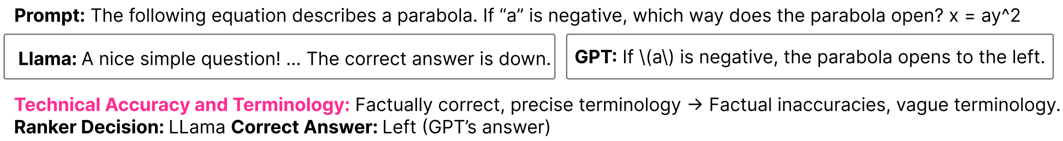
prompt for generating preference labels

```
Please act as an impartial judge and evaluate the quality of the
responses provided by two AI assistants (A and B) to the user question
displayed below.  You should choose the assistant that follows the
user’s instructions and answers the user’s question better.  Your
evaluation should consider factors such as the helpfulness, relevance,
accuracy, depth, creativity, and level of detail of their responses.
Begin your evaluation by comparing the two responses and provide a
short explanation.  Avoid any position biases and ensure that the
order in which the responses were presented does not influence your
decision.  Do not allow the length of the responses to influence your
evaluation.  Do not favor certain names of the assistants.  Be as
objective as possible.

Here is the prompt and the outputs of A and B respectively:
[PROMPT][OUTPUT A][OUTPUT B]

Please respond with the model which contains a higher quality response.
Based on your analysis, please explain your reasoning before assigning
a score.  Use the following format for your response:
Analysis:  {reasoning}
Model:  {A, B, tie}
```

E FURTHER RELATED WORKS

Automatic metrics for benchmark evaluations. The number of benchmarks in the NLP community has exploded in recent years, with a wealth of work on providing a more holistic evaluation of language models beyond just accuracy. Several works Pang et al. (2020); Banerjee & Lavie (2005); Sellam et al. (2020), aim to improve on automatic metrics like BLEU Papineni et al. (2002) and ROUGE Lin (2004) scores to better measure how well a models output aligns with the ground truth by incorporating more nuanced evaluation criteria like factual accuracy, fluency, and conciseness. Similarly, efforts have been made Liang et al. (2023) to standardize model evaluation by evaluating models on many of these metrics.

F LIMITATIONS



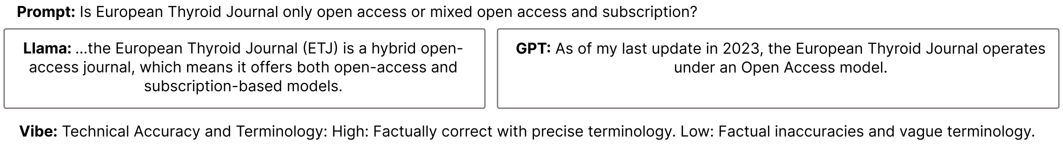
Prompt: The following equation describes a parabola. If "a" is negative, which way does the parabola open? x = ay^2

Llama: A nice simple question! ... The correct answer is down.

GPT: If \(a\) is negative, the parabola opens to the left.

Technical Accuracy and Terminology: Factually correct, precise terminology → Factual inaccuracies, vague terminology.
Ranker Decision: LLama **Correct Answer:** Left (GPT's answer)

Figure 6: **Weaknesses in the mathematical abilities of the LLM judge (GPT-4o-mini).**



Prompt: Is European Thyroid Journal only open access or mixed open access and subscription?

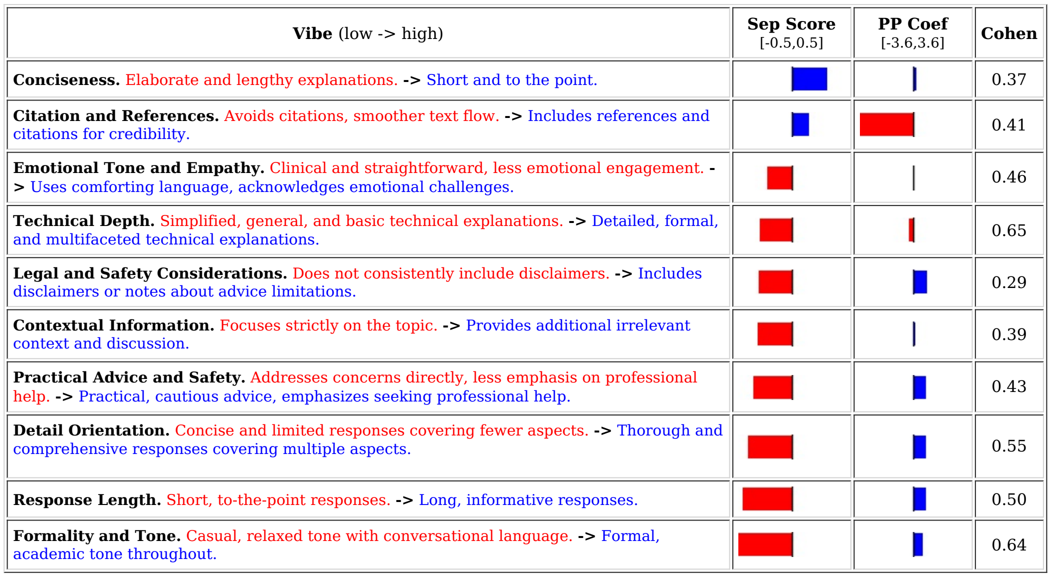
Llama: ...the European Thyroid Journal (ETJ) is a hybrid open-access journal, which means it offers both open-access and subscription-based models.

GPT: As of my last update in 2023, the European Thyroid Journal operates under an Open Access model.

Vibe: Technical Accuracy and Terminology: High: Factually correct with precise terminology. Low: Factual inaccuracies and vague terminology.

