Framing Brief:

Chosen Domain: Sleep patterns, quality, and their impact on overall well-being

Main Research Question: How do different sleep factors (duration, quality, timing, consistency) affect physical health, mental well-being, and cognitive performance, and what interventions most effectively improve sleep outcomes?

Sub Questions:

1. What is the relationship between sleep duration and mental health outcomes (anxiety, depression, stress)?
2. How does sleep quality (measured by sleep stages, interruptions) affect physical health markers?
3. What role does sleep consistency (regular schedule) play vs total sleep hours?
4. How do lifestyle factors (exercise, caffeine, screens, diet) affect sleep quality and duration?
5. What sleep interventions (CBT-I, sleep hygiene, supplements) show the strongest evidence for improving well-being?
6. How do individual differences (age, chronotype, health conditions) moderate sleep-wellbeing relationships?

These sub questions allow for exploration in multiple ways from quantitative associations to modifiable behaviors, treatment effectiveness, and personal factors.

Scope:
We'll include empirical research with quantitative findings, focus more on the adult population (18+), more recent publications that are peer-reviewed and from reputable sources, span multiple health domains, and explore diverse populations mostly through the USA.

We're excluding exploring specific clinical sleep disorders like narcolepsy, pediatric-only studies, non-human studies, opinion articles, and sleep caused by specific medical conditions, and any extreme cases. We aim to only explore avenues that are generalizable and will benefit the greater number of people.

Two Tasks: Claim–evidence extraction (CEE) and cross-source synthesis (CSS)

Prompt Kit:

Prompt Card 1:

Prompt name: CEE - Prompt A (Baseline)
Intent: Extract verifiable claims from empirical sleep research papers with supporting evidence with minimal guardrails.

Inputs (what you provide):

- PDF of research paper with quantitative findings
- The prompt

Outputs (required structure):

- Table with 3 columns: claim, evidence, and citation

Constraints / guardrails (no fabricated citations; cite chunk IDs; etc.):

- Extract 5 claims

When to use / when not to use:

- Use when testing the baseline of LLM capabilities
- Don't use when wanting precise citations, papers without clear quantitative findings, or production-ready outputs

Failure modes to watch for:

- Missing, vague, or wrong citations
- Direct quotes that are really just paraphrasing
- Incomplete information
- Inconsistent formatting

Prompt Text:
Extract 5 claims from this research paper attached about sleep and present them in a table with three columns: Claim, Evidence, and Citation.

Prompt Card 2:

Prompt name: CEE - Prompt B (Improved)
Intent: Extract verifiable claims from empirical sleep research papers with supporting evidence and precise citations with structured requirements and guardrails to ensure high quality output.

Inputs (what you provide):

- PDF of research paper with quantitative findings
- The prompt

Outputs (required structure):

- A table with 5 rows and 3 columns of claim, direct evidence/quote, and citation with format of (Paper#, Section_Name, Para#)

Constraints / guardrails (no fabricated citations; cite chunk IDs; etc.):

- The 5 claims must be from the Results or Discussion
- Can only use information explicitly stated in the paper
- Must include statistics, confidence intervals, etc when available
- Direct quotes must be verbatim
- If it's unsure about a claim or lacking clear evidence, skip it
- If there's less than 5 distinct claims, explain why and provide fewer
- LLM must take on a role as a research assistant and follow the provided example

When to use / when not to use:

- Use when accurate citations in a specific format is a must, direct quotes must be verbatim, and when it's okay to have less than 5 claims because of quality reasons
- Don't use when paper lacks a clear structure or opinion pieces without primary data

Failure modes to watch for:

- Wrong paragraph numbers cited
- Quotes being cited from wrong sections or when it's really from the abstract

Prompt Text:

ROLE: You are a research assistant trained in extracting evidence-backed claims from a sleep research paper and academic citation practices.

TASK: Extract exactly 5 claims from the Results or Discussion section.

CONTEXT: You are analyzing peer-reviewed research to identify specific, evidence-backed claims about sleep and well-being. Each claim must be directly supported by explicit evidence from the paper.

OUTPUT FORMAT (use this exact table structure):

Claim	Direct Evidence/Quote	Citation
[State the specific claim]	[Exact quote or data from paper]	(Paper#, Section_Name, Para#)
[Next claim]	[Supporting evidence]	(Paper#, Section_Name, Para#)
[Next claim]	[Supporting evidence]	(Paper#, Section_Name, Para#)
[Next claim]	[Supporting evidence]	(Paper#, Section_Name, Para#)
[Next claim]	[Supporting evidence]	(Paper#, Section_Name, Para#)

INSTRUCTIONS:
1. Read the entire paper, focusing on the Results and Discussion sections
2. Identify 5 distinct, substantive claims about sleep and its effects
3. For each claim, extract the direct evidence (quote, statistic, or data)
4. Cite using the format: (Paper#, Section_Name, Paragraph_Number)
   - Example: (Paper1, Results, Para_3) or (Paper2, Discussion, Para_12)
5. Number paragraphs sequentially within each section

