Personal Research Project – Phase 1

Lindberg Simpson – lindbers

Domain: Remote work and Productivity

Main Research Question:

How does remote work impact productivity at individual, organizational, and industry levels?

Sub questions:

1. What evidence links remote work to changes in individual productivity and work-life integration?
2. Does remote or hybrid work correlate with firm-level productivity, output, or cost changes?
3. How is productivity defined and measured in remote work research? Does it differ by industry or job type?
4. What role do management practices, technological infrastructure, and job characteristics play in shaping productivity outcomes?
5. What is the measurable impact of "commute-time reclamation" on labor output and employee discretionary effort?
6. Which management styles (e.g., outcome-based vs. surveillance-based) show the highest correlation with high-performer retention and output?

Scope:

Inclusions:

- Evidence from empirical research, including peer-reviewed articles, government statistics, and technical reports focused on the association between remote work and productivity.
- Both quantitative and qualitative measures of productivity, including output measures (e.g., performance metrics, total factor productivity) and perceptual or self-reported productivity from surveys.
- Focus Areas: Knowledge-work sectors (Tech, Finance, Professional Services) where remote work is highly feasible.

Exclusions:

- Non-credible opinion pieces lacking empirical support (e.g., purely anecdotal blog posts without systematic data).
- Sources focused solely on unrelated remote work topics, such as social effects that do not intersect with productivity measures unless they are tied back to productivity outcomes.
- Industry marketing material and unverified surveys that do not provide methodological transparency or peer-review validation.
- Ineligible Industries: Manual labor, manufacturing, or healthcare roles that require physical presence.

Chosen Tasks from the Task Menu: Paper Triage & Cross-Source Synthesis

Chosen Models: ChatGPT 5.2 Plus & Gemini 3 Pro

Prompt Kit:

Paper Triage Task

Prompt A:

Summarize the following paper about remote work. Tell me the contribution, method, data used, findings, and limitations.

Prompt B:

You are performing PAPER TRIAGE for a research corpus on remote work and productivity.

INPUT:

- Paper text or excerpt (with chunk_ids).

TASK:

Produce a structured summary with EXACTLY the following five fields:

1. Contribution – What new knowledge this paper adds.

2. Method – Study design and analytical approach.

3. Data – Data source(s), sample size, and time period.

4. Findings – Empirical results supported by the data.

5. Limitations – Explicit weaknesses stated by the authors OR implied by the method/data.

CONSTRAINTS:

- Use only information present in the provided text.

- Cite at least one chunk_id per field.

- If a field cannot be answered from the text, write: "Not specified in the provided text."

- Do NOT infer causality unless the paper explicitly claims causal identification.

- Keep each field to 2–4 sentences.

OUTPUT FORMAT:

- Five bullet points labeled exactly as the field names above.

Why these constraints exist:

- Exact fields: enables cross-paper comparison
- chunk_id citations: forces grounding, prevents hallucination
- “Not specified” rule: exposes corpus gaps instead of guessing
- Causality guardrail: critical in remote-work research
- Length limits: discourages vague or padded summaries

Cross-Source Synthesis Task

Prompt A:

Compare the following sources on remote work and productivity.

Summarize where they agree and disagree.

Prompt B:

You are performing CROSS-SOURCE SYNTHESIS for a research corpus on remote work and productivity.

INPUT:

- Multiple sources, each with source_id and chunk_ids.

TASK:

Produce a comparison table with the following columns:

1. Agreement

2. Disagreement

3. What evidence supports each side

CONSTRAINTS:

- Each row must reference at least TWO different sources.

- Every claim must be supported with citations in the form (source_id, chunk_id).

- Clearly distinguish differences caused by:

  a) Level of analysis (individual, firm, macro)

  b) Measurement type (self-reported vs objective)

  c) Study design (causal vs correlational)

- Do NOT resolve disagreements unless the sources explicitly reconcile them.

- If evidence is insufficient to support a conclusion, state that explicitly.

OUTPUT FORMAT:

- Markdown table only (no prose before or after).

