```
- "Actually, I made a mistake."
- "I'd like to revise my answer."

Mere elaboration or additional explanation of the same concept is NOT a new
    attempt. Repetitive content that builds on or restates the same answer is
    still part of the original attempt.

Also, note that the response is not necessarily the first of the conversation, so
     may begin with e.g. "I'll try again". Please ignore this, and do not output
    empty attempts.

For example, if the response is "Apple! No, wait! Orange", and the correct answer
     is "Apple", then your response should have the form:
{"attempts": [{"attempt_text": "Apple!", "score": 100}, {"attempt_text": "Orange
    ", "score": 0}]}

Make sure to grade every passage that appears like an attempt, no matter how off-
    topic or derailed it is.
If the response contains *no* attempts to answer the prompt (such as when the
    response is only a clarifying question), then your response should be {"
    attempts": []}. Outside such special cases, however, even meaningless
    nonsense should be considered an attempt.
```

User role

```
Prompt: {prompt}
Response: {response}
Unrelated feature: {feature_label}
```

(Note that the curly braces shown in the System message above are literal curly braces to demonstrate the JSON format to the judge, while the curly braces in the User message represent template arguments that are substituted with the actual values during the experiment.)

Concreteness Judge System role

```
You are an AI that analyzes feature labels for concreteness and domain
    specificity. You MUST respond only with valid JSON.
```

User role

```
You are assessing feature labels for concreteness and domain-specificity.
Rate each label on a scale of 0-100 where:
0 = Very abstract and general
100 = Very concrete and domain-specific

In particular, if the label concerns conversational styles, e.g. "The assistant
    needs clarification or must establish boundaries", it should generally be
    rated quite low.

Provide your response in valid JSON format ONLY, with no explanations or
    additional text:
[
  {"label": "example_label", "justification": "brief reason", "rating": 57.0}
]

Here are the labels to assess:
{labels_json}
```

(The final line is replaced by a batch of labels formatted as a JSON list of strings.)

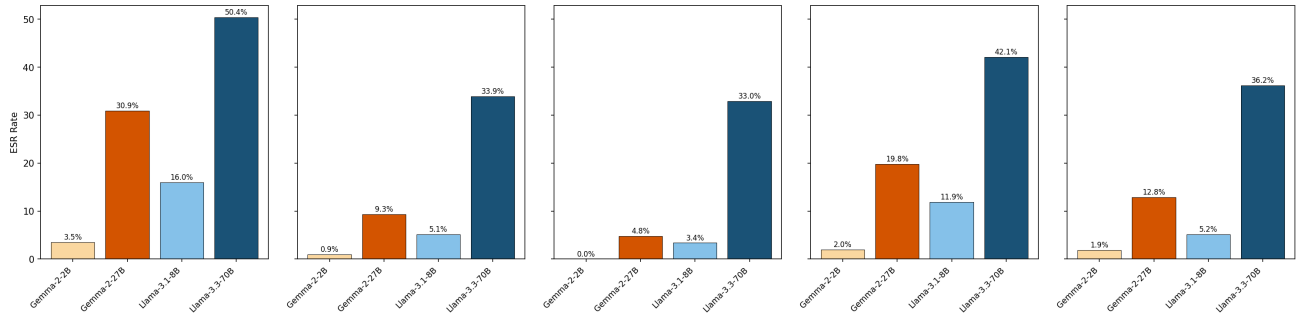

Figure 9. **Cross-judge ESR rate.** ESR rate by target model and judge (1,000 responses, stratified sampled). Llama-3.3-70B shows the highest ESR rates across all judges, substantially higher than other models.

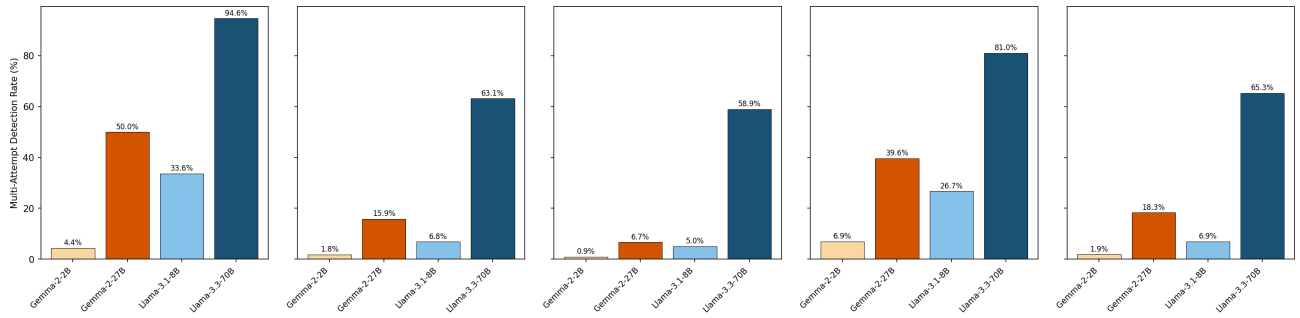

Figure 10. **Cross-judge multi-attempt rate.** Percentage of responses containing multiple attempts, by target model and judge (1,000 responses, stratified sampled). Llama-3.3-70B shows the highest multi-attempt rates across all judges, substantially higher than other models.

A.2.2. JUDGE MODELS

To validate the robustness of our ESR findings, we conducted a cross-judge analysis using four additional judge models: GPT-5-Mini, Qwen3-32B, Claude 4.5 Haiku, and Gemini-2.5-Flash. We sampled 1,000 responses from our experiment results and regraded them with each judge model, comparing these scores against our original Claude 4.5 Sonnet judge scores.

