Research Plan in current stage (2-3 weeks):

Designing Benchmark Queries to Evaluate LLM Deep Research Capabilities in the Energy Domain

Objective:

To develop a benchmark of 100 queries that assess large language models (LLMs) on their ability to perform deep research in the energy domain. This stage focuses on query design, which includes both manual and AI-assisted generation.

Query Set Composition:

Total: 100 queries

* 50 manually authored by researchers (students)
* 50 generated via LLM APIs (e.g., OpenAI)

Domain coverage:

* 40% general energy-related questions (single-subdomain or broad topic)
* 60% cross-subdomain questions (multi-domain)

Difficulty balance:

* 30% Easy
* 40% Medium
* 30% Hard

Design Principles:

1. Diversity and Domain Coverage – Cover major subdomains: renewable energy, fossil fuels, nuclear, grid & storage, energy policy, economics, and environmental impact.
2. Difficulty Levels – Ensure balanced distribution; define criteria for Easy, Medium, Hard based on reasoning complexity.
3. Reasoning-Oriented – Favor queries that require synthesis, multi-step reasoning, or explanation. Avoid trivia or single-fact recall.
4. Clarity and Specificity – Queries should be clearly phrased with a defined research direction.
5. Real-World Relevance – Focus on current issues, realistic scenarios, or challenges in the energy sector.
6. Format Variety – Use different formats: comparative, analytical, hypothetical, and descriptive.

Query Metadata Schema (for each query)

* ID
* Query Text
* Category ("General" or "Cross-Subdomain")
* Subdomain(s) (e.g., Renewable, Policy, etc.)
* Difficulty (Easy, Medium, Hard)
* Source (Manual / AI-generated)

Workflow:

1. Understand the Subdomains & Principles
   1. Review the energy subdomains and design goals.
2. Manual Query Generation
   1. Use the taxonomy to write balanced queries (topic & difficulty)
   2. Submit queries with metadata tags in the required schema
3. AI-based Query Generation
   1. Use GPT API or similar tool to generate prompts fitting the subdomain and difficulty
   2. Curate, revise, and filter out unqualified queries
4. Review & Tagging
   1. Ensure queries are clear, novel, and correctly tagged
   2. Format in spreadsheet or JSON (Preferred) following the metadata schema
5. Submit Final Deliverables
   1. 100 fully formatted queries with metadata
   2. Notes on how queries were generated and reviewed
   3. Any observations or issues encountered

End Goal:

A high-quality, labeled set of 100 benchmark queries that comprehensively evaluate deep research capabilities of LLMs in the energy domain, suitable for future evaluation and expansion into other domains.