Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 2, April 2026
DOI : https://doi.org/10.33395/sinkron.v10i2.16020
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Comparative Evaluation of Rule-Based and Large Language Models for Financial Transaction Extraction in Chatbots

Verdymas Atma Yulianto[1)*], Erba Lutfina [2)], Galuh Wilujeng Saraswati[3)]
[1,2,3)] Program Studi Sistem Informasi, Fakultas Ilmu Komputer, Universitas Dian Nuswantoro, Indonesia.
[1)]atmaverdymas@gmail.com, [2)] erba.lutfina@dsn.dinus.ac.id, [3)] galuhwilujengs@dsn.dinus.ac.id

Submitted :Mar 26, 2026 | **Accepted** : Apr 19, 2026 | **Published** : Apr 21, 2026

Abstract: The increasing use of digital financial services requires tools that allow users to record and manage personal financial transactions efficiently. However, many conventional applications still rely on form-based interfaces that may reduce user engagement in daily financial recording. This study evaluates a conversational personal financial recording system that uses natural language processing to convert informal user messages into structured transaction data. The system was developed using the prototyping method and implemented through a messaging-based interface, a backend service, and a natural language processing module. The evaluation used a dataset of 300 conversational financial messages annotated with intent, amount, and category. The study compares a rule-based baseline with two open-source large language models, using intent accuracy, entity extraction metrics, output validity, user acceptance testing, and the system usability scale. The results show that open-source large language models achieved the best performance across the natural language processing evaluation, with strong intent classification, high entity extraction quality, and complete output validity. The user acceptance testing involving 30 participants produced an average success rate of 97.3%, while the system usability scale score reached 82.5, indicating excellent usability. These findings suggest that prompt-constrained large language models can improve conversational financial recording by providing reliable structured extraction and an accessible user experience.

Keywords: chatbot; financial management; messaging applications; natural language processing; prototyping method

INTRODUCTION

The rapid development of information technology has significantly transformed the way individuals access financial services and manage personal finances. Digital platforms now allow users to perform financial transactions, monitor spending, and retrieve financial information through mobile banking, e-wallets, and messaging applications (Dwi Astuti & Soleha, 2023; Mahastanti & Utoyo, 2022; Rachmawati et al., 2023). However, although access to digital financial services has increased, personal financial recording remains a persistent behavioral challenge because many users still rely on memory, postpone data entry, or abandon financial recording when the interface is perceived as burdensome. This situation reflects a practical and scientific challenge in which personal financial data is often expressed in natural language, while most finance applications still require users to convert that information into rigid and structured form fields.

This behavioral barrier is further reflected in the broader national landscape, according to the National Survey of Financial Literacy and Financial Inclusion (SNLIK) 2025 published by the Financial Services Authority (Otoritas Jasa Keuangan), Indonesia's financial literacy index reached 66.46%, while financial inclusion reached 80.51% (Otoritas Jasa Keuangan & Badan Pusat Statistik, 2025). These results indicate that although access to financial services has increased significantly, the level of financial literacy among the public remains relatively limited. Financial literacy plays an important role in helping individuals plan budgets, control spending behavior, and make appropriate financial decisions (Dwi Astuti & Soleha, 2023; Ricaldi et al., 2022). Previous studies also emphasize that financial recording behavior is strongly associated with financial awareness and long-term financial planning (Mahastanti & Utoyo, 2022; Rachmawati et al., 2023).

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 2, April 2026
DOI : https://doi.org/10.33395/sinkron.v10i2.16020
e-ISSN : 2541-2019
p-ISSN : 2541-044X

From a scientific standpoint, the main challenge is not only user convenience but also the transformation of ambiguous conversational input into valid structured data. Financial messages are often brief, informal, and context-dependent. They may contain abbreviations, implicit amounts, multiple transactions in a single sentence, or incomplete category information. In NLP terms, this requires reliable intent interpretation and entity extraction, followed by structured output generation that is consistent with a predefined schema. Recent studies on conversational AI and information extraction show that extracting meaningful entities and relations from free text remains nontrivial, even for modern large language models (LLMs), because errors may occur in structured output generation, reproducibility, and handling of domain-specific language. In financial NLP, these challenges are amplified by informal numerical expressions, abbreviated currency formats, and context-dependent transaction descriptions that complicate entity extraction and intent interpretation (Chandrakala et al., 2024; Dagdelen et al., 2024; Dave & Chowanda, 2024; Dietrich & Hollstein, 2025; Ferrera et al., 2025; Ouaddi et al., 2025).

