UNIVERSITAS
GADJAH MADA
Analisis Perbandingan Kualitas Pembuatan Flashcard Berbasis Large Language Model Dari Video Youtube
Dan Dokumen Pada Sistem Pembelajaran Spaced Repetition
Devrangga Hazza Mahiswara, Dinar Nugroho Pratomo, S.Kom., M.IM., M.Cs.
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

INTISARI

ANALISIS PERBANDINGAN KUALITAS PEMBUATAN *FLASHCARD* BERBASIS *LARGE LANGUAGE MODEL* DARI VIDEO YOUTUBE DAN DOKUMEN PADA SISTEM PEMBELAJARAN *SPACED REPETITION*

Devrangga Hazza Mahiswara

22/498430/SV/21212

Retensi pengetahuan jangka panjang merupakan tantangan fundamental dalam pembelajaran. Sistem *Spaced Repetition* berbasis *flashcard* terbukti efektif mengatasi tantangan tersebut. Namun, pembuatan *flashcard* secara manual menjadi hambatan yang dapat diatasi melalui otomasi berbasis *Large Language Model* (LLM). Penelitian ini menganalisis dan membandingkan kualitas pembuatan *flashcard* yang dihasilkan oleh tiga model LLM, yaitu Gemini 2.5 Flash, GPT-4o mini, dan Claude Haiku 4.5, dari video YouTube dan dokumen (PDF/DOCX). *Flashcard* tipe *basic* dipilih dalam penelitian ini, yaitu hanya melibatkan pertanyaan dan jawaban. Pengujian dilakukan menggunakan 8 sumber, yaitu empat video YouTube dan empat dokumen, yang disediakan oleh dua ahli bidang Boga dan Pariwisata. Kualitas *flashcard* dinilai oleh kedua ahli secara *cross-rating* melalui Google Form berbasis skala Likert 1-5 pada empat aspek, yaitu Relevansi, Kejelasan, *Coverage*, dan *Hallucination Rate*, sedangkan rancangan *prompt* divalidasi melalui koefisien Aiken's V. Latensi dan biaya operasional diukur langsung dari log teknis. Hasil analisis menunjukkan bahwa ketiga model menghasilkan *flashcard* berkualitas tinggi dan setara, dengan selisih skor komposit hanya 2,2 poin. Gemini 2.5 Flash memperoleh skor komposit tertinggi (84,1) dengan keunggulan pada Coverage (86,2%) dan Kejelasan, Claude Haiku 4.5 unggul pada Akurasi Faktual, sedangkan GPT-4o mini paling efisien (biaya $0,0008 per *deck* dan latensi tercepat 8,17 detik). Reliabilitas antar-penilai tergolong rendah akibat *ceiling effect*, sehingga perbedaan kecil antar model tidak dapat diinterpretasikan sebagai keunggulan mutlak. *Flashcard* dari dokumen dan dari video YouTube memperoleh kualitas yang setara. Temuan ini menjadi acuan berbasis data bagi pengembang sistem pembelajaran *spaced repetition* dalam memilih model LLM sesuai prioritas kualitas, kecepatan, dan efisiensi biaya.

Kata kunci : evaluasi LLM, *flashcard*, pembuatan pertanyaan otomatis, *large language model*, *spaced repetition*

UNIVERSITAS
GADJAH MADA
Analisis Perbandingan Kualitas Pembuatan Flashcard Berbasis Large Language Model Dari Video Youtube
Dan Dokumen Pada Sistem Pembelajaran Spaced Repetition
Devrangga Hazza Mahiswara, Dinar Nugroho Pratomo, S.Kom., M.IM., M.Cs.
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

ABSTRACT

COMPARATIVE ANALYSIS OF FLASHCARD GENERATION QUALITY BASED ON LARGE LANGUAGE MODELS FROM YOUTUBE VIDEOS AND DOCUMENTS IN A SPACED REPETITION LEARNING SYSTEM

Devrangga Hazza Mahiswara

22/498430/SV/21212

Long-term knowledge retention is a fundamental learning challenge. Flashcard based Spaced Repetition systems have proven effective in addressing this challenge. However, the manual creation of flashcards remains a barrier that can be overcome through Large Language Model (LLM)-based automation. This study analyzes and compares the quality of flashcards generated by three LLMs, namely Gemini 2.5 Flash, GPT-4o mini, and Claude Haiku 4.5, from YouTube videos and documents (PDF/DOCX). The basic flashcard template, involving only questions and answers, was used in this study. Testing was conducted using 8 sources, namely four YouTube videos and four documents, provided by two domain experts in Culinary and Tourism. Flashcard quality was assessed by both experts through cross-rating on a Google Form based on a 1-5 Likert scale across four aspects, namely Relevance, Clarity, Coverage, and Hallucination Rate, while the prompt design was validated using Aiken's V coefficient. Latency and operational cost were measured directly from technical logs. The analysis results indicate that the three models produced high-quality and comparable flashcards, with a composite-score difference of only 2.2 points. Gemini 2.5 Flash obtained the highest composite score (84.1) with advantages in Coverage (86.2%) and Clarity, Claude Haiku 4.5 led in Factual Accuracy, while GPT-4o mini was the most efficient (cost of $0.0008 per deck and the fastest latency of 8.17 seconds). Inter-rater reliability was low due to a ceiling effect, so the small differences between models cannot be interpreted as an absolute advantage. Document-sourced and video-sourced flashcards obtained equivalent quality. These findings serve as a data-driven reference for developers of spaced repetition learning systems in selecting an LLM according to priorities of quality, speed, and cost efficiency.

Keywords: automatic question generation, flashcard, large language model, LLM evaluation, spaced repetition