AI MODEL DEVELOPMENT

Group 4 : Dhiksha Rathis, Shreya Verma

PHASE 1 - PROMPT KIT

1. Overview

This prompt kit contains structured prompts for two research tasks:

- **Task 1:** Claim-Evidence Extraction: Testing grounding and citation accuracy
- **Task 2:** Cross-Source Synthesis: Testing comparative analysis capability

Each task includes:

- **Prompt A (Baseline):** Minimal instruction to test default model behavior
- **Prompt B (Structured):** Enhanced with guardrails, format constraints, and uncertainty handling

Models: Claude Opus 4.5, Claude Sonnet 4.5, GPT-5, Gemini 3

2. Task 1: Claim-Evidence Extraction

2.1 Purpose :

Extract specific claims about LLM carbon emissions with verbatim supporting evidence and traceable citations. This task tests:

- Grounding accuracy: Does the model stick to source content?
- Citation fidelity: Are citations correctly formatted and traceable?
- Uncertainty handling: Does the model acknowledge when evidence is insufficient?

2.2 Prompt 1A: Claim-Evidence Extraction (BASELINE)

Extract 5 claims about carbon emissions or energy consumption from this text. For each claim, provide the direct quote or evidence that supports it.

2.3 Prompt 1B: Claim-Evidence Extraction (STRUCTURED)

You are a research assistant extracting claims for a systematic review on LLM carbon footprints. TASK: Extract exactly 5 claims with supporting evidence from the provided text.

2.3.1 DOMAIN FOCUS: Claims should relate to:

- Energy consumption of ML/LLM systems
- Carbon emissions from training or inference
- Measurement methodologies and their limitations
- Hardware efficiency and carbon intensity factors
- Lifecycle emissions (embodied + operational)

2.3.2 REQUIRED OUTPUT FORMAT:

Produce a table with exactly 5 rows and 3 columns

Claim	Direct Quote or Snippet	Citation

2.3.3 CRITICAL RULES:

1. EVIDENCE GROUNDING:

- The "Direct Quote or Snippet" MUST be copied VERBATIM from the source text.
- Use quotation marks for exact quotes.
- If paraphrasing is necessary, indicate with [paraphrased: ...].

2. CLAIM-EVIDENCE ALIGNMENT:

- Each claim MUST be DIRECTLY supported by its paired evidence.
- Do NOT pair a claim with tangentially related evidence.

3. CITATION ACCURACY:

- Use the EXACT source_id and chunk_id provided in the input.
- Do NOT invent, modify, or guess citation identifiers.

4. UNCERTAINTY HANDLING:

- If the text contains FEWER than 5 relevant claims, explicitly state: "Only N claims found in this text."
- Do NOT fabricate claims to reach 5.

5. PRESERVE HEDGING:

- If the source uses hedging language (e.g., "may," "suggests," "approximately"), PRESERVE it in your claim.
- Do NOT convert "may reduce" to "reduces."

6. NO FABRICATION:

- Do NOT generate claims not present in the text.
- Do NOT generate quotes not present in the text.
- If uncertain, skip the claim rather than guess.

3. Task 2: Cross-source synthesis

3.1 Purpose :

Compare multiple sources on a specific topic to identify methodological agreements and disagreements. This task tests:

- Comparative reasoning: Can the model identify genuine differences?
- Multi-source citation: Are both sources correctly referenced?
- Nuanced analysis: Does the model capture subtle methodological distinctions?

3.2 Prompt 2A: Cross-source synthesis (BASELINE)

Compare these two sources on the topic of "Green Computing in AI (Sustainability)". Identify where they agree and where they disagree.

3.3 Prompt 2B: Cross-source synthesis (STRUCTURE)

You are synthesizing research on LLM carbon measurement methodologies for a systematic literature review. TASK: Compare two sources on the topic "{topic}" and identify specific points of agreement and disagreement.

3.3.1 REQUIRED OUTPUT FORMAT:

Produce a table with exactly 5 columns and 3 rows

Aspect	Agreement	Disagreement	Evidence (Source 1)	Evidence (Source 2)

3.3.2 CRITICAL RULES:

1. EVIDENCE REQUIREMENT:

- EVERY cell in the Evidence columns MUST cite specific text from that source.
- Use format: "[quote or close paraphrase]"
- Do NOT make claims about a source without textual evidence.

2. AGREEMENT DEFINITION:

- "Agreement" = Both sources make compatible or mutually supporting claims.
- Similarity is NOT the same as agreement — they must address the same point.

3. DISAGREEMENT DEFINITION:

- "Disagreement" = Sources explicitly contradict each other, OR
- Sources use incompatible methods/assumptions for the same measurement.
- If NO disagreement exists for an aspect, write "No disagreement found."

4. BALANCED COVERAGE:

- Both sources must be cited in each row.
- Do NOT favor one source over the other.

5. NO INFERENCE BEYOND SOURCES:

- Do NOT infer what a source "would say" — only report explicit content.
- If a source does not address an aspect, write "Not addressed in this source."

4. Guardrails & Rubric

Guardrail Type	Research Justification
Verbatim quotes	Prevents the “helpful modification” behavior where models round numbers or paraphrase (observed in GPT-5 baseline)
Explicit citation format	Neither model reliably provides citations without specification (Henderson et al., 2020 on reporting standards)
Uncertainty acknowledgment	Aligns with Green AI principles of honest reporting (Schwartz et al., 2020)
No fabrication rules	Directly addresses hallucination risk in research contexts
Preserved hedging	Maintains scientific accuracy, critical for carbon estimates with uncertainty

5. Evaluation Criteria: Scoring Rubric (1-4 Scale)

Score	Label	Definition
4	Excellent	Correctly grounded; citations accurate; structure perfect; uncertainty stated when appropriate
3	Good	Mostly correct; minor omissions OR minor citation/format issues; usable with light editing
2	Fair	Partially correct; key omissions OR weak grounding OR vague citations; needs significant revision
1	Poor	Not usable; hallucinated claims, fabricated citations, or fails required structure

6. References

[1] Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and Policy Considerations for Deep Learning in NLP. ACL 2019.

[2] Luccioni, A.S., Viguier, S., & Ligozat, A.L. (2022). Estimating the Carbon Footprint of BLOOM. arXiv:2211.02001.

[3] Patterson, D., et al. (2021). Carbon Emissions and Large Neural Network Training. arXiv:2104.10350.

[4] Schwartz, R., Dodge, J., Smith, N.A., & Etzioni, O. (2020). Green AI. Communications of the ACM.

[5] Henderson, P., et al. (2020). Towards the Systematic Reporting of the Energy and Carbon Footprints of ML. arXiv:2002.05651.