The difficulty of this extraction task is one reason why many existing financial management applications still default to manual data entry through structured forms. This mechanism requires users to repeatedly input transaction attributes such as amount, category, and description. Previous studies indicate that form-based financial recording systems often reduce usability and discourage users from consistently recording their daily financial activities (Mahastanti & Utoyo, 2022; Rachmawati et al., 2023). In addition, most chatbot studies in the literature have focused on customer service, helpdesk automation, education, or generic question-answering, while relatively few have examined conversational systems as tools for personal financial transaction extraction (Eriana & Subariah, 2025; Putri Oktavianita & Andreas Sutanto, 2024; Sasmita et al., 2025; Sezgin et al., 2024). Despite these developments, empirical evidence on how different approaches perform in transforming conversational financial messages into structured transaction records remains limited in the current literature on conversational information extraction and financial NLP (Dagdelen et al., 2024; Dave & Chowanda, 2024). In particular, the relative capability of rule-based extraction and open large language models to interpret informal financial expressions and generate reliable structured outputs has not been widely examined in conversational personal finance contexts (Ferrera et al., 2025; Ouaddi et al., 2025).

Addressing these combined behavioral and technical limitations, recent advances in Artificial Intelligence (AI), particularly Natural Language Processing (NLP), provide new opportunities to improve interaction between users and digital systems. NLP enables computer systems to interpret human language and extract meaningful information from textual input, allowing users to interact with applications using natural conversational messages (Bai et al., 2025; Puspitasari et al., 2024; Wobst et al., 2025). On this basis, this study evaluates a rule-based baseline together with two open large language models, Qwen2.5 and Llama 3.1, for financial transaction extraction in conversational chatbots. The system is designed to convert user messages into structured JSON transaction records and to assess not only extraction performance, but also output validity, error patterns, and system usability.

This study proposes a comparative evaluation of financial transaction extraction in conversational chatbots. The research examines the relative performance of a rule-based extraction approach and open-source large language models in interpreting informal financial messages and generating structured transaction records. In addition to the technical evaluation, the study implements a prototype chatbot developed using the prototyping method to support iterative system refinement and usability validation. The system is further evaluated through user acceptance testing and usability assessment to examine its practical feasibility for conversational personal financial recording (Alda, 2023; Alfath et al., 2024; Pressman, 2010).

LITERATURE REVIEW

Natural Language Processing has shifted from fixed-rule text handling toward methods that better capture meaning, context, and domain-specific language. In conversational systems, this shift matters because user input is often informal, short, and inconsistent, so language understanding must go beyond surface patterns (Puspitasari et al., 2024; Wobst et al., 2025).

Chatbot studies show a clear divide between rule-based systems and more adaptive approaches. Rule-based designs offer stronger control and consistency, while newer intent-centered and embedding-based methods reduce manual rule creation and improve coverage of diverse utterances. Recent work also shows that intent detection remains difficult when data are sparse or heterogeneous, making clustering, embeddings, and prompt-based methods increasingly relevant (Chandrakala et al., 2024; Ferrera et al., 2025).

Intent detection research further indicates that no single approach dominates across domains. LLM-generated augmentation can help classifier training in low-data settings, while comparative studies show that GPT, BERT, LLaMA, and RoBERTa vary in performance depending on the task and dataset. This suggests that model choice should follow the data characteristics and the target domain rather than assume one universal solution (Benayas et al., 2024; Ouaddi et al., 2025).

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 2, April 2026
DOI : https://doi.org/10.33395/sinkron.v10i2.16020
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Named entity recognition and information extraction are even more central to financial transaction extraction because the system must identify amounts, categories, and descriptions, not only user intent. CRF and BiLSTM-CRF remain effective in specific domains, but their performance is often limited by linguistic variation and small datasets. Multi-task approaches such as DIET reduce the separation between intent and entity extraction, while financial NER studies show that numeric values and informal expressions remain difficult to extract reliably (Annisa et al., 2024; Dave & Chowanda, 2024; Widiyanti et al., 2023; Wildannissa Pinasti & Lya Hulliyyatus Suadaa, 2024; Zahra et al., 2022).