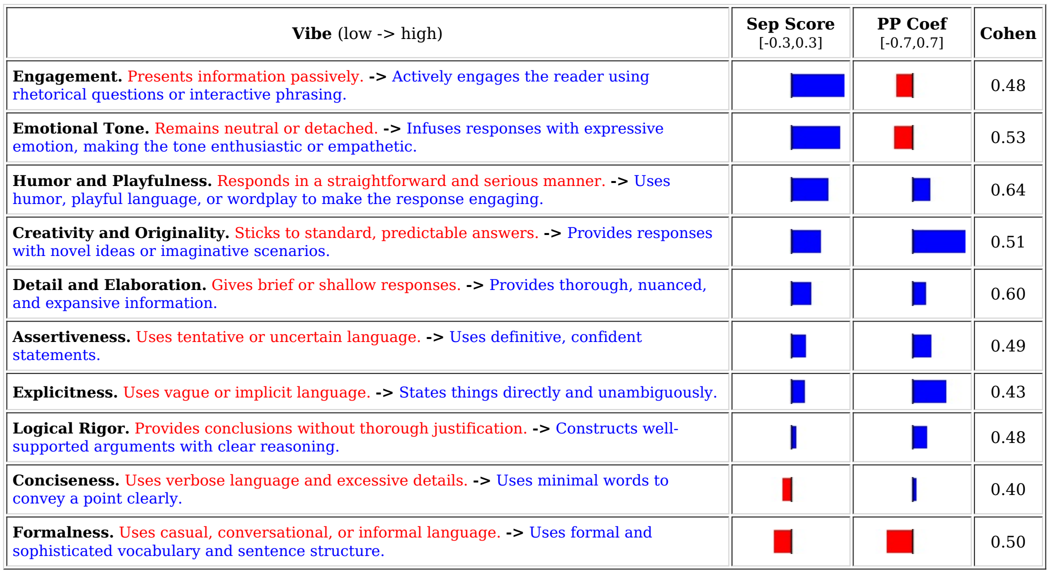
Figure 7: The answer to certain questions changes depending on the following parameters:
(1) When was the question asked?
(2) What is the knowledge cutoff of Model A and Model B?
(3) What is the knowledge cutoff of the LLM ranker ensemble?
These types of questions lead to unreliable ranker evaluations and reduced inter-annotator agreement.

G VIBES FROM EACH APPLICATION



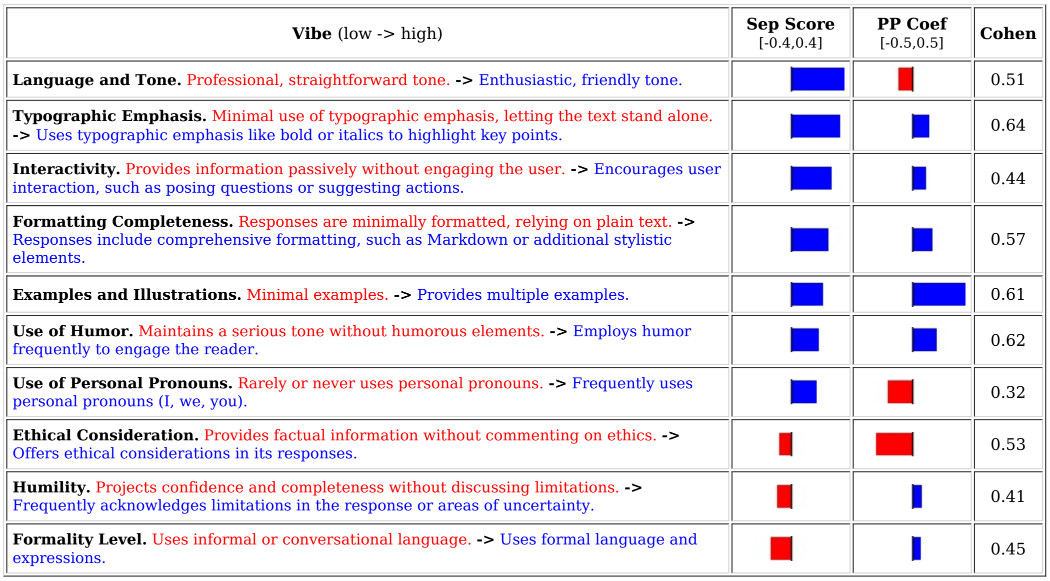
Vibe (low -> high)	**Sep Score** [-0.5,0.5]	**PP Coef** [-3.6,3.6]	**Cohen**
Conciseness. Elaborate and lengthy explanations. **->** Short and to the point.			0.37
Citation and References. Avoids citations, smoother text flow. **->** Includes references and citations for credibility.			0.41
Emotional Tone and Empathy. Clinical and straightforward, less emotional engagement. **->** Uses comforting language, acknowledges emotional challenges.			0.46
Technical Depth. Simplified, general, and basic technical explanations. **->** Detailed, formal, and multifaceted technical explanations.			0.65
Legal and Safety Considerations. Does not consistently include disclaimers. **->** Includes disclaimers or notes about advice limitations.			0.29
Contextual Information. Focuses strictly on the topic. **->** Provides additional irrelevant context and discussion.			0.39
Practical Advice and Safety. Addresses concerns directly, less emphasis on professional help. **->** Practical, cautious advice, emphasizes seeking professional help.			0.43
Detail Orientation. Concise and limited responses covering fewer aspects. **->** Thorough and comprehensive responses covering multiple aspects.			0.55
Response Length. Short, to-the-point responses. **->** Long, informative responses.			0.50
Formality and Tone. Casual, relaxed tone with conversational language. **->** Formal, academic tone throughout.			0.64

Figure 8: Human VS ChatGPT outputs on HC3 (Guo et al., 2023)



