EdgePrompt: Engineering Guardrail Techniques for Offline LLMs in K-12 Educational Settings

Riza Alaudin Syah*
alaudinsyah@graduate.utm.my
Universiti Teknologi Malaysia
Johor Bahru, Malaysia

Christoforus Yoga Haryanto*
cyharyanto@zipthought.com.au
ZipThought
Melbourne, VIC, Australia

Emily Lomempow
ZipThought
Melbourne, VIC, Australia

Krishna Malik
Independent Researcher
Jakarta, Indonesia

Irvan Putra
Independent Researcher
Jakarta, Indonesia

Abstract

EdgePrompt is a prompt engineering framework that implements pragmatic guardrails for Large Language Models (LLMs) in K-12 educational settings through structured prompting inspired by neural-symbolic principles. The system addresses educational disparities in Indonesia's Underdeveloped, Frontier, and Outermost (3T) regions by enabling offline-capable content safety controls. It combines: (1) content generation with structured constraint templates, (2) assessment processing with layered validation, and (3) lightweight storage for content and result management. The framework implements a multi-stage verification workflow that maintains safety boundaries while preserving model capabilities in connectivity-constrained environments. Initial deployment targets Grade 5 language instruction, demonstrating effective guardrails through structured prompt engineering without requiring formal symbolic reasoning components.

CCS Concepts

• **Social and professional topics** → **K-12 education**; • **Applied computing** → **Computer-assisted instruction**; • **Computing methodologies** → *Natural language generation.*

Keywords

Large Language Models, Edge Computing, K-12 Education, AI Safety, Prompt Engineering, Content Filtering, Offline AI, Educational Technology, Guardrails

ACM Reference Format:
Riza Alaudin Syah, Christoforus Yoga Haryanto, Emily Lomempow, Krishna Malik, and Irvan Putra. 2025. EdgePrompt: Engineering Guardrail Techniques for Offline LLMs in K-12 Educational Settings. In *Proceedings of 2nd PromptEng Workshop at the ACM WebConf'25 (PromptEng'25).* ACM, New York, NY, USA, 2 pages. https://doi.org/XXXXXXX.XXXXXXX

*Both authors contributed equally to this research.

PromptEng'25, Sydney, NSW

ACM ISBN 978-1-4503-XXXX-X/2025/04
https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Recent advances in LLM guardrails have shown promise for domain-specific control [3, 4], yet resource-constrained environments require balancing safety with offline capabilities. Inspired by neural-symbolic architectures [4], EdgePrompt implements "structured prompt-based controls enforcing content boundaries while maintaining validation across edge deployments" through: (1) **Structured Prompting:** templates encoding safety constraints with formal validation rules, (2) **Multi-stage Validation:** sequential prompt-based checks with explicit boundary conditions, and (3) **Edge Deployment Compatibility:** optimized mechanisms for resource–constrained operation.

Our framework addresses core challenges in K-12 education: (1) maintaining robust content safety in offline environments, (2) enabling sophisticated assessment with edge-based validation, and (3) ensuring consistent distributed evaluation. Initial deployment targets Grade 5 language instruction in Indonesia's "3T" regions (Frontier, Outermost, and Disadvantaged) [1, 5], where limited internet access [2] necessitates edge LLMs for evaluation while leveraging cloud LLMs for content generation.

2 Methodology

We implement a rigidly structured validation pipeline leveraging cloud and edge LLMs for distinct operational roles. The architecture enforces safety through multi-stage template validation, explicit constraint propagation, and formalized evaluation protocols, as shown in Fig. 1. The core components implement template processing, staged validation, and lightweight integration.

2.1 Teacher-Driven Content Generation

(1) **Question Template Definition:** (a) Template: "Write a descriptive paragraph about [topic] using at least 3 sensory details", (b) Answer boundaries: "Response must be 50-100 words, school-appropriate vocabulary", and (c) Learning outcome: "Student demonstrates ability to use descriptive language (Grade 5 Language Arts Standard 5.2)"
(2) **Cloud/Edge LLM Pipeline:** (a) Rubric: "4 points: Uses 3+ sensory details, proper length, grade-level vocabulary", (b) Format: Simplified scoring criteria for offline LLM processing, and (c) Validation: "Check: word count 50-100, presence of sensory words, appropriate vocabulary"