Large language models offer a different route because they can produce structured outputs such as JSON and perform joint extraction tasks in a single pipeline. At the same time, recent studies note concerns about hallucination, reproducibility, and inconsistent structured output, which become more critical in financial contexts where small extraction errors can affect data integrity. Open-source models such as Qwen and Llama are therefore attractive because they allow controlled prompting, local use, and adaptation to informal or localized language (Batura et al., 2025; Chen et al., 2025; Dietrich & Hollstein, 2025; Yan et al., 2025).

Financial NLP studies show that text-based methods are increasingly useful for market analysis, text mining, and automated service support, while chatbot studies in finance and other domains still mostly focus on general interaction or form-based workflows. In parallel, prototyping remains a suitable approach for chat-based systems because it supports iterative refinement and usability testing. However, empirical evidence is still limited on how rule-based extraction compares with open-source LLMs when both are used to transform conversational financial messages into validated structured records. This study addresses that gap by evaluating a rule-based baseline alongside Qwen2.5 and Llama 3.1 for financial transaction extraction in conversational chatbots, while also considering output validity and usability in a prototype system (Alda, 2023; Alfath et al., 2024; Bai et al., 2025; Eriana & Subariah, 2025; Mahastanti & Utoyo, 2022; Rachmawati et al., 2023; Sasmita et al., 2025; Sezgin et al., 2024; Tandrio & Fianty, 2026; Wobst et al., 2025).

METHOD

System Development Method

This research applies the prototyping development method to iteratively design and refine the chatbot system. The prototyping model consists of several stages: communication, quick planning, modeling quick design, construction of prototype, and deployment with feedback, which are commonly used to support iterative system refinement and user-centered development in software engineering (Alda, 2023; Pressman, 2010).

The process begins with the communication stage where the researchers identify the problems related to personal financial recording and gather user requirements for a conversational financial management system. After the requirements are identified, the quick planning stage is conducted to determine the main system features, chatbot interaction flow, and the overall architecture of the application.

Next, the modeling quick design stage focuses on creating a preliminary system design, including the chatbot interaction model, financial transaction data structure, and the integration between frontend, backend, and AI services. Based on this design, the construction of prototype stage is carried out to develop the initial version of the chatbot that implements the core functionality such as receiving user messages, processing natural language input, and storing financial transaction data.

Finally, the system enters the deployment, delivery, and feedback stage where the prototype is tested through user interactions. Feedback obtained from this stage is used to improve the system iteratively until the chatbot system achieves the expected functionality and usability.

System Architecture

The system architecture describes how the main components interact to process user messages and generate financial transaction records. The architecture consists of four components the user, frontend interface, backend service, and AI service. Users communicate with the system through a chat interface on the frontend, implemented using Vue.js version 3, a reactive component-based framework suitable for real-time conversational interfaces. The frontend captures user messages and sends them to the backend through REST-based API endpoints that support message processing, transaction recording, transaction history retrieval, and validation responses.

The backend functions as the central processing layer that manages requests, validates user input, communicates with the AI service, and stores transaction data in the database. It is implemented using Laravel version 12, which supports structured routing, authentication, and API management. The AI service applies natural language processing using large language models to interpret user messages and extract financial information. The extracted data is returned to the backend for validation, storage, and delivery to the user through the frontend interface.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 2, April 2026
DOI : https://doi.org/10.33395/sinkron.v10i2.16020
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Rule-Based Method

Fig. 1 Rule-based financial transaction extraction pipeline

The rule-based baseline extracts financial transaction attributes using deterministic linguistic rules. The pipeline consists of several sequential steps including text normalization, keyword detection, monetary value extraction using regular expressions, category mapping, and JSON transaction generation. Token normalization is applied to standardize informal expressions and abbreviations commonly used in conversational financial messages. Keyword detection is used to identify indicators of income or expense such as purchase verbs, payment expressions, or income-related terms. Monetary values are then extracted using numeric and currency patterns, while transaction categories are determined through a predefined keyword–category mapping table. After the relevant elements are detected, the extracted values are assembled into the same structured JSON schema used by the LLM approaches. This rule-based pipeline serves as a transparent baseline representing traditional information extraction techniques.

