Evaluation of Deep–Research Report Generation Agent

Group 6

May 11 , 2025

1 Dataset Generation

To generate our own dataset of reports generated from our agent, we passed queries such as “Machine Learning for Frequency Excursion Prediction in High Inverter-Based Resource Grids”, “A Comparative Study of Superconducting Qubit Designs: Transmons, Flux Qubits, and Gatemons”, to our agent and made reports on it. We took specific and meticulous topics that would require research papers and web pages for report generation. We have generated 36 reports in total and used them as our datapoints for the evaluation.

Our agent uses Gemini 2.0 flash to invoke LLM calls. To compare our agent, we used Gemini 2.5 Flash and asked it to generate reports on the same topics that we did for our deep research agent as there wasn’t a pre-defined dataset relevant for our evaluation. This acted as a baseline that we compared our model to.

2 Evaluation Criteria

We used the following criteria to evaluate our report generation:

1. **ROUGE and BERTScore**
   - *ROUGE* (Recall-Oriented Understudy for Gisting Evaluation): ROUGE-1, ROUGE-2, ROUGE-L, and a combined ROUGE score.
   - *BERTScore*: Precision, Recall, and F_1, computed via the Huggingface Evaluate `bertscore` module with `model_type=roberta-base`.
2. **LLM evaluation over 4 human metrics**
   We prompt an LLM model (ChatGPT o4-mini-high) to systematically score a deep-research report across four parameters (1–5):
   - *Accuracy*: Are the key facts and results stated in the report correct?
   - *Depth*: Does the content of the report provide a shallow understanding of the topic or does it bring a nuanced outlook?
   - *Structure*: Is the report well-structured with well-demarcated sections?
   - *Use of Sources*: Are the cited sources relevant and credible?

3. **Chain of thought rating**
   We again use an LLM model (ChatGPT o4-mini-high) to evaluate the chain of thought of our agent on a scale of 1–5 on three metrics:
   - *Clarity*: Is the purpose and justification of each step clearly stated?
   - *Logical Flow*: Are the steps coherent, and is the flow of thoughts sensible?
   - *Conciseness*: Is the chain of thought devoid of unnecessary or redundant steps?
4. **Readability and Formality**
   Classic readability formulas (Flesch, Flesch–Kincaid, SMOG, Gunning Fog, Automated Readability Index) and a spaCy-based formality score (0–200).

3 Results and Inference

3.1 ROUGE and BERTScore

Metric	Mean
ROUGE-1	0.5460
ROUGE-2	0.2053
ROUGE-L	0.1569
Combined ROUGE	0.2493

ROUGE Metrics

Metric	Mean
Precision	0.85815
Recall	0.84887
F_1	0.85347
Final Combined Score	0.67218

BERTScore Metrics A lower ROUGE score and higher BERTScore imply that the two reports have fewer exact text overlaps but a high semantic match. This shows that our model produces reports similar in meaning to the baseline.

3.2 Readability and Formality

Readability Metrics

Formality Score The readability indices suggest that a high education level is needed to understand the text, consistent with the fact that we are sourcing from research papers and journals and the queries are science oriented.

Metric	Mean
Flesch Reading Ease	28.62
Flesch–Kincaid Grade Level	12.62
SMOG Index	14.39
Gunning Fog Index	16.95
Automated Readability Index	12.29

Metric	Mean
Formality score	160.56

3.3 LLM Evaluation on Human Metrics

Metric	Mean
Accuracy	3.83
Depth	2.77
Structure	4.04
Use of Sources	2.84
Overall Average	3.37

A good level of Accuracy shows effective fact extraction resulting from effective scraping and RAG retreival. Lower Depth reflects limited query tree levels due to limited resources and computational power. Strong Structure is due to prompt-generated outlines based on the sub query tree that we generate. Lower Use of Sources stems from imperfect in-line citations.

3.4 Chain of Thought Evaluation

Metric	Average
Clarity	4.16
Logical Flow	4.07
Conciseness	2.48

High Clarity and Logical Flow reflect well-structured task and subtask generation. Lower Conciseness arises from repetitive statements (e.g. cached abstracts, error messages).