AI MODEL DEVELOPMENT 1st February 2026

Group 4 : Dhiksha Rathis, Shreya Verma

PHASE 1 - ANALYSIS MEMO

Domain: Green AI, sustainable computing

Main Research Question: How do we accurately measure and compare the carbon footprint of different LLMs across their lifecycle?

1. Executive Summary

This memo analyzes 32 evaluation runs across four models (Claude Opus 4.5, Claude Sonnet 4.5, GPT-5, Gemini 3), two tasks (Claim-Evidence Extraction, Cross-Source Synthesis), and two prompt variants (Baseline, Structured).

Key Findings:

Finding	Evidence
Structured prompts improve all models by **30% average**	Baseline: 3.11 → Structured: 3.99
Claude Opus 4.5 has strongest baseline performance	3.75 avg; zero failure tags
GPT-5 is most prompt-sensitive	+36% improvement; but highest baseline failures
Gemini 3 struggles with format compliance	4 STRUCT failures at baseline
Citation accuracy is **universal weakness**	All models except Opus need explicit rules

2. Failure Pattern Analysis

Pattern 1: Citation Fabrication (FABCITE)

Attribute	Detail
Frequency	1 occurrence
Affected Model	GPT-5
Task	Claim-Evidence Extraction (TC1B)
Example	Model invented chunk_07 when chunk_05 was provided
Root Cause	Model generates plausible-looking identifiers without verification
Fix Applied	“Use the EXACT source_id and chunk_id provided. Do NOT invent identifiers.”
Result	0 FABCITE failures with structured prompt

Pattern 2: Format Non-Compliance (STRUCT)

Attribute	Detail
Frequency	6 occurrences
Affected Models	Gemini 3 (4), Sonnet 4.5 (2)
Task	Both tasks
Example	Produced bullet lists instead of requested tables

Root Cause	Models default to prose/list format without explicit structure
Fix Applied	Explicit markdown table template with example row
Result	0 STRUCT failures with structured prompt

Pattern 3: Overconfidence (OVERCONF)

Attribute	Detail
Frequency	1 occurrence
Affected Model	GPT-5
Task	Cross-Source Synthesis (TC2A)
Example	Changed “estimates suggest” → “studies show”
Root Cause	Model simplifies for readability, removing epistemic hedging
Fix Applied	“PRESERVE hedging language (may, suggests, approximately)”
Result	0 OVERCONF failures with structured prompt

Pattern 4: Generic Claims (GENERIC)

Attribute	Detail
Frequency	2 occurrences
Affected Model	GPT-5
Task	Both tasks
Example	“Training uses significant energy” (no specific numbers)
Root Cause	Model summarizes at high level, losing paper-specific details
Fix Applied	“Include SPECIFIC numbers where reported (e.g., X kWh, Y tonnes CO2)”
Result	0 GENERIC failures with structured prompt

Pattern 5: Claim-Evidence Misalignment (MISALIGN)

Attribute	Detail
Frequency	1 occurrence
Affected Model	Gemini 3
Task	Cross-Source Synthesis (TC2A)

Example	Attributed Patterson's claim to Strubell
Root Cause	Multi-source context confuses attribution
Fix Applied	"EVERY cell in Evidence columns MUST cite specific text from THAT source"
Result	0 MISALIGN failures with structured prompt

3. Model Comparison

3.1 Quantitative Summary

Model	Baseline Avg	Structured Avg	Δ Improvement	Failure Count (Baseline)
Claude Opus 4.5	**3.75**	4.00	+6.7%	**0**
Claude Sonnet 4.5	3.00	4.00	+33.3%	2
GPT-5	2.94	4.00	+36.1%	4
Gemini 3	2.75	3.94	+43.3%	5

3.2 Qualitative Assessment

Dimension	Opus 4.5	Sonnet 4.5	GPT-5	Gemini 3
Default Grounding	Excellent	Strong	Moderate	Moderate
Citation Behavior	Good baseline	Needs prompting	Prone to fabrication	Needs prompting
Format Compliance	Excellent	Moderate	Excellent	Weak
Hedging Preservation	Excellent	Good	Moderate	Good
Prompt Responsiveness	Low (already good)	High	Very High	Very High

3.3 Model Strengths & Weaknesses

Claude Opus 4.5 -

- Best baseline grounding (zero failures)
- Natural uncertainty acknowledgment
- Consistent citation behavior
- Higher cost tier

Claude Sonnet 4.5

- Matches Opus with structured prompts
- Cost-effective
- Format compliance issues at baseline
- Needs explicit citation rules

GPT-5

- Excellent format compliance
- Highly responsive to constraints
- Citation fabrication risk
- Loses hedging language
- Tends toward generic summaries

Gemini 3

- Good grounding once structured
- Handles multi-source well
- Persistent format issues
- Attribution confusion

4. Phase 2 Design Recommendations

4.1 Model Selection

Role	Model	Rationale
Primary	Claude Opus 4.5	Best grounding; lowest failure rate; most reliable for research
Fallback	Claude Sonnet 4.5	Matches Opus with structured prompts; lower cost
Avoid	GPT-5 (without heavy constraints)	Citation fabrication risk too high

4.2 Retrieval Pipeline

Component	Specification	Rationale
Chunk size	512 tokens	Sufficient for verbatim quote extraction
Overlap	128 tokens	Preserve context across boundaries
Metadata	source_id, chunk_id, section	Enable precise citations
Top-k	5-8 chunks	Balance relevance and context window

5. Limitations

Limitation	Impact	Mitigation
4 test cases	May miss edge cases	Expand to 20+ in Phase 2
Single evaluator	Potential scoring bias	Document rubric; inter-rater check
English sources only	Limits generalizability	Acknowledge scope
API versions may change	Results may not replicate	Document exact model versions

6. Conclusion

Phase 1 evaluation demonstrates that:

1. **Prompt engineering is essential** — 30% average improvement across all models
2. **Model choice matters for baseline reliability** — Opus 4.5 requires less prompt engineering
3. **Citation handling requires explicit design** — No model defaults to research-grade citations
4. **Structured prompts eliminate most failure modes** — 11 baseline failures → <1 structured

These findings provide a strong foundation for Phase 2's RAG system, where the structured prompt templates and citation requirements will be directly integrated.