CONSTRAINTS:
- Use ONLY information explicitly stated in the paper

- Include specific statistics, effect sizes, confidence intervals, or p-values when available
- Direct quotes must be verbatim and enclosed in quotation marks
- Do NOT infer, extrapolate, or add information not in the paper
- If a claim lacks clear evidence in the paper, skip it and find another claim
- If you cannot find 5 distinct claims with clear evidence, explain why and provide fewer

QUALITY CHECKS:
- Each citation must be verifiable (I should be able to find the exact text at that location)
- Each claim should be specific, not vague generalizations
- Evidence must directly support the claim, not tangentially relate to it

EXAMPLE ROW:
| Short sleep significantly increases anxiety risk in adults | "Very short sleep (<5 hours) was associated with a 40% higher risk of anxiety symptoms (adjusted IRR = 1.40, 95% CI [1.23–1.59])" | (Paper1, Results, Para_3) |

Now analyze the paper and extract 5 claims:

Prompt Card 3:

Prompt name: CSS - Prompt A (Baseline)
Intent: Compare 2 sleep research papers to find where they agree or disagree, with minimal guardrails.

Inputs (what you provide):
- 2 PDFs of research papers that have topical overlap
- The prompt

Outputs (required structure):
- A table with agreements, disagreements, and evidence from each side

Constraints / guardrails (no fabricated citations; cite chunk IDs; etc.):
- Compare the two papers given
- Create a table

When to use / when not to use:
- Use when testing the baseline of LLM capabilities, have papers with overlapping topics, and when a quick comparison is wanted
- Don’t use when wanting precise citations, real synthesis, or production-ready output

Failure modes to watch for:
- Missing, vague, or wrong citations
- Restating evidence without much comparison
- Focus more on one paper than the other
- Comparing non-comparable findings

Prompt Text:

Compare these two papers attached about sleep and create a table showing where they agree and disagree and evidence supporting each side.

Prompt Card 4:

Prompt name: CSS - Prompt B (Improved)
Intent: Execute cross-source synthesis identifying agreements and disagreements between two papers with precise citations and guardrails.

Inputs (what you provide):

- 2 PDFs of research papers that have topical overlap
- The prompt

Outputs (required structure):

- A table with the columns dimension, agreement between papers, disagreement/different emphasis, evidence from paper A, and evidence from paper B

Constraints / guardrails (no fabricated citations; cite chunk IDs; etc.):

- All comparisons must comeo from explicit statements from the papers
- Do not fabricate anything
- Include numbers, statistics, etc when comparing
- If a dimension isn't found state "not addressed" or "not comparable"

When to use / when not to use:

- Use when synthesizing between two papers, want to explore where two papers converge and diverge, and want high quality output
- Don't use when the two papers are completely unrelated or want just a summary instead of a comparative analysis

Failure modes to watch for:

- Missing, vague, or wrong citations
- Restating evidence without much comparison
- Focus more on one paper than the other
- Comparing non-comparable findings

Prompt Text:
ROLE: You are a research analyst specializing in evidence synthesis and comparative analysis of scientific literature.

TASK: Compare and contrast these two papers to identify areas of agreement, disagreement, and different emphases regarding sleep quality versus sleep quantity.

CONTEXT: These papers both examine relationships between sleep and health/cognitive outcomes, but they may emphasize different predictors (quality vs quantity) or reach different conclusions. Your synthesis should help readers understand where the evidence converges and diverges.

OUTPUT FORMAT (use this exact table structure):

Dimension	Agreement Between Papers	Disagreement/Different Emphasis	Evidence from Paper A	Evidence from Paper B
[Topic]	[What both agree on]	[How they differ or emphasize differently]	[Specific finding + citation]	[Specific finding + citation]

REQUIRED DIMENSIONS TO ANALYZE:
1. **Primary research focus**: What main question does each paper address?
2. **Key predictor variable**: Does each paper emphasize sleep quality, quantity, or both equally?
3. **Outcome measures**: What health/cognitive outcomes are examined?
4. **Main quantitative findings**: What are the key statistical results?
5. **Practical implications**: What do the authors conclude or recommend?