LLM Extraction Method

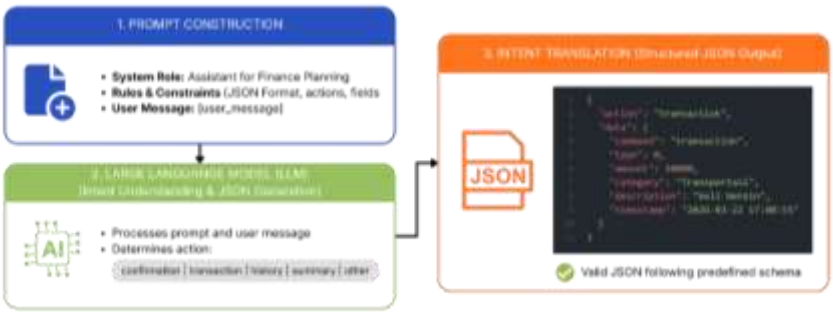

Fig. 2 Natural language processing pipeline for conversational financial message extraction

The NLP pipeline processes conversational user messages through prompt construction, model inference, structured JSON generation, and output validation. In this study, two large language models are evaluated, namely Qwen2.5 and Llama 3.1. Both models are used in an inference setting without additional fine-tuning so that the comparison focuses on their capability to interpret conversational financial messages using prompt instructions. The generated output is converted into a predefined JSON structure containing fields such as command, type, amount, category, description, and timestamp. The backend then validates the generated structure before storing the transaction record.

Fig. 3 Large language models prompt engineering in extracting financial transaction

Prompt engineering is applied to control how the models interpret user messages and produce structured outputs. The prompt defines the role of the model, the extraction task, response rules, and the required output schema. The model is instructed to identify financial transaction attributes including transaction type, monetary amount, category, and description from the conversational message. The response must follow a predefined JSON format without additional explanatory text. Output constraints and predefined category lists are included in the

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 2, April 2026
DOI : https://doi.org/10.33395/sinkron.v10i2.16020
e-ISSN : 2541-2019
p-ISSN : 2541-044X

prompt to reduce response variability and ensure consistent outputs across both Qwen2.5 and Llama 3.1 during the evaluation process.

System Environment

The system is developed and tested in a local development environment consisting of a backend service, database, and AI processing service. The development machine is equipped with an Intel Core i7-13650HX processor, 32 GB RAM, and an NVIDIA GeForce RTX 5050 Laptop GPU, which is sufficient to support backend execution and API-based interaction with the LLM service.

The backend service manages API requests, transaction processing, and database storage, while the AI service performs natural language interpretation through the LLM inference API. This environment allows the system to simulate real conversational interactions and evaluate the NLP pipeline during development.

Dataset

The evaluation dataset is stored in a CSV file with four columns text, intent, amount, and category. The text column represents the user message, while intent, amount, and category are used as ground truth annotations for evaluation. The dataset contains 300 records and the intent label includes two classes expense and income representing outgoing and incoming financial transactions.

Table 1
Example of Dataset Records

id	text	type	amount	category
1	bayar ukt semester ini 4.5jt	expense	4500000	Pendidikan
2	gaji pokok bulan april cair 6.500.000	income	6500000	Gaji
3	isi bensin motor di spbu 35k	expense	35000	Transportasi
4	makan siang nasi padang lauk rendang 25rb	expense	25000	Makanan & Minuman
5	beli token pln 100.000	expense	100000	Listrik & Air
…	…	…	…	…
300	jual item game online laku 150.000	income	150000	Bonus

In the collected dataset, approximately 77% of the records represent expense transactions and 23% represent income transactions. The columns intent, amount, and category are used as the ground truth annotations, while the text column serves as the input message for the extraction models.

The dataset is stored in CSV format and processed using a Python evaluation script. Each row is sequentially evaluated by the rule-based parser, Qwen2.5, and Llama 3.1, and the predicted outputs are compared with the ground truth annotations to calculate the evaluation metrics.

Experimental Design and Evaluation

The experimental evaluation in this study is divided into three main parts, namely User Acceptance Testing, System Usability Scale, and Rule-Based and LLM Evaluation. User Acceptance Testing is used to measure whether the chatbot functions correctly in practical usage scenarios. System Usability Scale is used to measure perceived usability from the user perspective. The SUS questionnaire consists of ten statements rated on a five-point Likert scale from strongly disagree to strongly agree. The score for each item is adjusted according to the SUS scoring procedure, summed across all items, and multiplied by 2.5 to obtain a final usability score ranging from 0 to 100. Rule-Based and LLM Evaluation is used to compare the performance of the rule-based baseline, Qwen2.5, and Llama 3.1 in intent detection, entity extraction, and JSON validity.

