AI MODEL DEVELOPMENT

1st February 2026

Group 4 : Dhiksha Rathis, Shreya Verma

PHASE 1 - EVALUATION SHEET

Evaluation Configuration

Task ID	Task Name	Prompt A (Baseline)	Prompt B (Structured)
T1	Claim-Evidence Extraction	CLAIM-EV BASELINE-V1	CLAIM-EV STRUCTURED-V1
T2	Cross-Source Synthesis	SYNTH-BASELINE V1	SYNTH-STRUCTURD-V1

Test Cases

Case ID	Task	Source(s)	Focus
TC1A	T1	Strubell et al. (2019)	Training carbon estimates
TC1B	T1	Luccioni et al. (2022)	BLOOM lifecycle data
TC2A	T2	Strubell + Patterson	Measurement methods comparison
TC2B	T2	Luccioni + Patterson	Hardware/carbon intensity assumptions

Total Runs

32 runs = 4 models × 2 prompt variants × 4 test cases

1. Scoring Rubric

1.1 Score Definitions (1 - 4 Scale)

Score	Label	Definition	Example
4	Excellent	Fully grounded; citations accurate; perfect structure; uncertainty acknowledged	All claims traceable to source; exact quotes used
3	Good	Mostly correct; minor omissions or format issues	4/5 claims correct; one minor citation error
2	Fair	Partial accuracy; weak grounding; vague citations	Claims correct but evidence doesn't directly support
1	Poor	Unusable; hallucinations; fabricated citations; structural failure	Made-up quotes; wrong source_ids

1.2 Evaluation Dimensions

Dimension	What It Measures	Weight
Groundedness	Are claims supported by source text?	25%
Structure	Does output match the required format?	25%
Citation Accuracy	Are citations correct and traceable?	25%
Usefulness	Is output usable for research?	25%

1.3 Failure Tags

Tag	Code	Definition
Hallucination	HALLUC	Claim fabricated; not in source
Fabricated Citation	FABCITE	Citation ID invented or wrong
Misalignment	MISALIGN	Claim-evidence pair mismatch
Overconfidence	OVERCONF	Hedging removed ("may" to "does")
Structure Failure	STRUCT	Wrong output format
Generic	GENERIC	Too vague to verify
Truncation	TRUNCATE	Incomplete output

2. Task 1: Claim-Evidence Extraction

2.1 Test Case TC1A: Strubell et al. (2019)

Run	Model	Prompt	Ground	Struct	Cite	Useful	Overall	Evidence/Notes
1	Claude Opus 4.5	Baseline	4	4	3	4	3.75	Correct claims; missing chunk_id in citations
2	Claude Opus 4.5	Structured	4	4	4	4	4.00	Perfect: “Training BERT on GPU is estimated to require 1,507 lbs of CO2” (strubell2019, chunk_03)
3	Claude Sonnet 4.5	Baseline	4	4	2	3	3.25	Good grounding; no citations provided
4	Claude Sonnet 4.5	Structured	4	4	4	4	4.00	Full compliance with format
5	GPT-5	Baseline	3	4	2	3	3.00	GENERIC - “Training uses significant energy” too vague
6	GPT-5	Structured	4	4	4	4	4.00	Responded well to constraints
7	Gemini 3	Baseline	3	3	2	3	2.75	STRUCT - Used bullet list instead of table
8	Gemini 3	Structured	4	4	3	4	3.75	Minor citation format inconsistency

2.2 Test Case TC1B: Luccioni et al. (2022)

Run	Model	Prompt	Ground	Struct	Cite	Useful	Overall	Evidence/Notes
9	Claude Opus 4.5	Baseline	4	4	3	4	3.75	Captured lifecycle distinction; citation incomplete
10	Claude Opus 4.5	Structured	4	4	4	4	4.00	“Total lifecycle: ~50.5 tonnes CO2eq” correctly extracted
11	Claude Sonnet 4.5	Baseline	4	4	2	3	3.25	Strong content; missing citations
12	Claude Sonnet 4.5	Structured	4	4	4	4	4.00	Excellent compliance
13	GPT-5	Baseline	3	4	1	3	2.75	FABCITE - Invented “chunk_07” when chunk_05 was provided
14	GPT-5	Structured	4	4	4	4	4.00	Corrected with explicit rules
15	Gemini 3	Baseline	3	3	2	3	2.75	STRUCTNumbered list format instead of table
16	Gemini 3	Structured	4	4	4	4	4.00	Full compliance achieved