INSTRUCTIONS:
1. Read both papers completely, focusing on Methods, Results, and Discussion sections
2. For each dimension, identify explicit agreements and disagreements
3. Extract specific evidence with statistics, effect sizes, or direct quotes
4. Cite using format: (Paper#, Section_Name, Para#)
   - Example: (PaperA, Results, Para_5) or (PaperB, Discussion, Para_8)
5. Distinguish between:
   - Direct contradictions (papers reach opposite conclusions on same question)
   - Different emphases (papers focus on different aspects)
   - Complementary findings (papers address related but distinct questions)

CONSTRAINTS:
- Base ALL comparisons on explicit statements from the papers
- Do NOT fabricate agreements or disagreements
- If papers address different outcomes/populations, note this explicitly
- Include specific numbers, statistics, or effect sizes when comparing
- If papers don't address a dimension, state "Not addressed" or "Not comparable"
- Use direct quotes when they clarify key differences

REASONING APPROACH (Chain-of-Thought):
For each dimension, think through:
- What does Paper A explicitly say about this?
- What does Paper B explicitly say about this?
- Do they address the same question or different questions?
- Where do findings overlap? Where do they diverge?
- What evidence supports each position?

QUALITY CHECKS:
- Every claim about agreement/disagreement is backed by cited evidence from both papers
- Citations are precise and verifiable
- Synthesis is balanced (not favoring one paper over the other)
- Distinctions between contradiction vs different focus are clear

EXAMPLE ROW:
| Primary predictor emphasized | Both examine sleep's impact on health outcomes | Paper A emphasizes sleep QUALITY as stronger predictor; Paper B focuses on optimal sleep DURATION and U-shaped relationship | "Sleep quality showed larger standardized effects on physical health scores than sleep duration (β = 0.34 vs β = 0.18)" (PaperA, Results, Para_4) | "Cognitive performance was optimized at 7-8 hours, with both shorter and longer durations showing significant decrements (linear coef = 0.045, p < 0.001; quadratic coef = -0.023, p < 0.001)" (PaperB, Results, Para_6) |

Now compare the two papers:

Analysis Memo:

Project: Personal Research Portal - Sleep and Well-being Domain
Date: February 1st, 2026
To: Professor Anand Rao
From: Evelynn Chen

Executive Summary:
The following memo details my results, findings, and learnings from Phase 1. This includes patterns like structured prompts improve citation frequency and accuracy and adding constraints to the prompt prevents the model from paraphrasing. Failures encountered were formatting inconsistencies and extra output and the model presenting hallucinations as factual. Next steps for Phase 2 are also detailed.

Patterns Found:
(1) Structured Prompts Improve Citation Frequency and Accuracy
Comparing both Prompt As to their respective Prompt Bs, I saw a big improvement in citation frequency and accuracy. Without an explicit citation format specified, models would cite by just the name(s) of the author(s) and year published, provide a vague citation like "Results section" or "Table 3", or just simply not include a citation. By specifying in Prompt B that we'd like citations to be in (Paper#, Section_Name, Para#) and by providing an example, we saw consistent citation attempts for all evidence pulled. The accuracy of citations was a little off by a couple paragraphs or identifying the wrong section. We also noticed that the model would have more inaccurate citations for papers following a two-column format than a one-column approach, which may be fixed once preprocessing steps are done to the papers.

(2) Constraints Prevent Model from Paraphrasing Direct Evidence
With both of the baseline Prompt As, oftentimes evidence pulled would have quotation marks around them (indicating they're direct quotes explicitly stated in the paper), but they were actually paraphrased or modified slightly. This could lead to a whole host of problems, especially for the cross-source synthesis task where the model wouldn't be comparing the true findings of paper A to paper B. In Prompt B, I added guardrails such as "Direct quotes must be verbatim…" and "Can only use information explicitly stated in the paper" which successfully forced the model to make all quotes direct from the papers.

Failures Encountered:
(1) Format Inconsistencies and Extra Output
For the Prompt As, the output format was not super clear, so I understand the inconsistencies there, but for the Prompt Bs when writing the prompts I expected the model to only give back exactly what I asked for (e.g. a table with the specified columns). It did successfully provide the table, but would provide bullet points, paragraphs, and explanations after about more key findings or evidence or offering to do more research. When thinking about how this output will flow straight into the next processing step in production, extra output such as the "Sure,

let me get started!" and extra information may add noise to the data and take away from the actual task at hand.

(2) Presenting Hallucinations as Factual
During the cross-source synthesis task, sometimes the model would create plausible-sounding comparisons that were not grounded in the papers, or would miss key information such as not finding quantitative findings when there were some. Without further review though, the model didn't admit this interpretation and presented with confidence, which can be misleading.

Impact on Phase 2:
As we consolidate our findings from Phase 1 and move onto Phase 2, we'll explore the following:

- Use more of Prompt B style prompts than Prompt A. Models need this more complex, optimized prompt style to better understand what I want.
- Design a better citation chunking and architecture system. Relying on the model to identify paragraphs and count them has proven to be inconsistent.
- Preprocess the papers before feeding them to the model to get a more consistent format.
- Have the model evaluate its own work based on the same 3 scoring metrics on the 1-4 scale.

Phase 1 taught me the importance of prompt and context engineering, and how models are so nondeterministic. I found areas of improvement to work on in Phase 2 as I start building out the initial RAG system.

Best regards,
Evelynn Chen
evelynnchen@cmu.edu