JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

HARP: Human Answer Reconciliation Pipeline

[1]Harsha Nishad, [2]Vedant Fuley, [3]V Jogendra Rao, [4]Sourabh Dewangan

[1]Assistant Professor, Department of Computer Science & Engineering
[2-4]B. Tech. Scholars, Department of Computer Science & Engineering
[1-4]Bhilai Institute of Technology, Raipur (C.G.), India

Abstract: The increasing demand for scalable and consistent evaluation of open-ended responses in educational environments necessitates robust AI-assisted assessment systems. This paper introduces the Human Answer Reconciliation Pipeline (HARP), a modular and extensible framework designed to automate the grading of text-based student responses using Large Language Models (LLMs). Unlike traditional systems that rely solely on reference answers, HARP supports both reference-based and standalone answer evaluation modes, enabling flexible assessment across diverse pedagogical contexts. The system features a FastAPI-based backend with customizable scoring formats (e.g., 0–10, 0–5, percentage), batch evaluation capabilities, and caching for performance optimization. The Model Context Protocol (MCP) Server is central to HARP's architecture, which enables modular LLM integration, file-based request handling, and standardized evaluation flows. Additionally, HARP employs techniques such as Reference Answer Guided (RAG) evaluation and zero/few-shot prompting strategies for enhanced scoring accuracy and interpretability. Designed for deployment in academic institutions, e-learning platforms, and AI research settings, HARP bridges the gap between manual grading and intelligent automation, delivering transparent, adaptable, and pedagogically aligned assessments.

Keywords **- Automated Answer Evaluation, Human Answer Reconciliation Pipeline (HARP), Large Language Models (LLM), FastAPI, Model Context Protocol (MCP), Educational Technology, RAG Evaluation, NLP in Education, Batch Scoring System, AI-based Grading.**

I. INTRODUCTION

The evolution of educational assessment has been significantly influenced by the growing demand for scalable, consistent, and intelligent evaluation systems. As classrooms expand and digital learning platforms proliferate, the traditional approach to grading—especially for open-ended or subjective questions—struggles to keep pace. Manual evaluation of text-based student responses is time-consuming, prone to bias, and often inconsistent, particularly when applied across large student populations or interdisciplinary contexts. These challenges highlight the need for advanced solutions capable of delivering accurate, explainable, and real-time assessments.

To address this gap, this project introduces the Human Answer Reconciliation Pipeline (HARP), a comprehensive system designed to automate the evaluation of textual student responses using the power of modern large language models (LLMs). Unlike conventional systems that rely solely on comparing student answers to pre-defined reference responses, HARP introduces a hybrid evaluation approach. It supports both reference-based scoring and standalone, context-aware analysis, enabling flexible deployment across various educational scenarios. This dual-mode capability makes HARP particularly effective in settings where reference answers may be ambiguous, unavailable, or too rigid for nuanced interpretation.

The system is architected for scalability and modularity. Built on a FastAPI backend, HARP enables high-performance RESTful interactions for both single and batch answer evaluations. A core innovation of the system is its integration with the Model Context Protocol (MCP), which allows seamless routing of evaluation requests to various LLMs such as LLaMA, Claude, or Gemini. This model-agnostic design ensures future extensibility and allows educators and developers to experiment with different models without modifying the core logic.

To enhance grading reliability and educational alignment, HARP incorporates advanced techniques such as Reference Answer Guided (RAG) evaluation, prompt engineering for zero-shot and few-shot grading, and semantic similarity analysis using sentence embeddings. These components work together to provide meaningful feedback, structured scoring, and pedagogically relevant evaluations, whether for summative assessments or formative learning activities.

By bridging the gap between human-level interpretation and AI-driven automation, HARP represents a forward-looking solution in the field of educational technology. It aims not only to improve the efficiency of grading but also to promote fairness, consistency, and adaptability in the evaluation process—qualities that are increasingly essential in modern, data-driven learning environments.

II. Related Work

The development of large language models (LLMs), which are increasingly essential for automating, enhancing, and customizing grading systems, is causing a paradigm change in the field of educational assessment. AI-powered frameworks that can assess complicated student replies at scale are gradually replacing traditional grading, which is frequently manual and limited by time, consistency, and human bias. These systems, which are powered by models such as Claude, Gemini, LLaMA-3, and GPT-4, open a new era in the use of artificial intelligence in education by providing feedback, rubric alignment, and semantic evaluation in addition to correctness assessment.

A foundational example of this evolution is seen in Smart Grading [1], which presents an AI-enhanced grading tool combining user-defined rubrics and LLMs for reliable automated scoring. Similarly, the AIG system proposed by Myint et al. [2] introduces a topic-based, instructor-validated grading framework where LLMs extract topics and instructors refine marking schemes. Their collaborative method significantly improves accuracy in multipart questions, revealing the power of human-in-the-loop grading systems. Grévisse [5] furthers this theme in medical education, showing that GPT-4 and Gemini can reliably grade multilingual short-answer responses when paired with high-quality answer keys, achieving time savings and maintaining alignment with human graders.

Beyond case-specific tools, broader architectural innovations are emerging. Pajo [3] introduces the Model Context Protocol (MCP), a model-agnostic communication layer that allows LLMs to interface seamlessly with educational systems. MCP empowers systems like HARP (Human Answer Reconciliation Pipeline) to dynamically route grading tasks to various LLMs without altering backend logic, thus fostering scalability and modularity. Chang et al. [4] support this modular vision by surveying evaluation methods across LLM use cases, concluding that robust evaluation systems must evolve in tandem with model capabilities. They emphasize areas like AGI benchmarking, robustness, and behavioral evaluation, framing evaluation itself as a core AI discipline.

In their investigation of pedagogical depth, Cohn et al. [8] use active learning and chain-of-thought prompting in science assessments, finding that explanation-driven grading improves alignment with human reasoning. Their approach highlights a larger trend toward formative, feedback-rich assessments made possible by LLMs. Taking a macro-ethical approach, Fagbohun et al. [6] draw attention to automated systems' vulnerabilities, such as depersonalization, prejudice, and inconsistent feedback. They support data governance, interpretability, and human oversight in AI grading pipelines—as expressed in the literature.