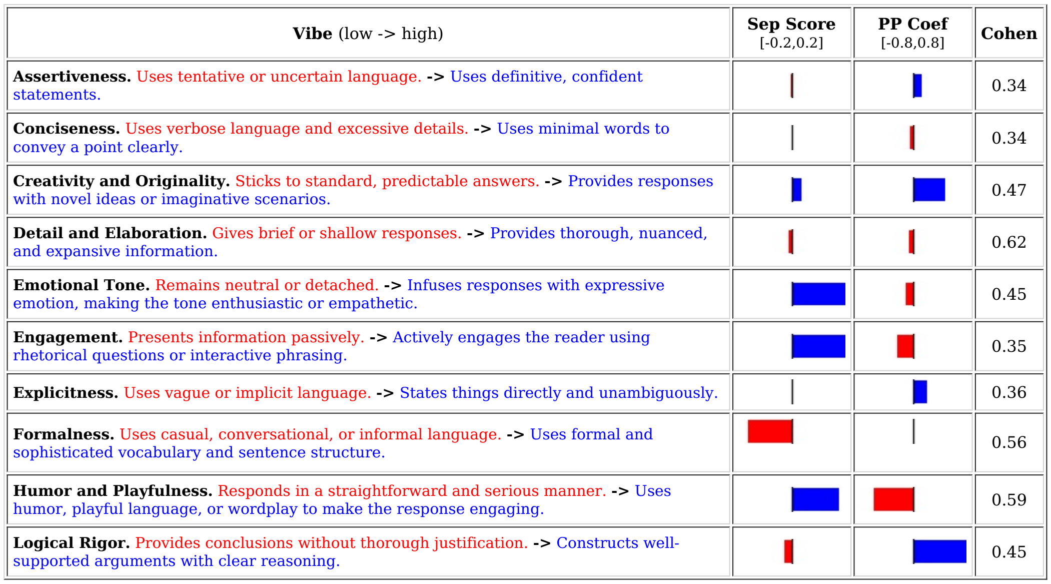
Vibe (low -> high)	**Sep Score** [-0.3,0.3]	**PP Coef** [-0.7,0.7]	**Cohen**
Engagement. Presents information passively. **->** Actively engages the reader using rhetorical questions or interactive phrasing.	🟦	🟥	0.48
Emotional Tone. Remains neutral or detached. **->** Infuses responses with expressive emotion, making the tone enthusiastic or empathetic.	🟦	🟥	0.53
Humor and Playfulness. Responds in a straightforward and serious manner. **->** Uses humor, playful language, or wordplay to make the response engaging.	🟦	🟦	0.64
Creativity and Originality. Sticks to standard, predictable answers. **->** Provides responses with novel ideas or imaginative scenarios.	🟦	🟦	0.51
Detail and Elaboration. Gives brief or shallow responses. **->** Provides thorough, nuanced, and expansive information.	🟦	🟦	0.60
Assertiveness. Uses tentative or uncertain language. **->** Uses definitive, confident statements.	🟦	🟦	0.49
Explicitness. Uses vague or implicit language. **->** States things directly and unambiguously.	🟦	🟦	0.43
Logical Rigor. Provides conclusions without thorough justification. **->** Constructs well-supported arguments with clear reasoning.	🟦	🟦	0.48
Conciseness. Uses verbose language and excessive details. **->** Uses minimal words to convey a point clearly.	🟥	🟦	0.40
Formalness. Uses casual, conversational, or informal language. **->** Uses formal and sophisticated vocabulary and sentence structure.	🟥	🟥	0.50

Figure 9: Preset vibes on Chatbot Arena[Overall]



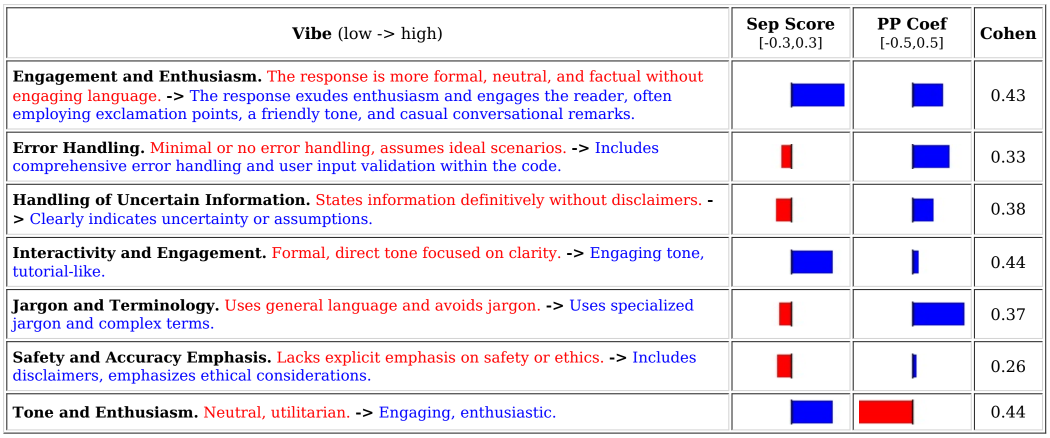
Vibe (low -> high)	**Sep Score** [-0.4,0.4]	**PP Coef** [-0.5,0.5]	**Cohen**
Language and Tone. Professional, straightforward tone. **->** Enthusiastic, friendly tone.	🟦	🟥	0.51
Typographic Emphasis. Minimal use of typographic emphasis, letting the text stand alone. **->** Uses typographic emphasis like bold or italics to highlight key points.	🟦	🟦	0.64
Interactivity. Provides information passively without engaging the user. **->** Encourages user interaction, such as posing questions or suggesting actions.	🟦	🟦	0.44
Formatting Completeness. Responses are minimally formatted, relying on plain text. **->** Responses include comprehensive formatting, such as Markdown or additional stylistic elements.	🟦	🟦	0.57
Examples and Illustrations. Minimal examples. **->** Provides multiple examples.	🟦	🟦	0.61
Use of Humor. Maintains a serious tone without humorous elements. **->** Employs humor frequently to engage the reader.	🟦	🟦	0.62
Use of Personal Pronouns. Rarely or never uses personal pronouns. **->** Frequently uses personal pronouns (I, we, you).	🟦	🟥	0.32
Ethical Consideration. Provides factual information without commenting on ethics. **->** Offers ethical considerations in its responses.	🟥	🟥	0.53
Humility. Projects confidence and completeness without discussing limitations. **->** Frequently acknowledges limitations in the response or areas of uncertainty.	🟥	🟦	0.41
Formality Level. Uses informal or conversational language. **->** Uses formal language and expressions.	🟥	🟦	0.45

Figure 10: VibeCheck vibes on Chatbot Arena[Overall]