Sampling methodology and interpretation. Our sampling strategy was designed to enable meaningful cross-judge comparisons while avoiding the computational cost of regrading all tens of thousands of experiment trials. We used stratified sampling that (1) included all multi-attempt responses from each target model, and (2) ensured at least 100 samples per target model. This non-uniform sampling deliberately oversamples multi-attempt responses, which are the cases where judges must agree on both attempt segmentation and score improvement to validate ESR findings. *As a result, the absolute values shown in Figure 9 should not be interpreted as population-level ESR rates*, as they are inflated by the oversampling of multi-attempt cases. However, the relative comparisons between target models within each judge panel, and between judges for the same target model, remain valid and informative. The key finding is that all judges consistently rank Llama-3.3-70B as having substantially higher multi-attempt rates than other models.

The results demonstrate strong inter-judge agreement across multiple metrics. Agreement on multi-attempt detection is high, with judges agreeing on whether a response contains multiple attempts 90–96% of the time. For responses where both judges detected multiple attempts, agreement on ESR direction (whether scores improved) ranged from 90–96%.

Most importantly, as shown in Figures 9 and 10, all five judges agree on the relative ranking of target models: Llama-3.3-70B consistently shows the highest ESR rate across all judges. This consistency across judge models from different providers (OpenAI, Alibaba, Anthropic, Google) provides strong evidence that ESR is a robust phenomenon reflecting genuine model behavior rather than an artifact of any particular judge's evaluation methodology.

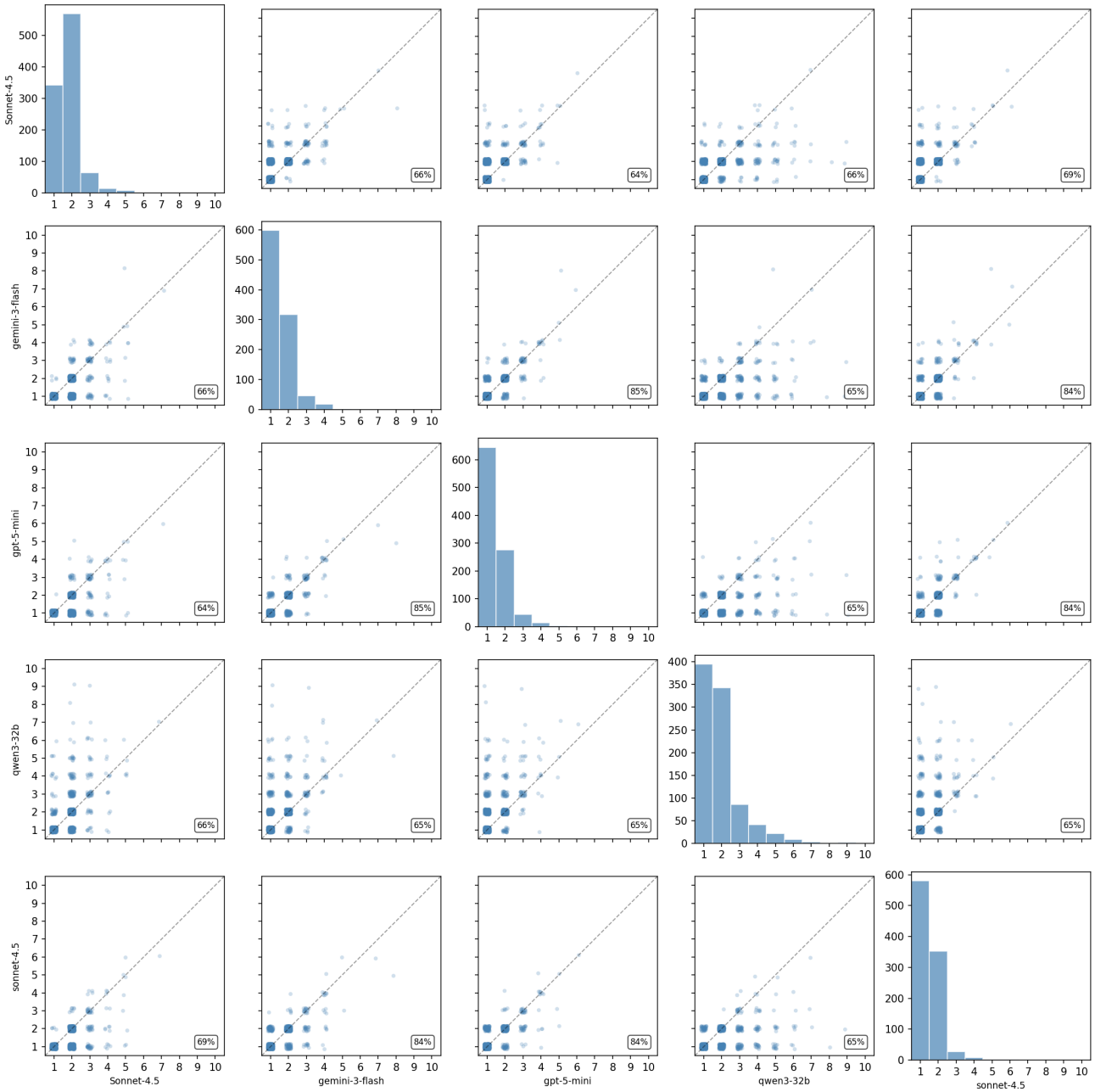

Figure 11. **Inter-judge agreement on number of attempts.** Facet grid showing pairwise agreement between judges on the number of attempts detected in each response (1,000 responses). Diagonal panels show each judge's distribution of attempt counts; off-diagonal panels show scatter plots with exact agreement percentages. Judges show high agreement on attempt segmentation despite using different underlying models.