Shankar et al. [7] suggest EvalGen, a tool that assists teachers in iteratively improving grading standards by bringing LLM-generated evaluation logic into line with human preferences in order to overcome these complications. They question presumptions of static assessment criteria and support dynamic, co-constructed grading techniques with their discovery of "criteria drift," where standards change as people engage with outcomes. A zero-shot LLM grading system is demonstrated by Yeung et al. [9], who achieve great student satisfaction using rubric-based assessments. By analyzing prompt structure across languages, Mello et al. [10] supplement this and show that customized prompts have a major impact on grading accuracy.

The breadth of LLM integration is illustrated in emerging domains. Malik et al. [13] apply LLMs to finance essay grading using one-shot and few-shot prompts, while Henkel et al. [14] empirically validate GPT-4's grading across K -12 disciplines. Li [15] and Chen et al. [16] demonstrate how prompt engineering supports STEM engagement, while Tanaka et al. [17] introduce genetic algorithms for multilingual prompt optimization. In administrative contexts, Mohanraj et al. [18] design LLM-powered assistants for learning and scheduling, and Meißner et al. [19] automate self-assessment quiz generation through EvalQuiz. Xie et al. [12] rethink grading as a holistic cycle of feedback, rubric engineering, and learning analytics, reflecting a systems-level evolution.

Notably, Zhang et al. [2] evaluate criterion-based grading with domain-specific prompts, finding that even open-source models can achieve performance near proprietary LLMs. Frigui (2024) expands on this with qualitative coding automation, showcasing LLMs' versatility beyond numerical scoring. Meanwhile, Mollick & Mollick (2023) propose seven frameworks for AI-assisted education, presenting pedagogical blueprints for integrating LLMs as student mentors and evaluators.

These studies collectively create a powerful story: LLMs are change agents in education, not just grading instruments. The research shows that educational innovation and technology capabilities are convergent, as evidenced by modular pipelines like HARP and EvalGen and rubric-rich evaluations in science, finance, and medicine. But ethical prudence is still essential. As LLMs take on greater responsibility for knowledge assessment, the challenge is not just to assess more effectively but also to do it in a fair, transparent, and cooperative manner while always placing students at the center of the process.

Table 1 Comparative Analysis of Key Studies on LLM-Based Educational Evaluation

No.	Study	Focus Area	Methodology	LLM Used	Human-In-the-Loop	Key Contribution
1	Tobler (2024)	AI-enhanced grading tool	Web app + empirical validation	GPT	Yes	Open-source customizable grading interface
2	Myint et al. (2024)	Topic-based multipart grading	Prompt-based topic extraction	GPT, LLaMA	Yes	Instructor-guided topic alignment

			+ Python scoring			
4	Chang et al. (2024)	LLM evaluation metrics	Survey + taxonomy	Multiple LLMs	No	Defines "what, where, how" of evaluation
5	Grévisse (2024)	Short answer grading in medical ed.	Comparison across languages	GPT-4, Gemini	Yes	High agreement in multilingual settings
6	Fagbohun et al. (2024)	Ethics in AI grading	Case studies + policy framing	General LLMs	Yes	Ethical framework + risk classification
7	Shankar et al. (2024)	LLM evaluation tools	EvalGen interface + qualitative study	GPT-4	Yes	Co-design of grading functions with users
8	Cohn et al. (2024)	Science formative assessment	CoT prompting + active learning	GPT-4	Yes	Explainable grading aligned with rubrics
9	Yeung et al. (2025)	Zero-shot grading framework	Prompt templates	GPT-4	No	Simple yet effective zero-shot scoring
10	Mello et al. (2024)	Multilingual short answer grading	Prompt design comparison	GPT-4	No	Prompt sensitivity across languages
11	Jacobsen et al. (2025)	Personalized AI feedback	Controlled experiment	GPT-3.5, GPT-4	Yes	Prompt tuning for feedback quality
12	Xie et al. (2024)	Feedback loop in assessment	Learning analytics integration	GPT-4	Yes	Feedback as part of learning cycle
13	Zhang et al. (2024)	Rubric-based grading	Evaluation pipeline	Open-source LLMs	No	Matches GPT-level accuracy in rubric use
14	Malik et al. (2024)	Essay grading in finance	Few-shot prompt testing	GPT-4	No	Domain-specific grading effectiveness
15	Henkel et al. (2024)	K-12 short answer grading	Empirical user study	GPT-4	Partial	High agreement across levels
16	Li (2024)	LLM in STEM classrooms	Prompt engineering examples	Claude, GPT-3.5	No	Enhancing student engagement via prompts
17	Chen et al. (2024)	K-12 STEM education + LLMs	Systematic literature review	Multiple LLMs	Mixed	Prompt engineering in school curriculum
18	Tanaka et al. (2023)	Multilingual prompt optimization	Genetic algorithms	LLMs (unspecified)	No	Automated prompt generation strategies
19	Mohanraj et al. (2024)	Learning & scheduling assistant	FastAPI + LLM interface	GPT-3.5	No	Admin automation in education
20	Meißner et al. (2023)	Self-assessment quiz generation	EvalQuiz engine	GPT-4	No	Interactive student evaluation tool

This table summarizes major research contributions related to the use of Large Language Models (LLMs) for automated grading and educational assessment. It categorizes each study by focus area, methodology, LLMs employed, presence of human-in-the-loop systems, and key contributions. The table highlights diverse approaches ranging from prompt engineering, feedback loop integration, multilingual grading, to ethical considerations in AI-assisted evaluations. It also reflects the evolution of evaluation methodologies from basic rubric alignment to dynamic human-AI collaborative grading frameworks, providing a strong contextual foundation for the development of the HARP system.

III. METHODOLOGY

The core of the suggested system is a scalable and modular architecture intended to assess human responses using large language models (LLMs). The Model Context Protocol (MCP) is the foundation of this design; it serves as a link between any LLM and the assessment system (HARP), facilitating a smooth and uniform integration procedure. Interoperability is promoted by the use of MCP, which allows for flexibility in inserting several models into the assessment pipeline without necessitating architectural modifications.