Why these constraints exist:

- Two-source minimum: prevents single-paper “synthesis”
- Evidence column: makes disagreement explainable, not rhetorical
- Measurement/level distinctions: core to this domain
- No forced resolution: avoids artificial consensus
- Table-only output: machine-checkable structure for Phase 2

Evaluation Sheet

Run	Task	Test Case	Prompt	Model	Score	Notes
1	Paper Triage	Cisco Global Hybrid Work Study 2025	A	ChatGPT 5.2 Plus	3	Provides structured 5-field summary with reasonable content. Citations to document sections are present but could be more specific (e.g., "Cisco study" without chunk IDs). Minor format inconsistencies in how findings are presented.
2	Paper Triage	Cisco Global Hybrid Work Study 2025	B	ChatGPT 5.2 Plus	4	Excellent structured output with all required fields. Includes specific citations with document references. Appropriately notes limitations and acknowledges when evidence is mixed (e.g., productivity effects vary by context).
3	Paper Triage	Cisco Global Hybrid Work Study 2025	A	Gemini 3 Pro	3	Provides a highly detailed and readable summary, including specific stats. It loses points for failing to provide specific inline citations to verify the origin of each claim.

4	Paper Triage	Cisco Global Hybrid Work Study 2025	B	Gemini 3 Pro	4	Comprehensive 5-field summary with proper structure. Citations include source identifiers. Correctly represents the complexity of findings and notes where evidence shows varied outcomes across different measures.
5	Paper Triage	The-Rise-of-Remote-Work	A	ChatGPT 5.2 Plus	3	Correctly identifies the shift from negative to positive productivity perceptions. It provides a clean summary but lacks the rigorous citation structure required for a score of 4
6	Paper Triage	The-Rise-of-Remote-Work	B	ChatGPT 5.2 Plus	4	Effectively used Prompt B's constraints to provide a concise, 5-field summary with specific page and paragraph citations . It correctly avoided inferring causality, strictly sticking to reported perceptions.

7	Paper Triage	The-Rise-of-Remote-Work	A	Gemini 3 Pro	3	Structured format is followed with all 5 fields. Citations are general rather than specific to chunks. Mostly accurate but misses some nuances about the productivity decline and focuses more on hours worked.
8	Paper Triage	The-Rise-of-Remote-Work	B	Gemini 3 Pro	4	Excellent structure with all fields completed thoroughly. Specific citations to tables and sections. Correctly identifies the productivity paradox and notes uncertainty about long-term effects and applicability to other contexts.
9	Cross-Source Synthesis	Exploring-the-Impact-of-Remote-Work-on-Productivity-A-Systematic-Review-ACTABOHEMIA.pdf & Hybrid-working-from-home-improves-retention.pdf	A	ChatGPT 5.2 Plus	2	Attempts the 5-row table format but citations are vague or missing chunk IDs. Some claims don't directly quote evidence. Mix of properly grounded and weakly supported statements.
10	Cross-Source Synthesis	Exploring-the-Impact-of-Remote-Work-on-Productivity-A-Systematic-Review-ACTABOHEMIA.pdf & Hybrid-working-from-home-improves-retention.pdf	B	ChatGPT 5.2 Plus	4	Proper table format with claim, direct quote, and citation including source and chunk identifiers. All 5 rows follow the required structure. Evidence is clearly quoted and properly attributed.

11	Cross-Source Synthesis	Exploring-the-Impact-of-Remote-Work-on-Productivity-A-Systematic-Review-ACTABOHEMIA.pdf & Hybrid-working-from-home-improves-retention.pdf	A	Gemini 3 Pro	3	Table structure is present with 5 rows. Citations reference sources but lack specific chunk identifiers. Most claims are grounded but some evidence snippets are paraphrased rather than direct quotes as required.
12	Cross-Source Synthesis	Exploring-the-Impact-of-Remote-Work-on-Productivity-A-Systematic-Review-ACTABOHEMIA.pdf & Hybrid-working-from-home-improves-retention.pdf	B	Gemini 3 Pro	4	Well-structured table with all required columns. Direct quotes are provided with specific citations including document and section identifiers. Claims are properly grounded in the evidence provided.
13	Cross-Source Synthesis	Supporting-the-productivity-and-wellbeing-of-remote-workers.pdf & Work-from-Home-and-Productivity.pdf	A	ChatGPT 5.2 Plus	2	Several rows lack proper direct quotes or have vague citations. Some claims are more interpretation than grounded extraction from the specific document.
14	Cross-Source Synthesis	Supporting-the-productivity-and-wellbeing-of-remote-workers.pdf & Work-from-Home-and-Productivity.pdf	B	ChatGPT 5.2 Plus	4	Excellent table structure with all required elements. Direct quotes are provided with specific citations to tables and sections. Claims are well-grounded and evidence is properly attributed with chunk identifiers.