3. Task 2: Cross-Source Synthesis

3.1 Test Case TC2A: Strubell + Patterson (Measurement Methods)

Run	Model	Prompt	Ground	Struct	Cite	Useful	Overall	Evidence/Notes
17	Claude Opus 4.5	Baseline	4	4	3	4	3.75	Identified key method differences; informal citation
18	Claude Opus 4.5	Structured	4	4	4	4	4.00	Perfect table: Agreement on GPU-hours metric; Disagreement on utilization assumptions
19	Claude Sonnet 4.5	Baseline	3	3	2	3	2.75	STRUCT - Prose format instead of table
20	Claude Sonnet 4.5	Structured	4	4	4	4	4.00	Correct format with dual citations
21	GPT-5	Baseline	3	4	2	3	3.00	OVERCONF - Changed “estimates suggest” to “studies show”
22	GPT-5	Structured	4	4	4	4	4.00	Preserved hedging when prompted
23	Gemini 3	Baseline	3	3	2	3	2.75	MISALIGN - Cited wrong source for one claim
24	Gemini 3	Structured	4	4	4	4	4.00	Correct attribution achieved

3.2 Test Case TC2B: Luccioni + Patterson (Hardware & Carbon Intensity)

Run	Model	Prompt	Ground	Struct	Cite	Useful	Overall	Evidence/Notes
25	Claude Opus 4.5	Baseline	4	4	3	4	3.75	Strong comparison; informal citations
26	Claude Opus 4.5	Structured	4	4	4	4	4.00	Key insight: Luccioni includes embodied; Patterson focuses operational
27	Claude Sonnet 4.5	Baseline	3	3	2	3	2.75	STRUCT - Bullet points instead of table
28	Claude Sonnet 4.5	Structured	4	4	4	4	4.00	Full format compliance
29	GPT-5	Baseline	3	4	2	3	3.00	GENERIC- “Both discuss carbon”, too vague
30	GPT-5	Structured	4	4	4	4	4.00	Specific aspects identified
31	Gemini 3	Baseline	3	3	2	3	2.75	STRUCT- Format non-compliance
32	Gemini 3	Structured	4	4	4	4	4.00	Correct table format

4. Performance (Structured Prompts)

Metric	Opus 4.5	Sonnet 4.5	GPT-5	Gemini 3	Average
Groundedness	4.00	4.00	4.00	4.00	4.00
Structure	4.00	4.00	4.00	4.00	4.00
Citation	4.00	4.00	4.00	3.75	3.94
Usefulness	4.00	4.00	4.00	4.00	4.00

Best Output (Structured) - Opus 4.5

Source: Strubell et al. (2019), chunk_03
Score: 4.00

Output:

Claim	Direct Quote or Snippet	Citation
BERT training produces substantial CO_2 emissions	"Training BERT on GPU is estimated to require 1,507 lbs of CO_2"	(Strubell et al., 2019, chunk_03)
Neural architecture search dramatically multiplies emissions	"The neural architecture search required to develop the Evolved Transformer consumed approximately 284,000 GPU hours, producing approximately 626,155 lbs of CO_2"	(Strubell et al., 2019, chunk_03)
Training costs can rival car lifetime emissions	"Training a single NLP model can produce as much carbon as five cars over their entire lifetimes"	(Strubell et al., 2019, chunk_03)
Energy estimates use GPU power calculations	"We use the formula: Energy (kWh) = GPU power (W) × Training time (hours) / 1000"	(Strubell et al., 2019, chunk_03)

Authors acknowledge estimation uncertainty	“Our analysis relies on estimates rather than direct measurements”	(Strubell et al., 2019, chunk_03)

Worst Output (Baseline) - GPT 5

1. Claim: BLOOM training consumed approximately 400,000 kWh of energy
   Evidence: The model used substantial electricity during training
   Citation: (luccioni2022, chunk_07) ← FABRICATED: chunk_05 was provided

2. Claim: France's grid has low carbon intensity around 50 gCO2/kWh
   Evidence: Training in France reduced emissions
   Citation: (luccioni2022, chunk_07) ← FABRICATED + IMPRECISE: actual value is 57 gCO2/kWh

5. Key Findings

Finding	Evidence
Opus 4.5 has strongest baseline grounding	Only model with 0 failure tags at baseline; 3.75 avg without structured prompts
All models converge with structured prompts	Structured avg: Opus 4.00, Sonnet 4.00, GPT-5 4.00, Gemini 3.94
GPT-5 most prompt-sensitive	Largest improvement (+36%) but also highest baseline failure count (4)
Gemini 3 weakest on format compliance	4 STRUCT failures; consistently used lists instead of tables
Citation accuracy is universal weakness	All models except Opus needed explicit citation format rules