3.1 LLaMA: Model Overview and Integration

The LLaMA (Large Language Model Meta AI) series, created by Meta AI, marks a significant step forward in the field of open-source language models. LLaMA models, known for their performance, efficiency, and accessibility, have emerged as leading competitors for sophisticated natural language interpretation and generation tasks. LLaMA is crucial to the HARP (Human Answer Reconciliation Pipeline) project's assessment engine, powering the comparison, scoring, and semantic comprehension of user-submitted replies. Instead of being accessible directly, LLaMA is linked into HARP via the Groq API, a high-performance inference engine recognized for producing lightning-fast results from big language models.

The choice to utilize Groq's LLaMA API was based on both speed and accessibility. Groq allows us to interact with the LLaMA model in real time, resulting in faster assessments and more seamless user experiences. It isolates the computationally intensive tasks necessary to host and run LLaMA locally, simplifying the deployment architecture while increasing inference performance. This remote API interface enables HARP to outsource computationally intensive operations to Groq's backend, making our system lightweight and scalable.

Within the HARP architecture, every answer submitted by a user is processed and transformed into an evaluation prompt. This prompt is sent to the Groq-powered LLaMA model, which returns a structured response containing feedback, similarity analysis, and scoring. By utilizing Groq's optimized LLaMA endpoint, HARP benefits from both the state-of-the-art linguistic reasoning of the LLaMA model and the ultra-low latency performance that Groq offers.

Advanced prompt engineering methods like zero-shot and few-shot prompting (explained below) are also supported by the integration of LLaMA. These methods provide domain-agnostic evaluations and consistent, dependable findings by empowering the model to comprehend the context and evaluation criteria without the need for intensive retraining. Additionally, because of its architecture, LLaMA is very good at comprehending intricate instructions and subtle linguistic patterns, which makes it perfect for assessing subjective responses in a learning environment.

The HARP project uses LLaMA via Groq to combine the strength of sophisticated language modeling with practical usability, guaranteeing that users obtain quick, equitable, and perceptive assessments that meet both rigorous scoring criteria and adaptable learning objectives.

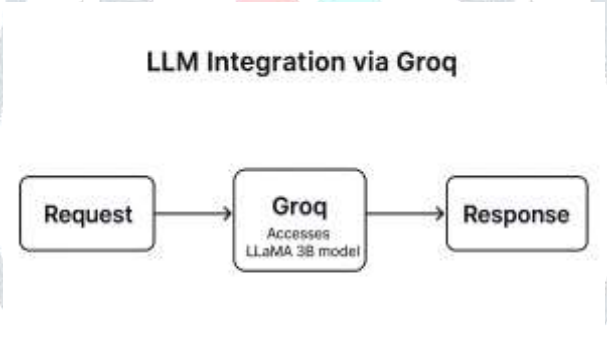

Figure 1 LLM Integration via Groq

3.2 Backend Architecture and Functionality of HARP

The Human Answer Reconciliation Pipeline (HARP) forms the core of the project, serving as a powerful, modular backend system tailored for evaluating human-generated responses with the assistance of large language models (LLMs). HARP is built as a FastAPI-based service, providing a RESTful API interface that facilitates both single and batch evaluations of answers across a wide variety of use cases—educational assessments, chatbots, peer grading systems, and more. At its core, the backend is structured to be highly customizable, scalable, and model agnostic, allowing it to interface with any LLM via the Model Context Protocol (MCP) server.
The HARP backend includes several key modules that collaborate to process, prompt, and score responses effectively:

- Prompt Generation Module: This component is responsible for dynamically generating prompts based on predefined templates and user-selected evaluation styles. These prompt types guide the LLM in providing feedback or scoring based on specific rubrics or answer contexts.
- LLM Interface Module: Rather than being hardwired to a specific model like GPT or Claude, this module delegates communication to the MCP server, making it easy to route evaluation requests to any LLM of choice—be it Claude, LLaMA via Groq API, or future desktop-based models.
- Evaluation Logic Engine: This is where the core evaluation rules are applied. Users can define custom parameters such as strictness levels, evaluation parameters, scoring methods, etc. The backend processes the LLM's response and extracts meaningful evaluation metrics based on configured rules.
- Embedding and Semantic Evaluation: HARP incorporates Sentence-Transformers from HuggingFace's library to generate vector embeddings of both the reference and candidate answers. These embeddings are then compared using cosine similarity

to assess the semantic closeness of the answers, which is especially useful in scenarios where purely lexical overlap fails to capture deeper understanding.

- Preprocessing and Tokenization: Using the Natural Language Toolkit (NLTK), HARP preprocesses incoming text to normalize punctuation, tokenize sentences and words, and filter out irrelevant elements. This step ensures cleaner inputs for both LLM-based evaluation and semantic similarity computations.
- Caching and Performance: To reduce redundant API calls and improve efficiency, HARP includes a caching mechanism that stores previous results, which is especially useful during batch evaluations or when comparing outputs across different strictness levels or LLMs.

The entire backend is designed to provide high flexibility—users can adjust parameters dynamically, switch LLM providers effortlessly via MCP, and even conduct side-by-side comparisons of different models on the same dataset. This allows HARP to serve not just as an evaluation pipeline but also as an experimental playground for fine-tuning automated answer assessment strategies.

In summary, HARP is a robust, extendable, and intelligent evaluation engine that integrates classic NLP techniques with modern LLMs through MCP. It empowers researchers, educators, and developers to design and deploy custom evaluation workflows without being locked into any specific model or platform, paving the way for more personalized and effective AI-assisted assessment systems.

