Supplementary Material: Zero-Shot Prompting Baseline

Setup

We evaluate a frozen `openai/gpt-oss-20b` model in a zero-shot, multi-label setting on nine NGSS-aligned science tasks. For each dataset, we build an *instruction preamble* by concatenating the task prompt, a concise exemplar response, and the analytic rubric. For every rubric element (label), we issue a separate binary query:

You are an expert classifier. Answer strictly "Yes" or "No".
Text: ⟨student response⟩
Question: Does the label "⟨label name⟩" apply?
Answer:

We compute $P(\text{Yes})$ by summing next-token softmax mass over tokenizer IDs corresponding to {"Yes", "yes", "YES", "1", "True", "true"} and normalize by the mass over the union of {Yes, No} token sets. A fixed threshold of 0.5 converts probabilities to binary predictions. We report Micro-/Macro-F1, Exact Match at threshold 0.5 (denoted *thr**), Cohen's κ averaged over labels, and mean Per-Label Accuracy.

Rubric Integration Example (Task 3: Gas-Filled Balloons)

This example mirrors the presentation in Section but shows precisely how we embed the scoring rubric into the zero-shot prompt.

Task prompt (abridged). Alice tested four gases (A–D), recording flammability, density, and volume under identical conditions. *Question:* Which gases, if any, could be the same? Use the table to explain.

Exemplar rationale (abridged). Gases A and D could be the same because both are flammable and have the same density (0.089 g/L); these are identifying properties.

Table 6: Scoring rubric for Task 3 (Gas-Filled Balloons), formatted to fit a single column.

ID	Perspective	Description
E1	DCI	Student states that Gases A and D could be the same substance.
E2	SEP+CCC	Supports the claim by pointing out that A and D have the same *flammability*.
E3	SEP+CCC	Supports the claim by pointing out that A and D have the same *density*.
E4	DCI	Indicates that *flammability* is a characteristic property for identifying substances.
E5	DCI	Indicates that *density* is a characteristic property for identifying substances.

Analytic rubric (embedded as plain text in the preamble).

Per-label question instantiation. Each rubric element becomes its own Yes/No query by substituting the label text:

Question: Does the label "E2: A and D have the same flammability (pattern across columns)" apply?
Answer:

This produces one probability per label, later thresholded at 0.5.

Results

Table 7 summarizes zero-shot performance across all tasks. As expected, some items exhibit high mean per-label accuracy but low F1 (e.g., *dry_ice_model*), reflecting label imbalance and conservative predictions. For *gas_filled_balloons*, Cohen's κ is undefined (NaN) due to degenerate predictions on at least one label.

Table 7: Zero-shot results with `gpt-oss-20b` (frozen).

Dataset	**Cohen's κ @0.5**	**Macro F1 @0.5**	**Micro F1 @0.5**	**Mean Per-label Acc @0.5**
anna_vs_carla	0.1708	0.2261	0.1650	0.735375
breaking_down_hydrogen_peroxide	0.1624	0.2850	0.2908	0.700600
carlos_javier_atomic_model	0.0561	0.1143	0.0951	0.603340
dry_ice_model	0.0000	0.0000	0.0000	0.792567
gas_filled_balloons	0.0000	0.0000	0.0000	0.333333
layers_in_test_tube	0.0000	0.5007	0.5281	0.358760
model_for_making_water	0.0937	0.2080	0.2006	0.734280
namis_careful_experiment	0.0000	0.0000	0.0000	0.566067
natural_sugar	0.0119	0.2743	0.2783	0.605040
Average	**0.0543**	**0.1787**	**0.1731**	**0.6033**