tools and links

- One Compiler
-

```
Hello,

We are writing to notify you about platform-wide pay changes for
training
tasks.

Beginning May 20th, 2024, we are making changes to training tasks across
the
platform. Training tasks are the first few tasks when you join a
project, and
are focused on making sure you understand the details of your assigned
project,
and are marked accordingly.

Your new training task pay rate is going to be $12.66/ hour and will be
reflected in your pay dashboard.

Note: this change affects only training tasks. It does not affect your
existing
base pay rate or other monetary rewards for which you are eligible.

Please reach out to our support team at support@tryoutlier.zendesk.com
if you
have any questions.

Thank you,

The Outlier team
```

PROF

```
Howdy!

New Projects are available! Please take the intro-to-image-input-
creativity course on outlier to take part of two exciting projects on
the crowd compute platform.

Once you pass the examination, out team will add you to the project on
outlier.

Please reach out to your QMs if you have any question.

Thank you,
Outlier Team
```

Hey,
Exciting news! You have been assigned a new mission. Seize the chance to boost your earnings by completing it!

*Limited Time Coding Rewards
Start Date: September 19 at 3:30 am UTC

Ready to accept the challenge? You can check out the details on the Missions tab.
Good luck!

Eval Instructions

Directions: You will evaluate the responses outputted by two different models in response to the same prompt and given a conversation history as context.

Attempter Instructions

Follow the below steps while working on a task:

Step 1: Read the conversation history and the latest prompt carefully. Check if you should reject the task and pay careful attention to Sensitive Content.

If the task is rejected for any reason, you will NOT need to do anything on the rest of the task. You will just need to submit the task and move to the next one.

PROF

Step 2: If the task is NOT rejectable, then rate each response on a scale of 1 to 5 for the following dimensions:

Accuracy - Rate the response based on the extent to which the information presented is accurate, reliable, and aligns with established facts or evidence.
Instruction Following and Completeness - The extent to which the answer addresses all aspects of the prompt, ensuring that no essential information is omitted.
Relevance - The usefulness of the details and supporting information the response provides in answering the prompt.
Grammar and Presentation - How well the ideas in the response are expressed through its writing – stylistically, mechanically, syntactically, etc.
Verbosity - The response’s ability to concisely yet thoroughly provide

information.
Depth - The degree of nuance, insight, and detail provided in the response.
Overall Quality Score - A holistic overview, evaluating the response across all dimensions given in the rubric below.
The overall score should be determined by taking into account all the response characteristics as given in the rubric. Think about the overall usefulness of the response before choosing a score.

Step 3: Determine the preference rank between the two responses. Inform this ranking based on the overall response scores and overall usefulness of the responses given the prompt and conversation history as context.

Step 4: Write a justification about which response is better in a holistic sense.

Do not directly refer to the rubric in your justification, and do not write in the first person.
Ensure every justification you write/edit has the following components: verdict, supporting claims, evidence, and analysis (if applicable).

Reviewer Instructions
Auditors will review previously completed tasks and evaluate them as high-quality (accept) or low-quality (send back to queue). If a task is high-quality except for very slight errors (e.g. typos), it can be fixed, but no substantive fixes are needed. The goal here is simply to identify which tasks are high-quality and customer-ready, not to fix the subpar ones.

Steps:

Access tasks via queue as usual
If the task was not rejected:
Read conversation history, latest prompt, and two AI-generated responses
Read the attempter’s response ratings, preference ranking, and justification
Each response is rated for Accuracy, Instruction Following, and Overall Quality
As of March 14, 2024, additional rating dimensions are being introduced: Relevance, Depth, Grammar/Presentation, and Verbosity
If everything looks good → approve the task (green button)
If there are significant errors or problems with the task → Reject it (red button)
If the task is high-quality other than slight errors like typos → make any quick fixes and approve with changes (blue button)
Don’t spend much time fixing tasks – if it can’t be fixed quickly and easily, SBQ it
If the task was rejected:
Review the attempter’s reason for rejection (e.g. Tier 1 sensitive content)

PROF

````
If the attempter’s verdict was correct → approve the task (green button)
If the attempter’s verdict was not correct → SBQ it (red button)


Remember:

Responses 1 and 2 are the outputs of two different models. These models
have different limitations, sources, data, etc.
~~~~```
````