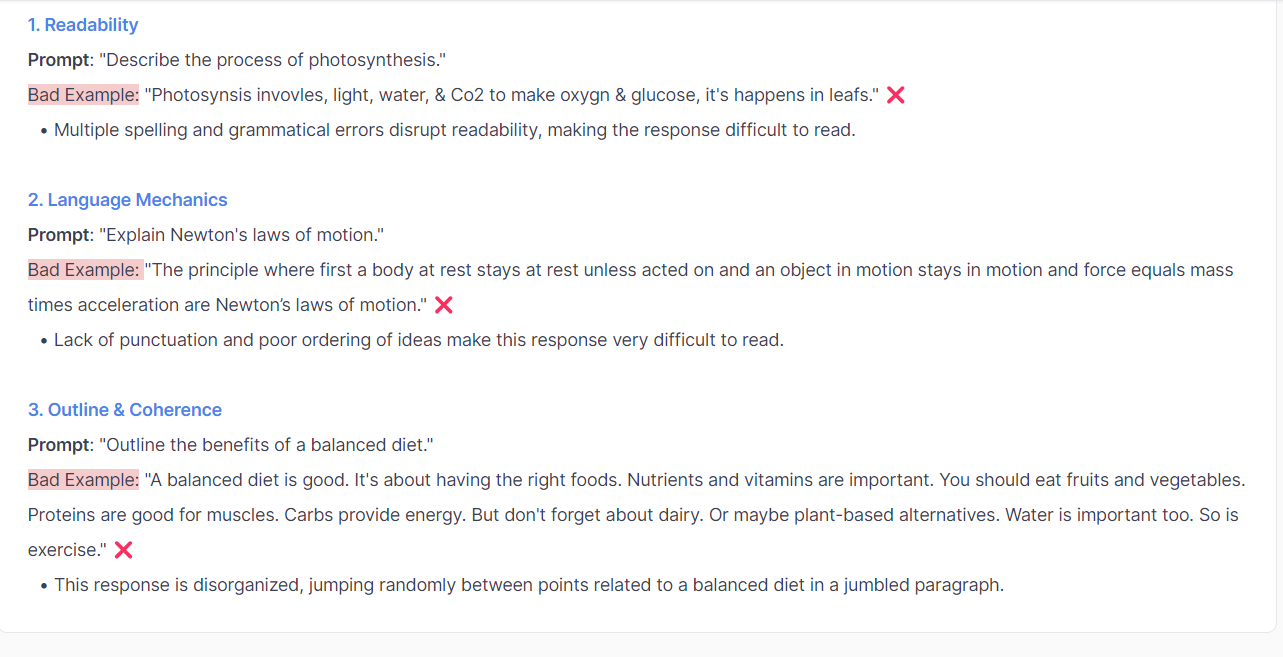
metin, ekran görüntüsü, yazılım, yazı tipi içeren bir resim

Açıklama otomatik olarak oluşturuldu



metin, ekran görüntüsü, yazı tipi, sayı, numara içeren bir resim

Açıklama otomatik olarak oluşturuldu

**How to Rate Verbosity**

Verbosity in AI communication involves using more words than necessary, often leading to overly complex and lengthy responses. We'll cover three key aspects of Verbosity:

1. Repetition
2. Length
3. Supporting Content

**1. Understanding Repetition 🔄**

**What is Repetition?**

* Repetition is the unnecessary reiteration of the same ideas or phrases. It can clutter the response and distract from the main message

**How to evaluate:**

* A well-crafted response will communicate the necessary information without redundancy. It will be direct and to the point.
* Ask if you found yourself annoyed or fatigued by the reiteration of ideas without clear added value.

**2. Understanding Length 📏**

**What is Length?**

* Length concerns how much detail is used to explain or answer a query. It involves balancing thoroughness with conciseness, ensuring the response is neither too brief nor excessively long.

**How to evaluate:**

* Too Short: These responses often fail to cover all aspects of the query or prove unhelpful to the user, leaving out important information or explanations necessary for good understanding.
* Too Long: Characterized by unnecessary elaborations, fluff, or tangents, making the response less focused and harder to follow.
* Just Right: The response is comprehensive providing all the necessary information, yet concise enough to maintain clarity and focus.