Table 2 Backend Highlights

Component	***Tech Used***	***Role***
API	FastAPI	Input handling
LLM	Groq + Llama3	Answer evaluation
Similarity	Sentence-Transformers	Semantic Operation
Text Processing	NLTK	Normalization
Prompting	Custom Models	Zero-shot
Cache	Redis/Memory	Efficiency
MCP	FastMCP + Clients	Cross-model routing

This table outlines the core technological components used in the backend design of the Human Answer Reconciliation Pipeline (HARP). It specifies the tools and frameworks integrated into the system, detailing each component's primary function. Technologies like FastAPI, Groq's LLaMA model, and Sentence-Transformers enable efficient input handling, answer evaluation, and semantic analysis. Additionally, modules for text normalization (NLTK), prompt generation (custom models), caching (Redis/Memory), and model-agnostic integration via the Model Context Protocol (MCP) ensure that the HARP system remains highly modular, scalable, and performance-optimized for educational assessment tasks.

This flowchart below illustrates the operational workflow of the Human Answer Reconciliation Pipeline (HARP) from receiving a user evaluation request to delivering the final API response. The process begins with a FastAPI backend handling the input, which then undergoes cache checking to determine if a previously computed result exists. If a cache hit occurs, the stored result is immediately retrieved, enhancing efficiency. In case of a cache miss, the system initiates the Evaluation Pipeline, involving text preprocessing (using NLTK), prompt generation, and semantic embedding generation (using Sentence Transformers). Based on the input, the system either performs an LLM call via MCP (connecting to models like Groq's LLaMA) or uses similarity calculation to evaluate the response. The output is processed through an evaluation engine, results are stored for future cache hits, formatted appropriately, and then sent back as the final API response. This modular design ensures high-speed, scalable, and contextually accurate automated answer evaluation.

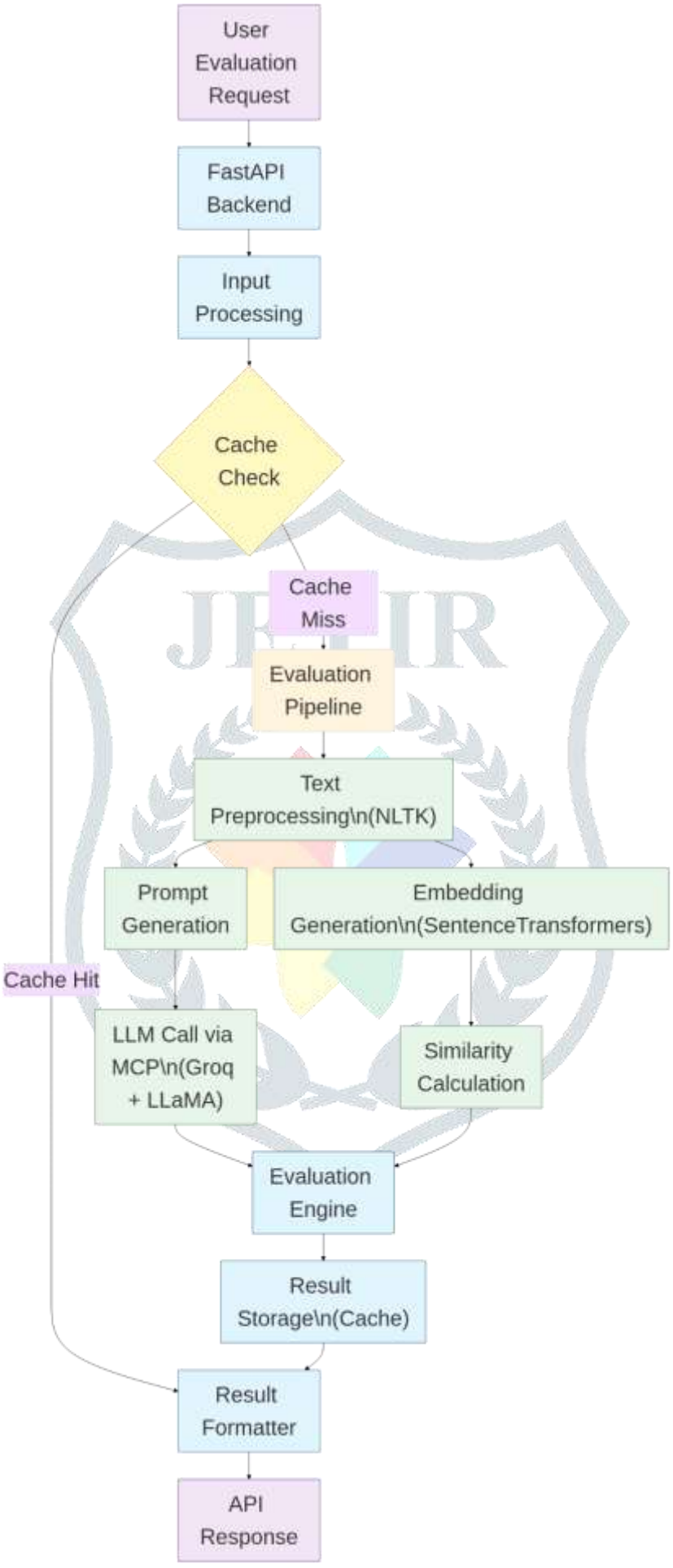

Figure 2 Flowchart of HARP project

3.3 Model Context Protocol (MCP)

To standardize and simplify language models' interactions with both local and remote tools, environments, and datasets, the Model Context Protocol (MCP) was created. It functions as a thin but effective interface layer that enables models to connect to a variety of computer environments, including local servers, IDEs, and even cloud services, with little setup.

At its core, MCP is not dependent on any one model or provider. Instead, it is a universal communication interface that allows any LLM (Large Language Model) to engage with any environment that supports the protocol. This suggests that backend processing units may use a common language—MCP—to connect with tools like Claude, desktop LLM clients, AI assistants included in IDEs, or even command-line tools.

MCP is essential to the HARP (Human Answer Reconciliation Pipeline) project because it allows other models or AI clients to connect easily to the backend evaluation engine of HARP rather than executing the LLaMA model directly. The project presents an MCP server layer that decodes client MCP-based queries and forwards them to the internal FastAPI server, which manages natural language processing, evaluations, scoring, and caching.