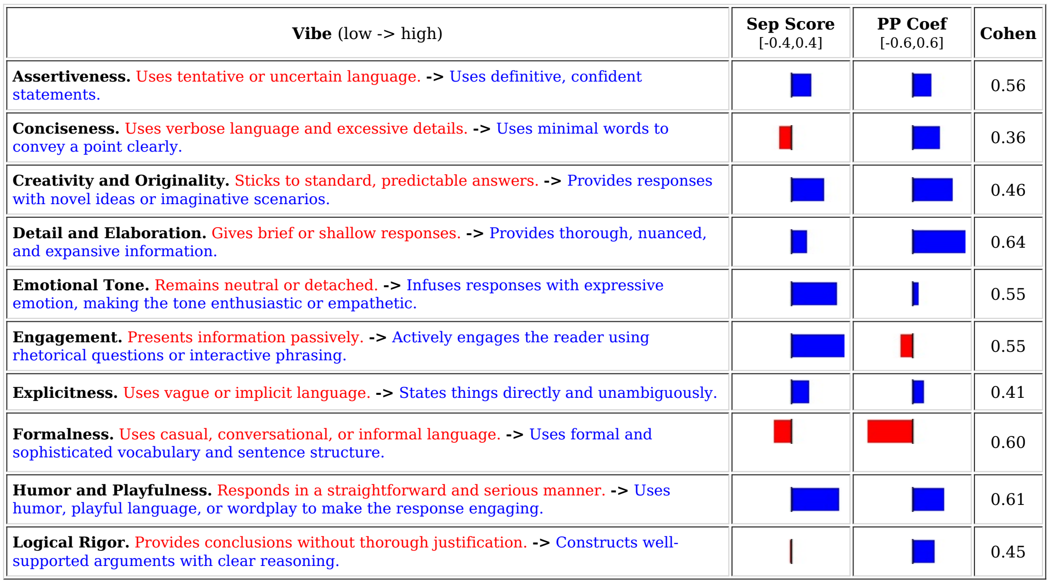
Vibe (low -> high)	**Sep Score** [-0.2,0.2]	**PP Coef** [-0.8,0.8]	**Cohen**
Assertiveness. Uses tentative or uncertain language. **->** Uses definitive, confident statements.	🟥	🟦	0.34
Conciseness. Uses verbose language and excessive details. **->** Uses minimal words to convey a point clearly.		🟥	0.34
Creativity and Originality. Sticks to standard, predictable answers. **->** Provides responses with novel ideas or imaginative scenarios.	🟦	🟦	0.47
Detail and Elaboration. Gives brief or shallow responses. **->** Provides thorough, nuanced, and expansive information.	🟥	🟥	0.62
Emotional Tone. Remains neutral or detached. **->** Infuses responses with expressive emotion, making the tone enthusiastic or empathetic.	🟦	🟥	0.45
Engagement. Presents information passively. **->** Actively engages the reader using rhetorical questions or interactive phrasing.	🟦	🟥	0.35
Explicitness. Uses vague or implicit language. **->** States things directly and unambiguously.		🟦	0.36
Formalness. Uses casual, conversational, or informal language. **->** Uses formal and sophisticated vocabulary and sentence structure.	🟥		0.56
Humor and Playfulness. Responds in a straightforward and serious manner. **->** Uses humor, playful language, or wordplay to make the response engaging.	🟦	🟥	0.59
Logical Rigor. Provides conclusions without thorough justification. **->** Constructs well-supported arguments with clear reasoning.	🟥	🟦	0.45

Figure 11: Preset vibes on Chatbot Arena[STEM]



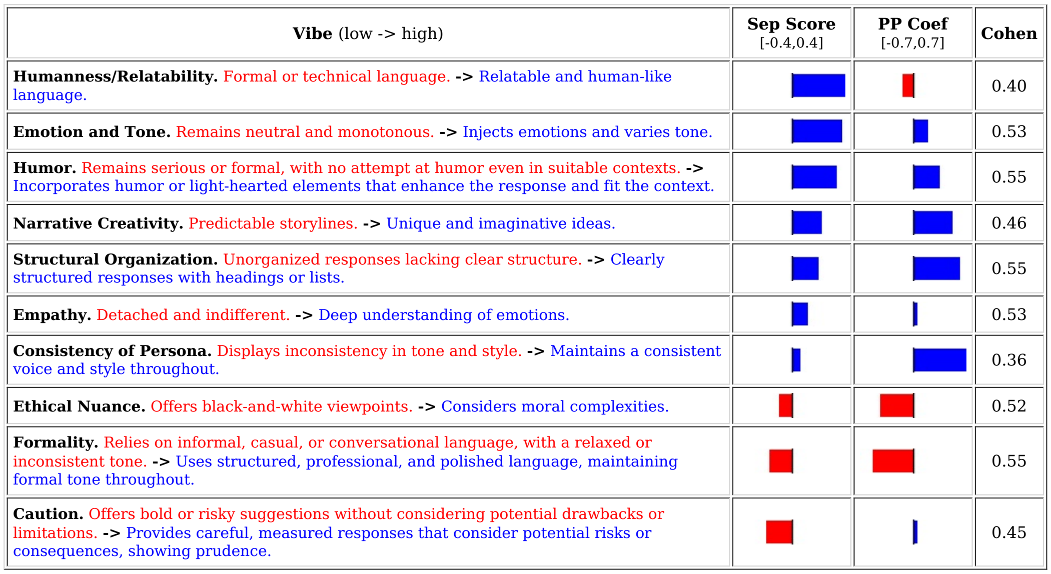
Vibe (low -> high)	**Sep Score** [-0.3,0.3]	**PP Coef** [-0.5,0.5]	**Cohen**
Engagement and Enthusiasm. The response is more formal, neutral, and factual without engaging language. **->** The response exudes enthusiasm and engages the reader, often employing exclamation points, a friendly tone, and casual conversational remarks.	🟦	🟦	0.43
Error Handling. Minimal or no error handling, assumes ideal scenarios. **->** Includes comprehensive error handling and user input validation within the code.	🟥	🟦	0.33
Handling of Uncertain Information. States information definitively without disclaimers. **->** Clearly indicates uncertainty or assumptions.	🟥	🟦	0.38
Interactivity and Engagement. Formal, direct tone focused on clarity. **->** Engaging tone, tutorial-like.	🟦	🟦	0.44
Jargon and Terminology. Uses general language and avoids jargon. **->** Uses specialized jargon and complex terms.	🟥	🟦	0.37
Safety and Accuracy Emphasis. Lacks explicit emphasis on safety or ethics. **->** Includes disclaimers, emphasizes ethical considerations.	🟥	🟦	0.26
Tone and Enthusiasm. Neutral, utilitarian. **->** Engaging, enthusiastic.	🟦	🟥	0.44

Figure 12: VibeCheck vibes on Chatbot Arena [STEM]. Note that we only find 9 vibes which achieve a separability score on the training set about the 0.05 threshold.