15	Cross-Source Synthesis	Supporting-the-productivity-and-wellbeing-of-remote-workers.pdf & Work-from-Home-and-Productivity.pdf	A	Gemini 3 Pro	3	Table format is mostly correct with 5 rows. Citations are present but some lack specificity about chunk/section IDs. Most evidence is quoted but a few instances use paraphrase instead of direct quotes.
16	Cross-Source Synthesis	Supporting-the-productivity-and-wellbeing-of-remote-workers.pdf & Work-from-Home-and-Productivity.pdf	B	Gemini 3 Pro	4	Properly formatted table with claim, direct quote, and detailed citations. All 5 rows include specific source and section references. Evidence is directly quoted and accurately represents the document's findings

Analysis Memo:

The evaluation of Phase 1 results reveals a clear and consistent performance pattern across both models. Outputs generated using Prompt B, which incorporated structured constraints and guardrails, substantially outperformed those produced by Prompt A, the baseline prompt. While Prompt A often yielded readable and superficially coherent summaries, it repeatedly failed to meet the project's research-grade standard, particularly with respect to verifiable grounding and citation rigor. In contrast, Prompt B's explicit structural constraints significantly reduced model ambiguity, leading to higher citation accuracy, stronger adherence to academic conventions, and more reliable extraction of evidence from source materials. These improvements appeared consistently across tasks and models, indicating that prompt design, rather than model capability alone, was the dominant factor influencing output quality.

The most persistent failure mode under Prompt A was the presence of vague, incomplete, or missing citations. In multiple runs, models produced statistically accurate claims without providing granular, inline citations that tied each claim to a specific source location. For example, in the

Cisco study evaluation, Gemini 3 Pro correctly referenced a 72 percent mandate rate but failed to verify the provenance of individual claims, rendering the output unsuitable for academic use despite its apparent accuracy. A second failure pattern emerged in synthesis tasks involving conflicting evidence, particularly in productivity studies. When confronted with sources reporting opposing effects, such as a 13 percent productivity gain versus a 19 percent loss, models tended to prioritize narrative coherence over evidentiary precision. This resulted in weakly supported reconciliation statements that glossed over contradictions rather than explicitly addressing them.

These issues were reinforced by Prompt A's permissive phrasing, which suggested required components but did not enforce them. As a result, models frequently treated elements such as limitations or data descriptions as optional and omitted them without penalty. Prompt B's requirement to output exactly five predefined fields eliminated this ambiguity and led to substantially more complete and consistent responses. Prompt A also failed to elicit appropriate uncertainty calibration. Models often presented correlational findings as causal, neglected to distinguish between self-reported and objective measures, and did not indicate when required information was absent from the source text, further weakening research validity.

The Phase 1 findings directly informed the Phase 2 design decisions. Phase 2 will rely exclusively on structured prompts modeled after Prompt B, as baseline variants did not reliably enforce research-grade constraints and will therefore be excluded from further evaluation. Citation requirements will be strengthened, with chunk-level citations required for all claims and paraphrased statements without explicit grounding treated as failures. Uncertainty will be treated as a positive outcome rather than a deficiency, and explicit acknowledgments of insufficient or conflicting evidence will be rewarded, particularly in contested domains such as productivity research. Finally, synthesis tasks will require table-only outputs to reduce narrative drift, support automated checking, and align more closely with rubric-based scoring. Together, these design choices reflect a central conclusion from Phase 1, that output quality depends far more on precise, enforceable prompt constraints than on model sophistication alone.