AI MODEL DEVELOPMENT

Group 4 : Dhiksha Rathis, Shreya Verma

PHASE 1 - ANALYSIS MEMO

1. Summary

32 runs evaluated across 4 models, 2 tasks, 2 prompt variants. Structured prompts improve all models by 30% average. Opus 4.5 leads baseline performance (3.75); all models converge near 4.0 with structured prompts.

2. Failure Patterns

Pattern	Model	Example	Fix
FABCITE	GPT-5	Invented chunk_07 when chunk_05 given	“Use EXACT IDs : do NOT invent”
STRUCT	Gemini, Sonnet	Lists instead of tables	Explicit markdown template
OVERCONF	GPT-5	“may reduce” became “reduces”	“Preserve hedging language”
GENERIC	GPT-5	“Training uses energy” (no numbers)	“Include specific numbers”
MISALIGN	Gemini	Wrong source attribution	“Cite specific text from THAT source”

3. Model Comparison

Performance

Model	Baseline	Structured	Delta	Rank
Opus 4.5	3.75	4.00	+6.7%	1st
Sonnet 4.5	3.00	4.00	+33%	3rd
GPT-5	2.94	4.00	+36%	4th
Gemini 3	2.75	3.94	+43%	4th

Characteristics

Dimension	Opus 4.5	Sonnet 4.5	GPT-5	Gemini 3
Baseline grounding	Excellent	Strong	Moderate	Moderate
Citation accuracy	High	Medium	Low	Medium
Format compliance	Excellent	Moderate	Excellent	Weak
Prompt sensitivity	Low	High	Very High	Very High

4. Key Findings

1. **Opus 4.5 dominates baseline**: Only model with zero failures at baseline
2. **All models converge with structure**: Proper guardrails eliminate performance gaps
3. **GPT-5 most prompt-sensitive**: Highest improvement (+36%) but most baseline failures
4. **Gemini weakest on format**: 4 STRUCT failures; needs explicit templates
5. **Citation accuracy is universal gap**: All models except Opus need explicit rules

5. What Worked

Technique	Result
Explicit table templates	0 STRUCT failures
“VERBATIM” instruction	0 quote modifications
“Do NOT invent” rule	0 FABCITE failures
“Only N found” fallback	0 forced hallucinations
Dual-citation requirement	Balanced synthesis

7. References

[1] Luccioni et al. (2022). Estimating the Carbon Footprint of BLOOM. arXiv:2211.02001.

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

[3] Strubell et al. (2019). Energy and Policy Considerations for Deep Learning in NLP. ACL 2019.