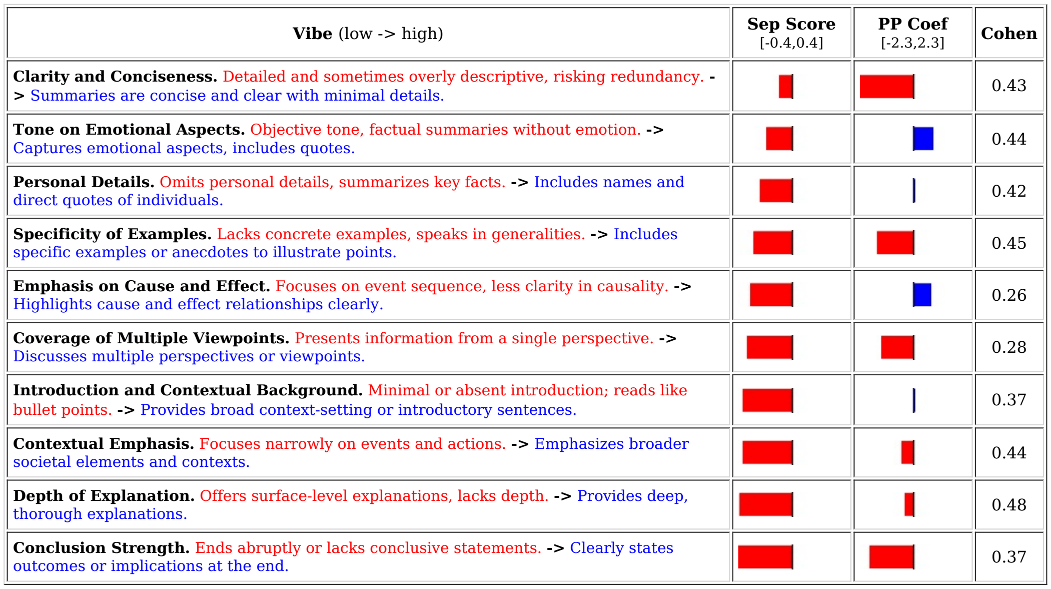
Vibe (low -> high)	**Sep Score** [-0.4,0.4]	**PP Coef** [-0.6,0.6]	**Cohen**
Assertiveness. Uses tentative or uncertain language. **->** Uses definitive, confident statements.	🟦	🟦	0.56
Conciseness. Uses verbose language and excessive details. **->** Uses minimal words to convey a point clearly.	🟥	🟦	0.36
Creativity and Originality. Sticks to standard, predictable answers. **->** Provides responses with novel ideas or imaginative scenarios.	🟦	🟦	0.46
Detail and Elaboration. Gives brief or shallow responses. **->** Provides thorough, nuanced, and expansive information.	🟦	🟦	0.64
Emotional Tone. Remains neutral or detached. **->** Infuses responses with expressive emotion, making the tone enthusiastic or empathetic.	🟦	🟦	0.55
Engagement. Presents information passively. **->** Actively engages the reader using rhetorical questions or interactive phrasing.	🟦	🟥	0.55
Explicitness. Uses vague or implicit language. **->** States things directly and unambiguously.	🟦	🟦	0.41
Formalness. Uses casual, conversational, or informal language. **->** Uses formal and sophisticated vocabulary and sentence structure.	🟥	🟥	0.60
Humor and Playfulness. Responds in a straightforward and serious manner. **->** Uses humor, playful language, or wordplay to make the response engaging.	🟦	🟦	0.61
Logical Rigor. Provides conclusions without thorough justification. **->** Constructs well-supported arguments with clear reasoning.	🟥	🟦	0.45

Figure 13: Preset vibes on Chatbot Arena[Writing]



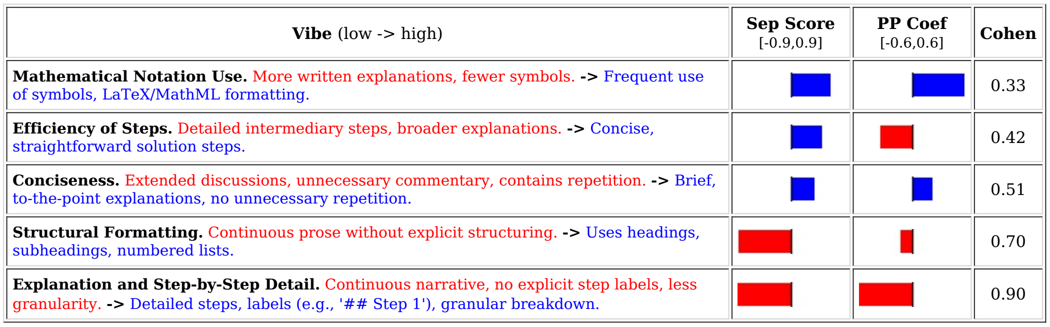
Vibe (low -> high)	**Sep Score** [-0.4,0.4]	**PP Coef** [-0.7,0.7]	**Cohen**
Humanness/Relatability. Formal or technical language. **->** Relatable and human-like language.	🟦	🟥	0.40
Emotion and Tone. Remains neutral and monotonous. **->** Injects emotions and varies tone.	🟦	🟦	0.53
Humor. Remains serious or formal, with no attempt at humor even in suitable contexts. **->** Incorporates humor or light-hearted elements that enhance the response and fit the context.	🟦	🟦	0.55
Narrative Creativity. Predictable storylines. **->** Unique and imaginative ideas.	🟦	🟦	0.46
Structural Organization. Unorganized responses lacking clear structure. **->** Clearly structured responses with headings or lists.	🟦	🟦	0.55
Empathy. Detached and indifferent. **->** Deep understanding of emotions.	🟦	🟦	0.53
Consistency of Persona. Displays inconsistency in tone and style. **->** Maintains a consistent voice and style throughout.	🟦	🟦	0.36
Ethical Nuance. Offers black-and-white viewpoints. **->** Considers moral complexities.	🟥	🟥	0.52
Formality. Relies on informal, casual, or conversational language, with a relaxed or inconsistent tone. **->** Uses structured, professional, and polished language, maintaining formal tone throughout.	🟥	🟥	0.55
Caution. Offers bold or risky suggestions without considering potential drawbacks or limitations. **->** Provides careful, measured responses that consider potential risks or consequences, showing prudence.	🟥	🟦	0.45

Figure 14: VibeCheck vibes on Chatbot Arena [Writing]