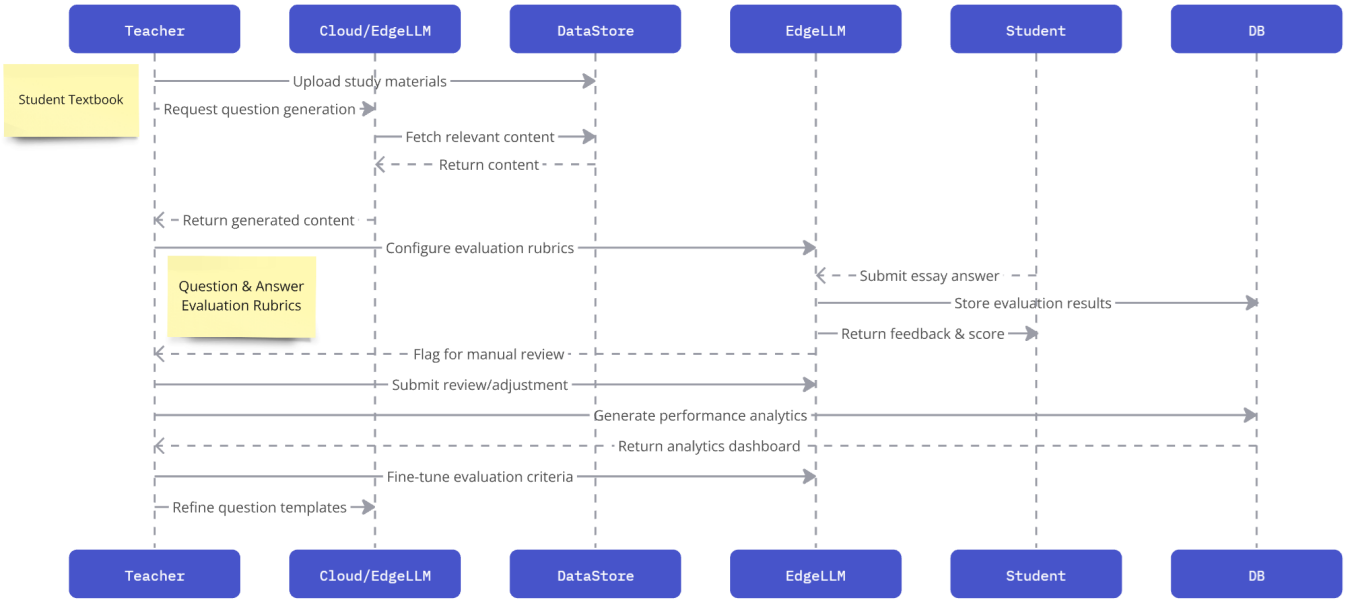

Figure 1: Sequence diagram of the system.

2.2 Student Answer Evaluation

(1) **Edge Validation:** (a) Question-answer match: "Does answer describe requested topic?", (b) Multi-stage checks: Length → vocabulary → content → scoring, and (c) Safety bounds: Filter inappropriate content, off-topic responses
(2) **Evaluation Process:** (a) Apply rubric: Count sensory details, check length requirements, (b) Score calculation: "3/4 points - meets length, 2 sensory details", and (c) Verify constraints: "Response meets safety and topic requirements"

2.3 Teacher Review System

(1) **Response Analysis:** (a) Flag responses: "80% confidence threshold for automated scoring", (b) Review triggers: Borderline scores, unusual patterns, and (c) Track patterns: "Common vocabulary errors, length issues"
(2) **System Updates:** (a) Adjust rubric: "Add specific examples of sensory language", (b) Optimize criteria: Update scoring weights based on review history, and (c) Refine templates: Clarify instructions based on common mistakes

Full implementation with the documentation and example prompts can be seen in the project GitHub repository.

2.4 Deployment Strategy

The EdgeLLM deployment architecture implements: (1) optimized edge runtime for Llama 3.2 3B with minimal resource footprint, (2) containerized environment ensuring consistent model behavior, (3) fault-tolerant storage system for offline operation, and (4) distributed validation protocol maintaining safety constraints.

2.5 Validation Strategy

Our validation framework integrates three key components:

Functional Metrics: Measuring guardrail effectiveness ($E(g)$) through valid response ratios, offline stability ($S(t)$) via operational reliability, and teacher workflow integration through systematic adoption tracking.

Performance Analysis: Evaluating edge deployment capabilities through resource utilization ($U(r)$), response latency profiles ($L(t)$), and scalability characteristics under varying loads.

System Validation: Building on SPADE's guardrail metrics [6], we extend evaluation to edge-specific indicators while maintaining pure prompt engineering approaches without model tuning. Insights from deployment optimization and constraint analysis inform our ongoing development of educational applications.

3 Resources

Project repository: https://github.com/build-club-ai-indonesia/edge-prompt. Teaching materials: https://buku.kemdikbud.go.id/.

References

[1] 2020. Peraturan Presiden (PERPRES) Nomor 63 Tahun 2020 Penetapan Daerah Tertinggal Tahun 2020-2024. https://peraturan.bpk.go.id/Details/136563/perpres-no-63-tahun-2020 Publication Title: Database Peraturan | JDIH BPK.
[2] Livia Kristianti Raka Adji. 2024. Indonesia's internet penetration hits 79.5 percent, trend continues. *Antara News* (Jan. 2024). https://en.antaranews.com/news/304593/indonesias-internet-penetration-hits-795-percent-trend-continues
[3] Yi Dong, Ronghui Mu, Gaojie Jin, Yi Qi, Jinwei Hu, Xingyu Zhao, Jie Meng, Wenjie Ruan, and Xiaowei Huang. 2024. Building Guardrails for Large Language Models. doi:10.48550/ARXIV.2402.01822
[4] Yi Dong, Ronghui Mu, Yanghao Zhang, Siqi Sun, Tianle Zhang, Changshun Wu, Gaojie Jin, Yi Qi, Jinwei Hu, Jie Meng, Saddek Bensalem, and Xiaowei Huang. 2024. Safeguarding Large Language Models: A Survey. doi:10.48550/ARXIV.2406.02622
[5] Kementerian Desa. 2025. Official Website of the Ministry of Villages. https://www.kemendesa.go.id
[6] Mohammad Niknazar, Paul V Haley, Latha Ramanan, Sang T. Truong, Yedendra Shrinivasan, Ayan Kumar Bhowmick, Prasenjit Dey, Ashish Jagmohan, Hema Maheshwari, Shom Ponoth, Robert Smith, Aditya Vempaty, Nick Haber, Sanmi Koyejo, and Sharad Sundararajan. 2024. Building a Domain-specific Guardrail Model in Production. doi:10.48550/ARXIV.2408.01452