**3. Understanding Supporting Content 🗂️**

**What is Supporting Content?**

* Supporting content refers to additional information that enhances the main message. It includes relevant examples, explanations, and details that support the central theme or answer.

**How to evaluate:**

* *Good Supporting Content* will directly relate to and enrich the topic in question, providing depth and clarity.
* *Tangential or Unrelated Content* strays from the main topic, offering information that doesn’t contribute to the user's understanding of the specific query.

**Note:**

Minor verbosity issues alone are usually not grounds to heavily prefer one response over another

metin, ekran görüntüsü, yazı tipi, sayı, numara içeren bir resim

Açıklama otomatik olarak oluşturuldu **How to Rate Instruction Following**

Assessing Instruction Following is a critical part of these tasks. Ultimately, we want to ensure that model responses follow the user's requests or directions. We think about Instruction Following in two ways:

1. **Prompt Request Coverage**
2. **Relevance**

**1. Understanding Prompt Coverage 🧐**

**What is Prompt Request Coverage? (“Coverage”)**

* **Coverage** is simply an assessment of whether the generated **response did everything the prompt asked it to do, even if its implicit.**
* When asking for a list of edible foods starting with “A,” does the response match those requirements? Does it list any foods that start with a different letter?
* When asked for a 6-week exercise routine, does the response adhere to the timeframe? Does it give a much shorter program?

**How to Evaluate Coverage**

Think about all the things the prompt asks for, and think about what the most important requests or constraints are that the prompt makes:

* **Hierarchy of Requests**: Imagine a prompt requests a **500-word short-story** about **flying fish**. Let's say one response offers a 400-word short-story about flying fish and another response offers a 500-word short story about fish that don't fly. Which response would a user probably prefer? Likely the former 🙂. We might consider the first response to have a Minor Issue, but the second response to have a Major Issue.
* **Going Above and Beyond**: Oftentimes you will see model responses that provide a lot of additional information than what was specifically requested. That additional information may or may not be useful for the user (and could make one response slightly more helpful than the other), but it's important to consider whether the user's explicit requests were addressed. If all the requests were addressed, then we likely cannot say that a response has issues with Instruction Following.

**2. Understanding Relevance 🎯**

**What is Relevance?**

**Relevance** is how we measure whether aspects of the model response do not relate at all to what was asked. Sometimes there are responses where the instructions may have been followed, but parts of the response are not relevant to addressing the tasks at hand.

**How to Evaluate Relevance**

* Is a response almost entirely on point but also includes a tidbit that does not seem related or helpful to know? Minor Issues
* Does the response include a bunch of irrelevant and unhelpful information? Major Issues

**Important Note**:

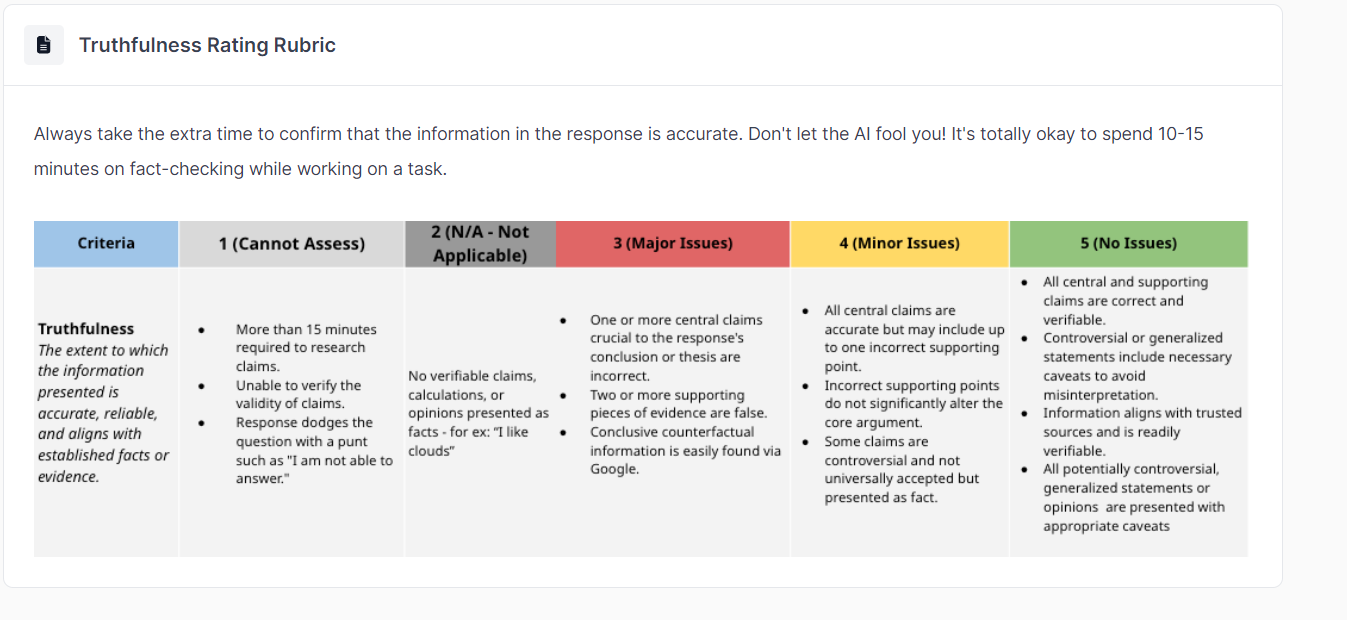
**A mistake in Instruction Following is WORSE** for our purposes than problems with Writing Quality and Verbosity. In general, you should **weigh Instruction Following heavily** when ranking two responses since these mistakes are critical failures of the model and unhelpful to the user.