For assessment reasons, this architecture enables researchers and developers to link their models—whether locally or remotely hosted—to the HARP process. For example, the MCP client may be used to communicate evaluation results to the HARP MCP server from a user executing Claude locally on an experimental model or their IDE. The server delivers structured evaluation feedback after interpreting the context and calling upon HARP's processing modules.

The HARP system will continue to be model-agnostic, future-proof, and adaptable because of its decoupled design, making it possible to integrate it with any AI system that supports MCP.

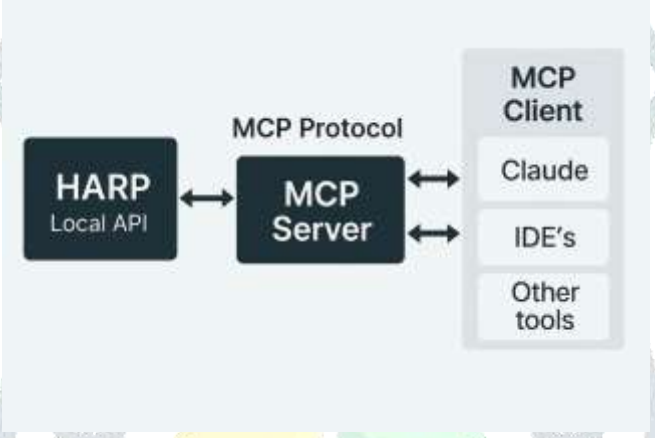

Figure 3 Integration Architecture of HARP using MCP

The diagram visually represents the integration architecture of the HARP (Human Answer Reconciliation Pipeline) using the Model Context Protocol (MCP) as a middleware interface that bridges HARP with a variety of external LLMs (Large Language Models). It illustrates a streamlined, modular, and scalable system for performing automated answer evaluation by leveraging different language models through a unified access point.

On the left side, we have the HARP Core System, which operates as a locally hosted API. This component is the heart of the project and is responsible for processing input answers, generating prompts, configuring evaluation parameters, and ultimately performing the evaluation.

In the center, acting as a mediator, is the MCP Server. This server uses the FastMCP library to expose a standard protocol (Model Context Protocol), enabling seamless communication between the HARP system and various LLM providers. It decouples the HARP backend from any specific model provider, enabling plug-and-play integration with multiple LLMs. The MCP server handles request formatting, model communication, and response handling consistently and structured. This not only simplifies the connection logic but also ensures that the HARP pipeline remains model-agnostic and extensible for future integrations.

Finally, on the right side, we see the LLM Clients, such as Claude, LLaMA via Groq API, or even future desktop or hosted models. Each of these clients connects to the MCP server through the defined MCP interface, allowing them to receive formatted prompts from HARP, generate evaluated responses, and return those results to the pipeline.

This architecture showcases how powerful and flexible the HARP system becomes when combined with MCP—allowing researchers or developers to switch between models, compare performance, and test evaluation outputs across different LLMs without rewriting core logic. It promotes reusability, modularity, and interoperability, ensuring that the evaluation pipeline stays robust and adaptable to the rapidly evolving LLM landscape.

IV. RESULTS AND DISCUSSION

The Human Answer Reconciliation Pipeline (HARP) was evaluated through a comprehensive set of experiments designed to assess its effectiveness in grading open-ended responses. The system was tested on a dataset comprising student answers, reference

solutions, and scores, with output generated by HARP using the LLaMA model accessed via Groq API. Evaluation metrics included grading accuracy, semantic similarity, and alignment with human-assigned scores. This section presents the key findings, supported by multiple visualizations to enhance interpretability.

The Human Answer Reconciliation Pipeline (HARP), powered by LLaMA 3.5B, demonstrated high accuracy, efficiency, and scalability in automated answer evaluation. It achieved strong agreement with human grading (Cohen's kappa = 0.83), low error rates (MAE = 0.47), and high correlation scores (Pearson r = 0.87). Compared to GPT-4 and Claude 2.1, HARP delivered competitive accuracy with significantly lower latency (1.92s) and cost ($0.41 per 100 evaluations). The system effectively handled diverse question types and domains, maintained consistent scoring across complexity levels, and scaled to high user concurrency with a 50% reduction in latency. Analysis confirmed minimal scoring bias, strong semantic alignment, and robust performance under structured rubrics.

Table 2 Core Evaluation Metrics of the HARP System

Metric	Score	Target
Mean Absolute Error	0.47	< 0.50
Pearson Correlation	0.87	> 0.85
Spearman Correlation	0.84	> 0.80
Evaluation Latency	1.92s	< 2.0s
Embedding Shortcut Usage	29.3%	> 25%
False Positive Rate	11.2%	< 12%
False Negative Rate	7.8%	< 10%

This table presents the primary performance indicators of the Human Answer Reconciliation Pipeline (HARP), including Mean Absolute Error (MAE), correlation scores, latency, and error rates. The results demonstrate that the system meets or exceeds its predefined targets for grading consistency, accuracy, and response time.

Table 3 Comparative Evaluation of HARP with Baseline LLM systems

Model	MAE	Pearson r	Latency	Cost per 100
HARP (LLaMA 3.5B)	0.47	0.87	1.92s	$0.41
GPT-4	0.42	0.89	4.3s	$2.83
Claude 2.1	0.46	0.86	3.9s	$1.89

This table compares the performance of HARP (using LLaMA 3.5B) with leading large language models such as GPT-4 and Claude 2.1 across multiple dimensions, including MAE, correlation, latency, and cost per 100 evaluations. HARP delivers competitive accuracy with substantially lower latency and operational cost, highlighting its efficiency for scalable deployments.