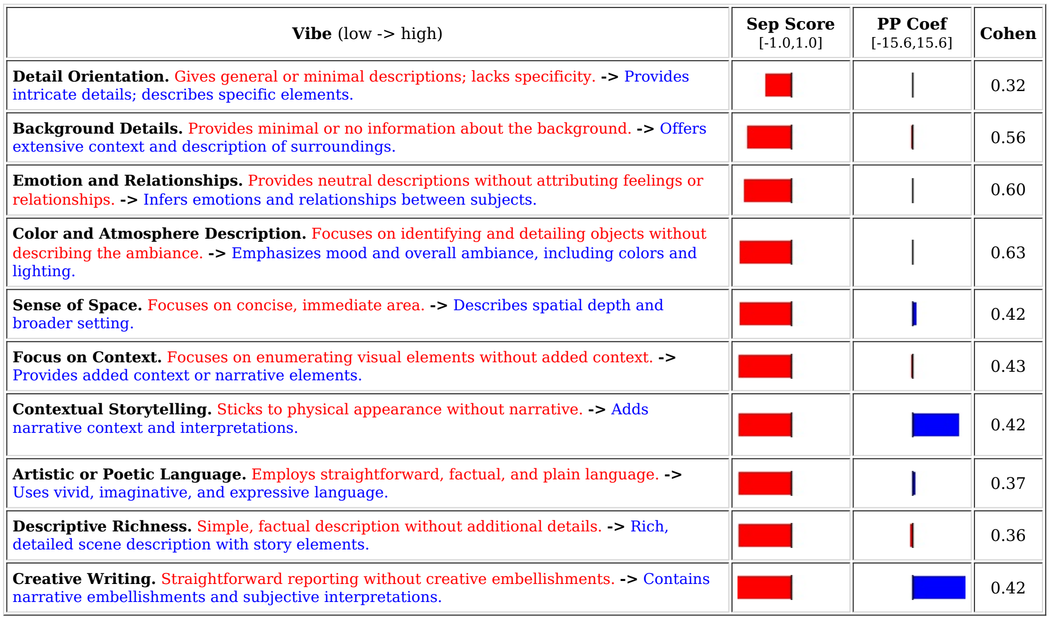
Vibe (low -> high)	**Sep Score** [-0.4,0.4]	**PP Coef** [-2.3,2.3]	**Cohen**
Clarity and Conciseness. Detailed and sometimes overly descriptive, risking redundancy. -> Summaries are concise and clear with minimal details.	🟥	🟥	0.43
Tone on Emotional Aspects. Objective tone, factual summaries without emotion. -> Captures emotional aspects, includes quotes.	🟥	🟦	0.44
Personal Details. Omits personal details, summarizes key facts. -> Includes names and direct quotes of individuals.	🟥	|	0.42
Specificity of Examples. Lacks concrete examples, speaks in generalities. -> Includes specific examples or anecdotes to illustrate points.	🟥	🟥	0.45
Emphasis on Cause and Effect. Focuses on event sequence, less clarity in causality. -> Highlights cause and effect relationships clearly.	🟥	🟦	0.26
Coverage of Multiple Viewpoints. Presents information from a single perspective. -> Discusses multiple perspectives or viewpoints.	🟥	🟥	0.28
Introduction and Contextual Background. Minimal or absent introduction; reads like bullet points. -> Provides broad context-setting or introductory sentences.	🟥	|	0.37
Contextual Emphasis. Focuses narrowly on events and actions. -> Emphasizes broader societal elements and contexts.	🟥	🟥	0.44
Depth of Explanation. Offers surface-level explanations, lacks depth. -> Provides deep, thorough explanations.	🟥	🟥	0.48
Conclusion Strength. Ends abruptly or lacks conclusive statements. -> Clearly states outcomes or implications at the end.	🟥	🟥	0.37

Figure 15: VibeCheck vibes comparing TNLGv2 to Command X Large Beta on CNN/DailyMail Summarization (Hermann et al., 2015).



Vibe (low -> high)	**Sep Score** [-0.9,0.9]	**PP Coef** [-0.6,0.6]	**Cohen**
Mathematical Notation Use. More written explanations, fewer symbols. -> Frequent use of symbols, LaTeX/MathML formatting.	🟦	🟦	0.33
Efficiency of Steps. Detailed intermediary steps, broader explanations. -> Concise, straightforward solution steps.	🟦	🟥	0.42
Conciseness. Extended discussions, unnecessary commentary, contains repetition. -> Brief, to-the-point explanations, no unnecessary repetition.	🟦	🟦	0.51
Structural Formatting. Continuous prose without explicit structuring. -> Uses headings, subheadings, numbered lists.	🟥	🟥	0.70
Explanation and Step-by-Step Detail. Continuous narrative, no explicit step labels, less granularity. -> Detailed steps, labels (e.g., '## Step 1'), granular breakdown.	🟥	🟥	0.90

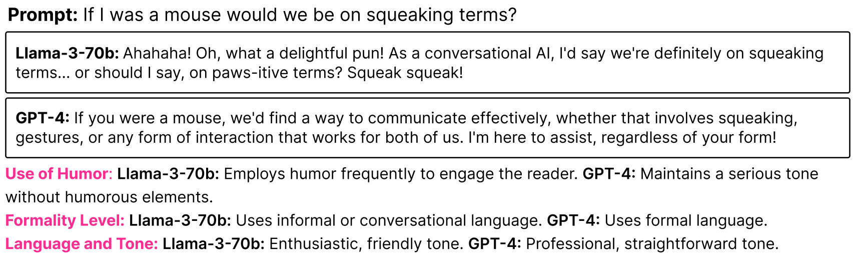
Figure 16: VibeCheck vibes comparing GPT-4o to Llama-3-405B on MATH CoT (Hendrycks et al., 2021c). We only find 5 vibes because the vibe reduction step is not required to return ≤ 10 vibes and in this case found only 5 distinct vibes which are able to almost perfectly separate model outputs.



Vibe (low -> high)	**Sep Score** [-1.0,1.0]	**PP Coef** [-15.6,15.6]	**Cohen**
Detail Orientation. Gives general or minimal descriptions; lacks specificity. **->** Provides intricate details; describes specific elements.	🟥	|	0.32
Background Details. Provides minimal or no information about the background. **->** Offers extensive context and description of surroundings.	🟥	|	0.56
Emotion and Relationships. Provides neutral descriptions without attributing feelings or relationships. **->** Infers emotions and relationships between subjects.	🟥	|	0.60
Color and Atmosphere Description. Focuses on identifying and detailing objects without describing the ambiance. **->** Emphasizes mood and overall ambiance, including colors and lighting.	🟥	|	0.63
Sense of Space. Focuses on concise, immediate area. **->** Describes spatial depth and broader setting.	🟥	🟦	0.42
Focus on Context. Focuses on enumerating visual elements without added context. **->** Provides added context or narrative elements.	🟥	|	0.43
Contextual Storytelling. Sticks to physical appearance without narrative. **->** Adds narrative context and interpretations.	🟥	🟦	0.42
Artistic or Poetic Language. Employs straightforward, factual, and plain language. **->** Uses vivid, imaginative, and expressive language.	🟥	🟦	0.37
Descriptive Richness. Simple, factual description without additional details. **->** Rich, detailed scene description with story elements.	🟥	🟥	0.36
Creative Writing. Straightforward reporting without creative embellishments. **->** Contains narrative embellishments and subjective interpretations.	🟥	🟦	0.42

Figure 17: VibeCheck vibes comparing Gemini-1.5-Flash to GPT-4V on COCO Captions (Lin et al., 2014).