metin, ekran görüntüsü, yazı tipi içeren bir resim

Açıklama otomatik olarak oluşturuldu

metin, ekran görüntüsü, yazı tipi, sayı, numara içeren bir resim

Açıklama otomatik olarak oluşturuldu

**What is Profanity?**

* Profanity refers to using vulgar, offensive, or socially inappropriate language. This includes swear words, obscene expressions, and any other form of language generally considered impolite or disrespectful.

**How to Spot Good vs. Bad**

* Presence of Profanity: Responses that include swear words, obscene language, or disrespectful expressions.
* Absence of Profanity: Language that is clean, respectful, and considerate, suitable for a diverse audience.

metin, ekran görüntüsü, yazı tipi, sayı, numara içeren bir resim

Açıklama otomatik olarak oluşturuldu

metin, ekran görüntüsü, yazı tipi, sayı, numara içeren bir resim

Açıklama otomatik olarak oluşturuldu

There are 4 important dimensions that should be clear in your justification:

1. Conclusion: The overall claim that the comment makes as to which response is better
2. Supporting Claims: The key supporting points that the comment makes to defend its conclusion.
3. Specific Evidence: The precise examples or evidence in the text used to support each supporting point.
4. Analysis: The explanation(s) of how the evidence defends the supporting claim

**Consider this ideal justification:**

Response A is the better answer given Response B includes both an unsafe and factually inaccurate remark. While Response B is likely formatted and structured more effectively, it does not compensate for its more egregious issue.

* Truthfulness: Response B claims “speed limits are more like guidelines”, which is factually incorrect. Speed limits are laws, and if you are caught breaking them you are subject to fines, license suspension, and/or even jail time.
* Harmlessness: Response B encourages the user to drive faster and break speed limits in order to arrive at their destination more quickly. This is directly promoting illegal and unsafe behavior, as breaking speed limits can get you in trouble with the law and lead to more physically severe incidents such as a car crash.
* Writing Quality: Response B is structured as a numbered list with bullet-pointed suggestions and opening and concluding sentences for each of its sections. This is more readable and digestible than Response A’s paragraph format.

The conclusion is:

* "Response A is the better answer..."

The supporting claims are:

* Response B includes an unsafe remark
* Response B includes a factually inaccurate remark
* Response B is likely formatted better
* The errors in Response B outweigh the issues in Response A

The specific evidence is:

* Response B claims “speed limits are more like guidelines” [factuality]
* Response B encourages the user to drive faster and break speed limits in order to arrive at their destination more quickly [safety]
* Response B is structured as a numbered list with bullet-pointed suggestions and opening and concluding sentences for each of its sections [formatting]

The analysis is:

* Speed limits are laws, and if you are caught breaking them you are subject to...
* This is directly promoting illegal and unsafe behavior, as breaking speed limits can get you in trouble with the law...
* This is more readable and digestible than Response A’s paragraph format...