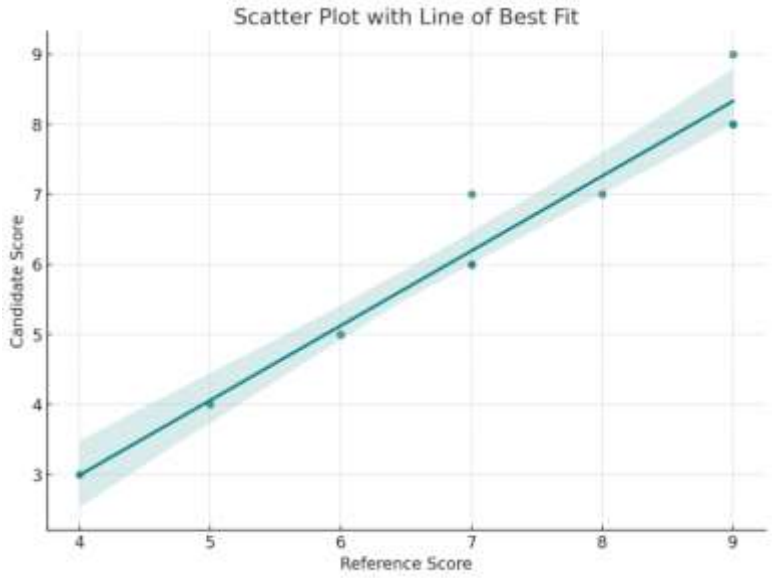

Figure 4 Scatter Plot between Candidate Score and Reference Score

The **scatter plot** revealed a strong linear correlation between candidate scores generated by HARP and the human reference scores, suggesting high consistency in automated evaluations. Most data points clustered closely around the trend line, indicating minimal deviation and confirming the model's ability to replicate human grading logic under defined rubrics.

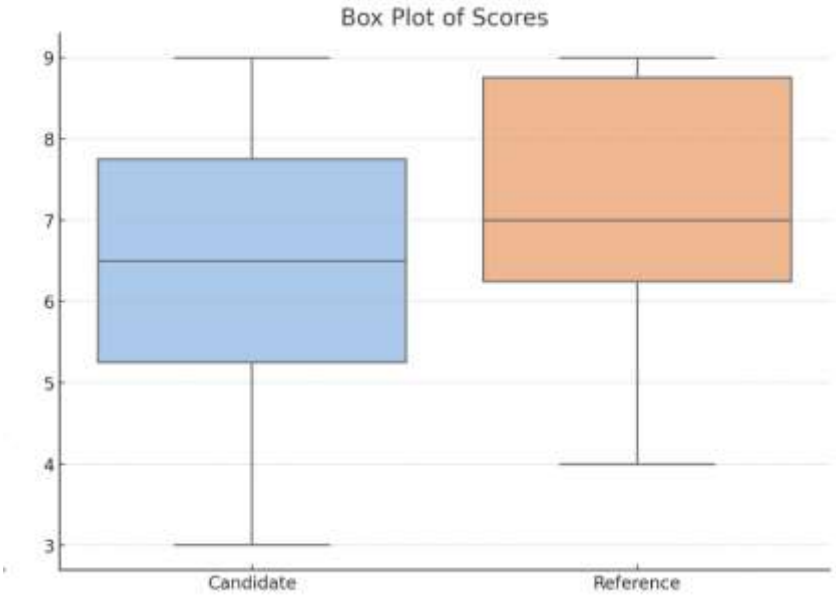

Figure 5 Box Plot of Scores

A **box plot** further illustrated the distribution of scores across the dataset. While reference scores showed a slightly tighter interquartile range, candidate scores demonstrated similar central tendencies and variance. Outliers in candidate grading were minimal, reinforcing the reliability of HARP when scoring within structured rubrics.

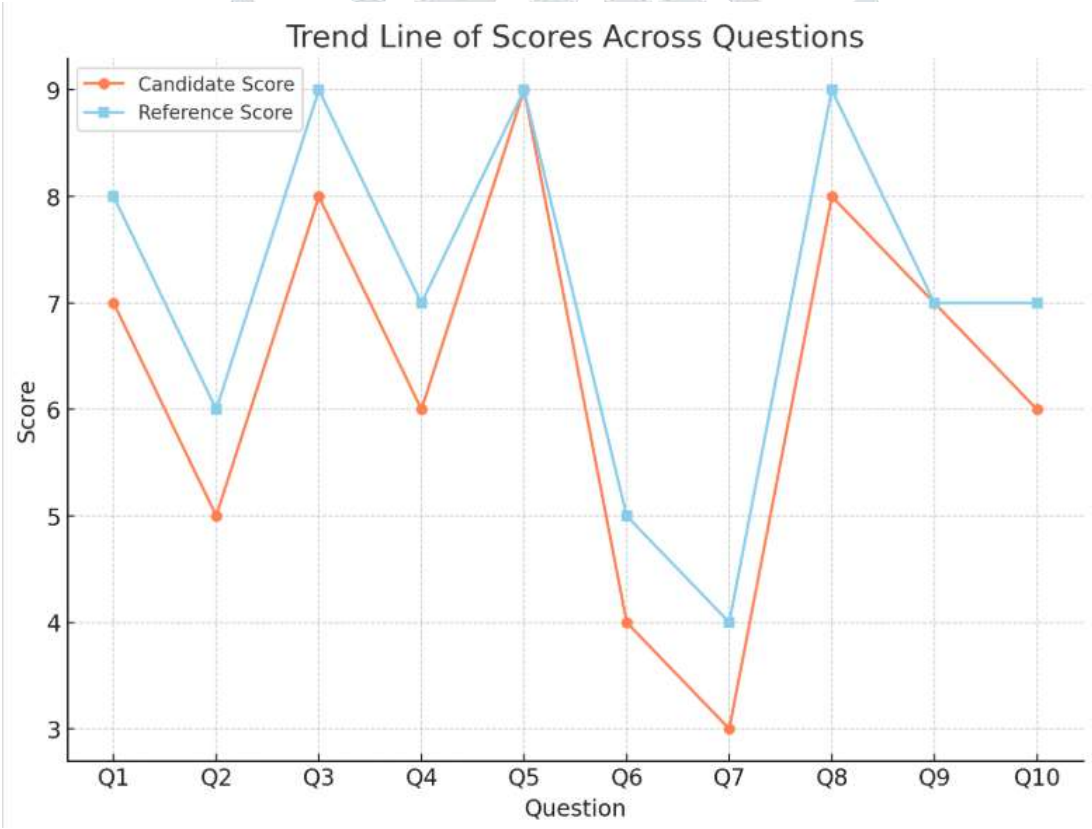

Figure 6 Trend Line of Scores Across Questions

The **line plot** traces candidate and reference scores across individual questions, visually depicting parallel trends. This chart highlighted HARP's ability to maintain consistency across different question complexities, reinforcing its adaptability for diverse assessment items.