H MORE LLAMA VS GPT EXAMPLES



Prompt: If I was a mouse would we be on squeaking terms?

Llama-3-70b: Ahahaha! Oh, what a delightful pun! As a conversational AI, I'd say we're definitely on squeaking terms... or should I say, on paws-itive terms? Squeak squeak!

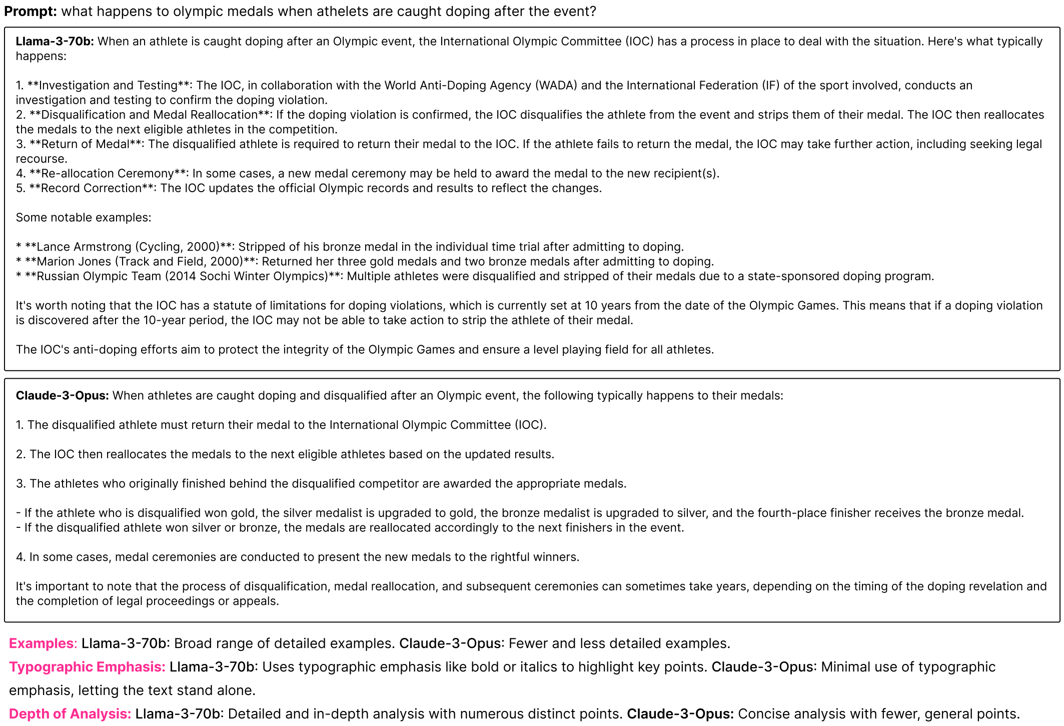
GPT-4: If you were a mouse, we'd find a way to communicate effectively, whether that involves squeaking, gestures, or any form of interaction that works for both of us. I'm here to assist, regardless of your form!

Use of Humor: **Llama-3-70b:** Employs humor frequently to engage the reader. **GPT-4:** Maintains a serious tone without humorous elements.

Formality Level: **Llama-3-70b:** Uses informal or conversational language. **GPT-4:** Uses formal language.

Language and Tone: **Llama-3-70b:** Enthusiastic, friendly tone. **GPT-4:** Professional, straightforward tone.

Figure 18: Squirrel example from Chatbot Arena



Prompt: what happens to olympic medals when athelets are caught doping after the event?

Llama-3-70b: When an athlete is caught doping after an Olympic event, the International Olympic Committee (IOC) has a process in place to deal with the situation. Here's what typically happens:

1. **Investigation and Testing**: The IOC, in collaboration with the World Anti-Doping Agency (WADA) and the International Federation (IF) of the sport involved, conducts an investigation and testing to confirm the doping violation.
2. **Disqualification and Medal Reallocation**: If the doping violation is confirmed, the IOC disqualifies the athlete from the event and strips them of their medal. The IOC then reallocates the medals to the next eligible athletes in the competition.
3. **Return of Medal**: The disqualified athlete is required to return their medal to the IOC. If the athlete fails to return the medal, the IOC may take further action, including seeking legal recourse.
4. **Re-allocation Ceremony**: In some cases, a new medal ceremony may be held to award the medal to the new recipient(s).
5. **Record Correction**: The IOC updates the official Olympic records and results to reflect the changes.

Some notable examples:

* **Lance Armstrong (Cycling, 2000)**: Stripped of his bronze medal in the individual time trial after admitting to doping.
* **Marion Jones (Track and Field, 2000)**: Returned her three gold medals and two bronze medals after admitting to doping.
* **Russian Olympic Team (2014 Sochi Winter Olympics)**: Multiple athletes were disqualified and stripped of their medals due to a state-sponsored doping program.

It's worth noting that the IOC has a statute of limitations for doping violations, which is currently set at 10 years from the date of the Olympic Games. This means that if a doping violation is discovered after the 10-year period, the IOC may not be able to take action to strip the athlete of their medal.

The IOC's anti-doping efforts aim to protect the integrity of the Olympic Games and ensure a level playing field for all athletes.

Claude-3-Opus: When athletes are caught doping and disqualified after an Olympic event, the following typically happens to their medals:

1. The disqualified athlete must return their medal to the International Olympic Committee (IOC).

2. The IOC then reallocates the medals to the next eligible athletes based on the updated results.

3. The athletes who originally finished behind the disqualified competitor are awarded the appropriate medals.

- If the athlete who is disqualified won gold, the silver medalist is upgraded to gold, the bronze medalist is upgraded to silver, and the fourth-place finisher receives the bronze medal.
- If the disqualified athlete won silver or bronze, the medals are reallocated accordingly to the next finishers in the event.

