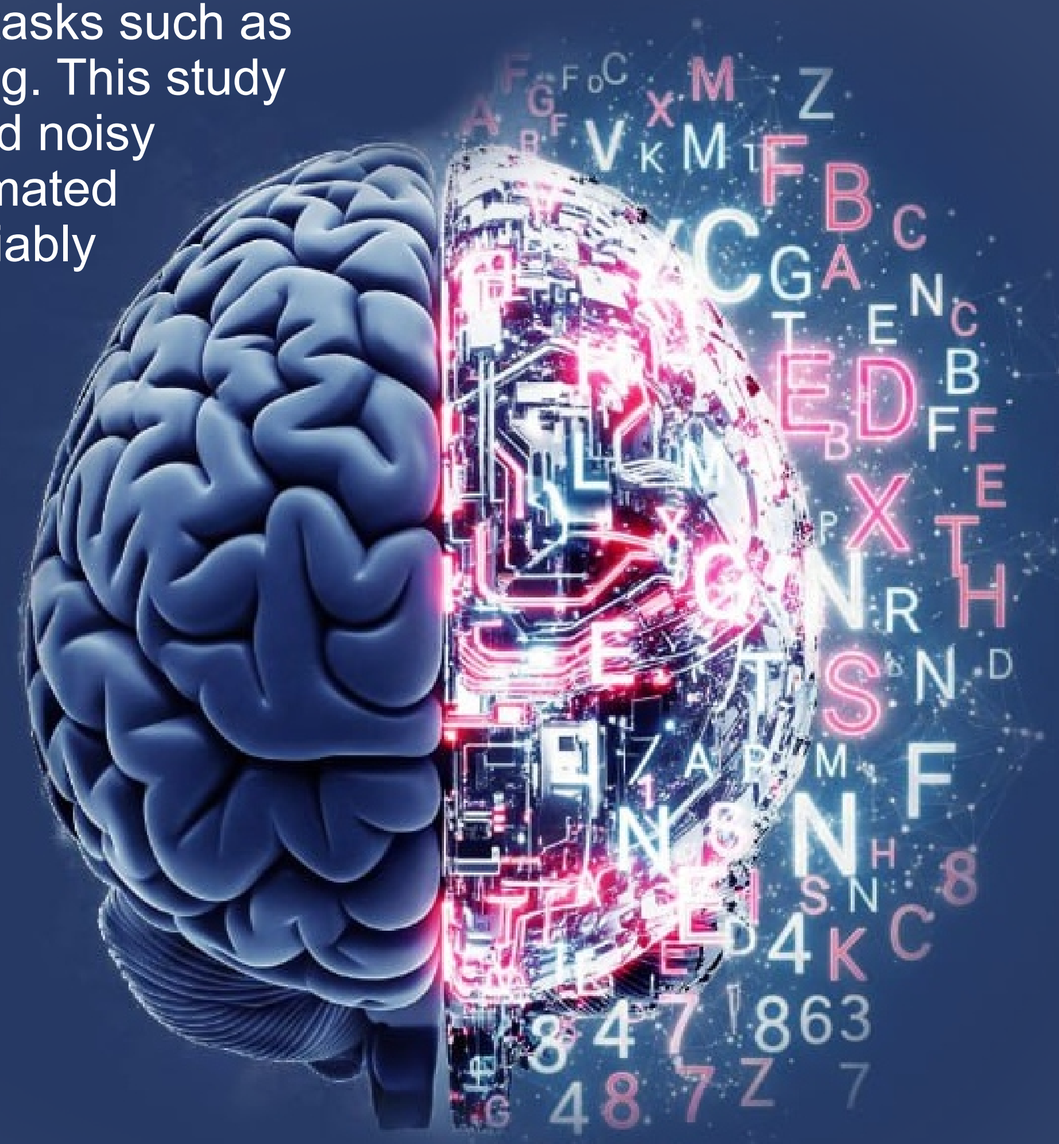
Evaluation of AI Text Generators

1. Introduction

OpenAI ChatGPT4.0 | Claude | Gemini | Command R+

Large language models (LLM) like GPT-4, Claude 3.7, Gemini, and Command R are increasingly used in real-world tasks such as ummarization, fact-checking, and decision-making. This study evaluates their performance under both clean and noisy input conditions using a mix of human and automated assessments. The goal is to understand how reliably these models operate beyond ideal settings.



2. Research Questions

Does source quality (clean vs. noisy text) affect the language structure and logical quality of AI-generated conclusions?

Can LLMs accurately read hybrid semantic constructs (e.g., process vs. decision) in real-world texts?

How do AI-generated final products compare to human-written outputs in terms of accuracy, harmony, and specific to a task performance (summarization, question answering, grammatical correction, fact-checking)?

3. Methodology

Study Design:
Hybrid evaluation using both human judgment and automated metrics
Focus: GPT-4, Claude 3.7, Gemini, Command R
Tasks: Summarization, fact-checking, grammar correction, decision analysis

Expert Evaluation (Researcher)

ExpertEvaluation(Researcher)
☑ **6criteria:**InputImpact, Semantics,Summarization, Decision-Making,Grammar, Factuality
☑ 5-pointscale,scoredby researcher ★★★★★

Human Evaluation (n=14)

HumanEvaluation(n=14)
☑ 5criteria:Accuracy, Coherence,Consistency, Fluency,TaskFit
☑ Blind,Likertscale(1–5), averagedpermodel/task

Automated Metrics

AutomatedMetrics:
☑ ROUGE(overlap),BLEU (precision),BERTScore (semantics)
☑ Benchmarkedwithhuman judgments

4. Key Findings

- ☑ Input Quality Matters
- ☑ Semantic Reasoning
- ☑ Human-Like Summaries
- ☑ Error Sensitivity
- ☑ Factual Validation
- ☑ Practical Decision-Making

5. Conclusions

1. **GPT-4 & Claude 3.7:** Most consistent, logical, and fluent.
2. **Command R:** Precise, best for grammar tasks.
3. **Gemini:** Fluent but weaker in reasoning.
4. **Input quality matters :** GPT-4 & Claude handle noise best.
5. **No model fits all tasks — choose based on context.**
6. **Limitations:** Small sample, English-only, static tests.
7. **Future:** Add dialogue, multilingual, bias studies.

Promoter: Prof. Dr. Koen Vanhoof
Sanaz Khoobi
Program: Master of Management in Data Science (2023–2025)

UHASSELT