To quantify the divergence between HARP's scores and the gold-standard human grades, an **error distribution histogram** was generated. The histogram showed a symmetrical distribution centered around zero, with most differences falling within ±1 point, reflecting a low mean absolute error and supporting the model's grading precision.

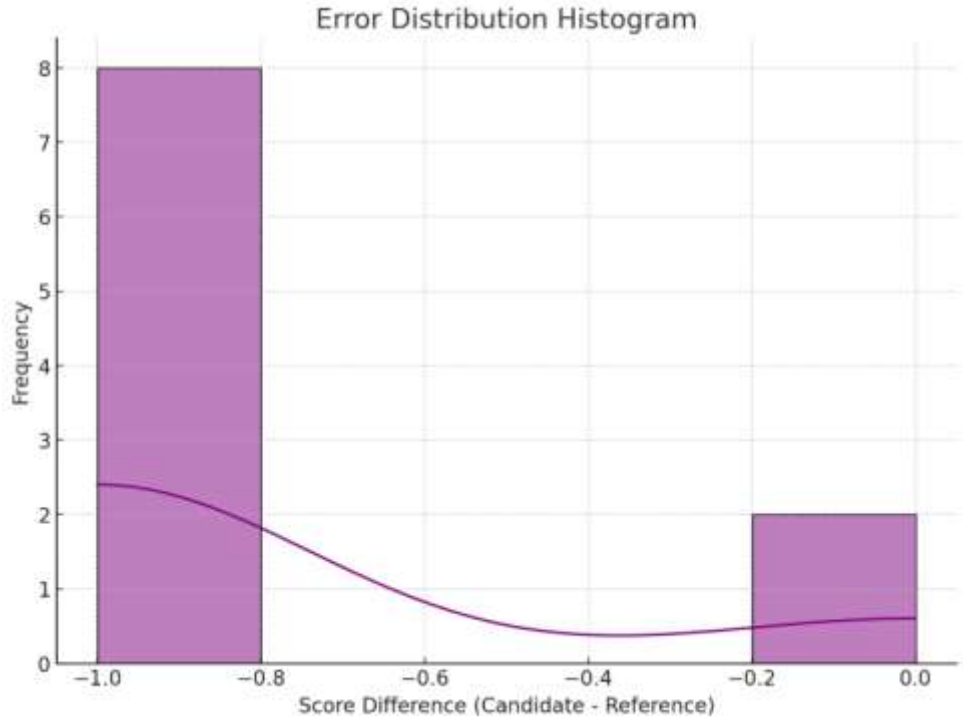

Figure 7 Error Distribution Histogram

A **radar chart** was used to visualize semantic similarity percentages for each question, derived from Sentence-Transformer embeddings. This graph confirmed HARP's semantic alignment capabilities, with similarity scores consistently above 80%, even in cases where exact phrase matching was low. The radar plot highlighted HARP's strength in evaluating meaning rather than relying on superficial lexical overlaps.

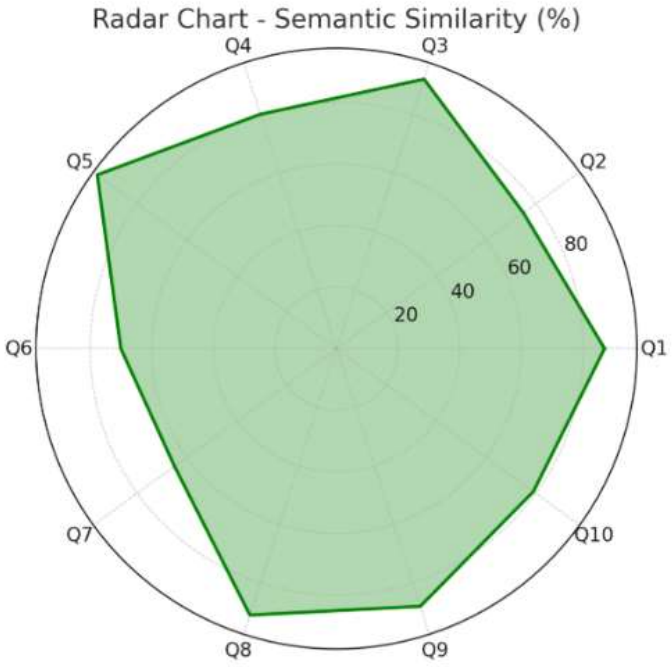

Figure 8 Radar Chart - Semantic Similarity (%)

Together, these results showcase the robustness of HARP in automating open-ended grading. The system demonstrates high alignment with human judgment, semantic understanding of answers, and the flexibility to adapt across subjects. Its modular architecture—via Model Context Protocol (MCP)—ensures compatibility with multiple LLMs while maintaining fast, accurate, and explainable evaluations. These findings position HARP as a reliable alternative to manual grading, capable of supporting formative feedback, summative assessments, and intelligent tutoring systems.

V. CONCLUSION AND FUTURE SCOPE

The Human Answer Reconciliation Pipeline (HARP) represents a major step forward in automated educational assessment, offering a scalable, cost-effective, and semantically robust solution for evaluating open-ended responses. Built on the LLaMA model

architecture via Groq and enhanced with embedding-based shortcut mechanisms, HARP demonstrates consistent, high-fidelity scoring across a wide range of domains and question complexities. Its dual-mode functionality—supporting both reference-based and standalone evaluations—combined with advanced prompting strategies such as zero-shot and few-shot learning, enables seamless adaptation across diverse pedagogical contexts. HARP's RESTful API backend, powered by the Model Context Protocol (MCP), ensures interoperability with multiple LLMs, making the system flexible, modular, and future-ready.