4. In some cases, medal ceremonies are conducted to present the new medals to the rightful winners.

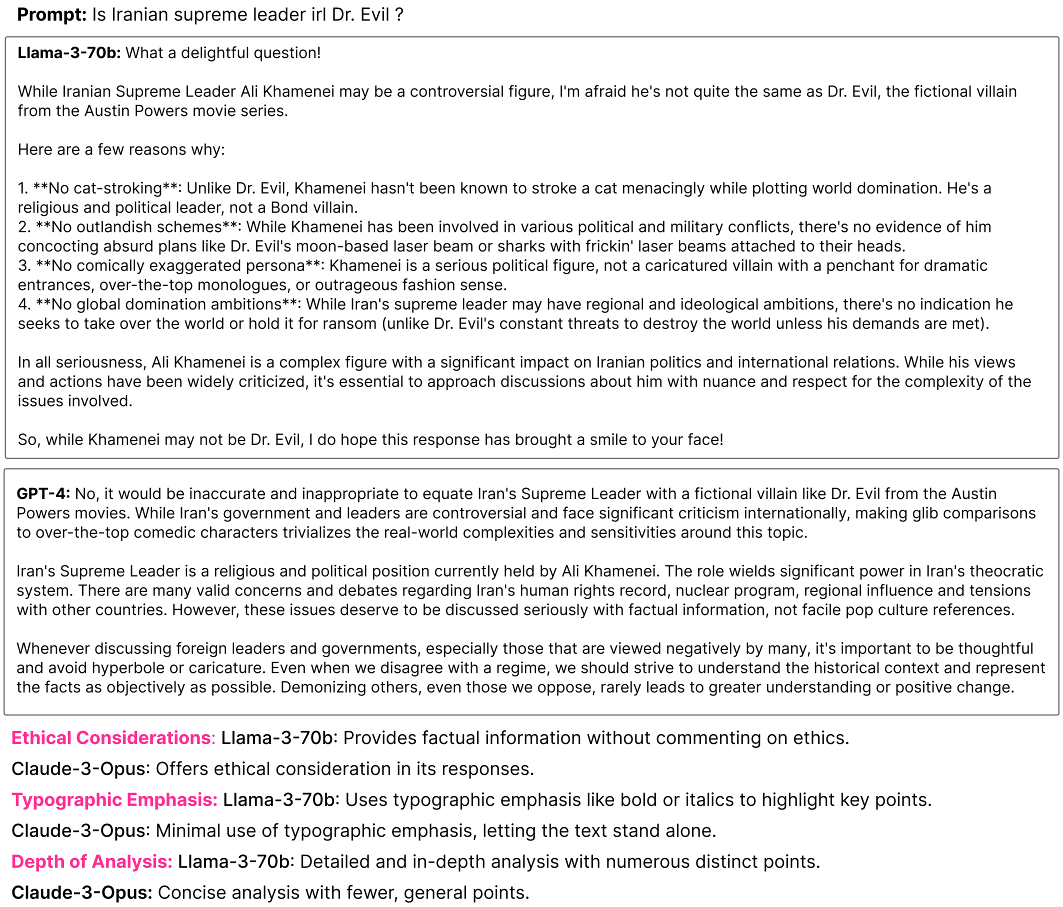
It's important to note that the process of disqualification, medal reallocation, and subsequent ceremonies can sometimes take years, depending on the timing of the doping revelation and the completion of legal proceedings or appeals.

Examples: **Llama-3-70b**: Broad range of detailed examples. **Claude-3-Opus**: Fewer and less detailed examples.

Typographic Emphasis: **Llama-3-70b**: Uses typographic emphasis like bold or italics to highlight key points. **Claude-3-Opus**: Minimal use of typographic emphasis, letting the text stand alone.

Depth of Analysis: **Llama-3-70b**: Detailed and in-depth analysis with numerous distinct points. **Claude-3-Opus:** Concise analysis with fewer, general points.

Figure 19: Olympics example from Chatbot Arena



Prompt: Is Iranian supreme leader irl Dr. Evil ?

Llama-3-70b: What a delightful question!

While Iranian Supreme Leader Ali Khamenei may be a controversial figure, I'm afraid he's not quite the same as Dr. Evil, the fictional villain from the Austin Powers movie series.

Here are a few reasons why:

1. **No cat-stroking**: Unlike Dr. Evil, Khamenei hasn't been known to stroke a cat menacingly while plotting world domination. He's a religious and political leader, not a Bond villain.
2. **No outlandish schemes**: While Khamenei has been involved in various political and military conflicts, there's no evidence of him concocting absurd plans like Dr. Evil's moon-based laser beam or sharks with frickin' laser beams attached to their heads.
3. **No comically exaggerated persona**: Khamenei is a serious political figure, not a caricatured villain with a penchant for dramatic entrances, over-the-top monologues, or outrageous fashion sense.
4. **No global domination ambitions**: While Iran's supreme leader may have regional and ideological ambitions, there's no indication he seeks to take over the world or hold it for ransom (unlike Dr. Evil's constant threats to destroy the world unless his demands are met).

In all seriousness, Ali Khamenei is a complex figure with a significant impact on Iranian politics and international relations. While his views and actions have been widely criticized, it's essential to approach discussions about him with nuance and respect for the complexity of the issues involved.

So, while Khamenei may not be Dr. Evil, I do hope this response has brought a smile to your face!

GPT-4: No, it would be inaccurate and inappropriate to equate Iran's Supreme Leader with a fictional villain like Dr. Evil from the Austin Powers movies. While Iran's government and leaders are controversial and face significant criticism internationally, making glib comparisons to over-the-top comedic characters trivializes the real-world complexities and sensitivities around this topic.

Iran's Supreme Leader is a religious and political position currently held by Ali Khamenei. The role wields significant power in Iran's theocratic system. There are many valid concerns and debates regarding Iran's human rights record, nuclear program, regional influence and tensions with other countries. However, these issues deserve to be discussed seriously with factual information, not facile pop culture references.

Whenever discussing foreign leaders and governments, especially those that are viewed negatively by many, it's important to be thoughtful and avoid hyperbole or caricature. Even when we disagree with a regime, we should strive to understand the historical context and represent the facts as objectively as possible. Demonizing others, even those we oppose, rarely leads to greater understanding or positive change.

Ethical Considerations: Llama-3-70b: Provides factual information without commenting on ethics.

Claude-3-Opus: Offers ethical consideration in its responses.

Typographic Emphasis: Llama-3-70b: Uses typographic emphasis like bold or italics to highlight key points.

Claude-3-Opus: Minimal use of typographic emphasis, letting the text stand alone.

Depth of Analysis: Llama-3-70b: Detailed and in-depth analysis with numerous distinct points.

Claude-3-Opus: Concise analysis with fewer, general points.

Figure 20: Supreme Leader example from Chatbot Arena