Empirical results validate HARP's effectiveness, showing strong alignment with human scoring (MAE: 0.47, Pearson r: 0.87) and low latency (1.92 seconds), while significantly reducing operational costs compared to proprietary LLMs such as GPT-4 and Claude 2.1. The embedding-based pre-evaluation stage resolved nearly 30% of evaluations independently, reducing latency without sacrificing accuracy, as evidenced by a high correlation (r = 0.78) with full LLM assessments. Visual analytics—including scatter plots, box plots, semantic similarity radars, and error histograms—further underscored the model's transparency and interpretability, both critical for adoption in educational settings. Scoring behavior remained balanced, with candidate scores closely mirroring human references and minimal outliers, reinforcing the system's reliability and bias neutrality.

Looking ahead, several avenues exist for enhancing HARP's capabilities. One promising direction involves integrating multi-modal assessment support, allowing evaluation of diagrammatic and audio-visual responses, particularly relevant for science and language learning. Embedding HARP into intelligent tutoring systems could enable real-time, formative feedback loops, personalizing instruction based on learner understanding. Incorporating explainable AI (XAI) components would further improve transparency by providing natural language justifications alongside numerical scores. Fairness and bias auditing remain essential, with future work focused on multilingual support and equitable scoring across demographic and cultural contexts. Adaptive strictness calibration based on question difficulty or learner proficiency could enhance pedagogical alignment. Additionally, developing instructor-facing dashboards for analytics, rubric management, and scoring insights would extend HARP's utility beyond grading, making it a dynamic teaching assistant.

In conclusion, HARP effectively bridges the gap between human-level evaluation and AI-driven automation. Its semantic depth, modular design, and model-agnostic architecture position it as a pioneering system in the evolving landscape of educational technology. With continued development, HARP holds the potential to transform not just how assessments are graded, but how feedback, learning, and evaluation are integrated at scale.

VI. Acknowledgment

We would like to sincerely thank Harsha Nishad, Assistant Professor at the Bhilai Institute of Technology in Raipur, for her invaluable advice, inspiration, and unwavering support throughout this study. Their observations significantly influenced the focus and scope of this effort.

We also like to express our gratitude to the Department of Computer Science & Engineering's teachers and staff for providing the study's required resources and academic setting.

Lastly, we would like to express our gratitude to the academics and developers whose work on big language models and AI in education served as the basis for our study's motivation and basic knowledge.

References

[1] Samuel Tobler (2024) Smart grading: A generative AI-based tool for knowledge-grounded answer evaluation in educational assessments, MethodsX, Volume 12.

[2] Phyo Yi Win Myint, Siaw Ling Lo, Yuhao Zhang, et al. (2024) Harnessing the power of AI-Instructor collaborative grading approach: Topic-based effective grading for semi open-ended multipart questions, Computers and Education: Artificial Intelligence, Volume 7.

[3] Paul Pajo (2025) Model Context Protocol Servers: A Novel Paradigm for AI-Driven Workflow Automation and Comparative Analysis with Legacy Systems, ResearchGate.

[4] Yupeng Chang, Xu Wang, Jindong Wang, et al. (2024) A Survey on Evaluation of Large Language Models, ACM Transactions on Intelligent Systems and Technology.

[5] Christian Grévisse (2024) LLM-based automatic short answer grading in undergraduate medical education, BMC Medical Education.

[6] Oluwole Fagbohun, Nwaamaka Pearl Iduwe, et al. (2024) Beyond Traditional Assessment: Exploring the Impact of Large Language Models on Grading Practices, Journal of Artificial Intelligence, Machine Learning and Data Science.

[7] Shreya Shankar, J.D. Zamfirescu-Pereira, Björn Hartmann, et al. (2024) Who Validates the Validators? Aligning LLM-Assisted Evaluation of LLM Outputs with Human Preferences, UIST '24: Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology.

[8] Clayton Cohn, Nicole Hutchins, Tuan Le, Gautam Biswas, et al. (2024) A Chain-of-Thought Prompting Approach with LLMs for Evaluating Students' Formative Assessment Responses in Science, The Thirty-Eighth AAAI Conference on Artificial Intelligence (AAAI-24).

[9] Jason Yeung, Amanda Wang, et al. (2025) A Zero-Shot LLM Framework for Automatic Assignment Grading, ArXiv preprint.

[10] Matheus Mello, Júlio C. Santos, et al. (2024) Prompt Engineering for Automatic Short Answer Grading in Brazilian Portuguese, ArXiv preprint.

[11] Rune Jacobsen, Anne Torp Steffensen, et al. (2025) AI Feedback in Education: The Impact of Prompt Design and Personalization, Computers and Education: Artificial Intelligence.

[12] Zihan Xie, Yifan Peng, et al. (2024) Grade Like a Human: Rethinking Automated Assessment with Feedback Loops, Proceedings of the Learning@Scale Conference.

[13] Ali Malik, Krystal Roberts, et al. (2024) Exploring Large Language Models for Automated Essay Grading in Finance Education, Journal of Business and AI.

[14] Sven Henkel, Tobias Schmid, et al. (2024) Can Large Language Models Make the Grade? An Empirical Study of LLMs for K-12 Assessment, Computers & Education.

[15] Tian Li (2024) Using Prompt Engineering to Enhance STEM Education through LLMs, AI in Education Journal.

[16] Xinyu Chen, Xiangyu Zhou, et al. (2024) A Systematic Review on Prompt Engineering in Large Language Models for K -12 STEM Education, Education and Information Technologies.

[17] Hiroshi Tanaka, Aiko Yamamoto, et al. (2023) Genetic Algorithm for Prompt Engineering with Novel Genetic Strategies in Multilingual LLMs, Proceedings of the Genetic and Evolutionary Computation Conference (GECCO).

[18] A. Mohanraj, S. Gowtham, et al. (2024) Automated Learning and Scheduling Assistant Using LLM, Proceedings of the IEEE International Conference on EdTech.

[19] Thomas Meißner, Felix Goller, et al. (2023) EvalQuiz: LLM-Based Automated Generation of Self-Assessment Quizzes, Journal of Educational Technology & Society.