The experiment is designed to compare the rule-based baseline, Qwen2.5, and Llama 3.1 on the same set of conversational financial inputs. The input data consists of natural language messages that express financial transactions in informal language, including explicit amounts, implicit amounts, single transactions, and messages containing multiple financial actions. Each sample is assigned a reference output that is used to assess whether the extracted transaction matches the expected structure.

The evaluation focuses on three aspects. The first is functional correctness, which is measured through User Acceptance Testing. The second is structured output validity, which checks whether the model produces JSON that follows the required schema. The third is NLP performance, which evaluates the extraction quality using accuracy, precision, recall, and F1-score. These metrics are commonly used in intent classification and entity extraction studies to assess model performance in conversational systems (Dave & Chowanda, 2024; Ouaddi et al., 2025; Sasmita et al., 2025).

For baseline comparison, the rule-based approach is used to show how far predefined lexical rules and regular expressions can process the same inputs without language modeling. The two LLMs are then compared against this baseline to determine whether they provide better extraction quality and more robust handling of informal

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 2, April 2026
DOI : https://doi.org/10.33395/sinkron.v10i2.16020
e-ISSN : 2541-2019
p-ISSN : 2541-044X

language. In addition, the comparison makes it possible to observe how well each model handles ambiguity, missing context, and mixed transaction expressions.

Table 2 Experimental Comparison Methods

Method	Type	Purpose
Rule-based	Baseline	Keyword and pattern extraction
Qwen2.5	LLM	Prompt-based financial extraction
Llama 3.1	LLM	Prompt-based financial extraction

System Evaluation Metrics

The NLP evaluation focuses on three aspects consist of intent detection, entity extraction, and JSON validity, measured for the rule-based baseline, Qwen2.5, and Llama 3.1.

Intent detection is evaluated using accuracy, since the task predicts one correct label between two classes, expense and income. Accuracy measures the proportion of correct predictions compared to all evaluated samples and is commonly used in classification evaluation (Sokolova & Lapalme, 2009).

$$Accuracy = \frac{Correct\ Predictions}{Total\ Predictions}$$

Entity extraction evaluates the ability to identify the transaction attributes amount and category from the user message. The extraction performance is measured using precision, recall, and F1-score, which are standard metrics in information extraction and named entity recognition studies (Ouaddi et al., 2025; Sokolova & Lapalme, 2009).

The evaluation uses the concepts of true positive (TP), false positive (FP), and false negative (FN). A TP occurs when the predicted entity matches the ground truth annotation, FP occurs when the predicted entity is incorrect, and FN occurs when the entity exists in the ground truth but is not extracted (Sokolova & Lapalme, 2009).

$$Precision = \frac{True\ Positive\ (TP)}{True\ Positive\ (TP) + False\ Positive\ (FP)}$$

$$Recall = \frac{True\ Positive\ (TP)}{True\ Positive\ (TP) + False\ Negative\ (FN)}$$

$$F1 - score = 2x\frac{Precision\ x\ Recall}{Precision + Recall}$$

In addition to extraction quality, JSON validity is measured to determine whether the generated output follows the predefined structured schema required by the backend system. This metric represents the proportion of responses that produce valid and parsable JSON objects.

$$JSON\ Validity = \frac{Valid\ JSON\ Outputs}{Total\ Outputs}$$

Besides the NLP evaluation, the overall system performance is also assessed through User Acceptance Testing (UAT) and the System Usability Scale (SUS) to evaluate functional correctness and user experience. UAT is conducted using several predefined interaction scenarios that represent common user activities such as login, transaction recording, history retrieval, and financial summary requests. Each scenario verifies whether the chatbot produces the expected system response and correctly processes the requested operation. Meanwhile, SUS is used to capture user perceptions regarding system usability after interacting with the chatbot. Participants complete a ten-item questionnaire using a five-point Likert scale, and the resulting scores are aggregated to indicate the overall usability level of the system. In this study, both UAT and SUS evaluations involve 30 participants, representing users who interact directly with the chatbot to perform the predefined scenarios before providing usability feedback.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 2, April 2026
DOI : https://doi.org/10.33395/sinkron.v10i2.16020
e-ISSN : 2541-2019
p-ISSN : 2541-044X

RESULT

System Implementation

The system provides a chatbot interface for recording personal financial transactions through conversational messages. Users authenticate their account with a token from the web application and then start the chatbot session using the /login <token> command. After authentication, the chatbot processes transaction messages and stores the extracted data in structured form.

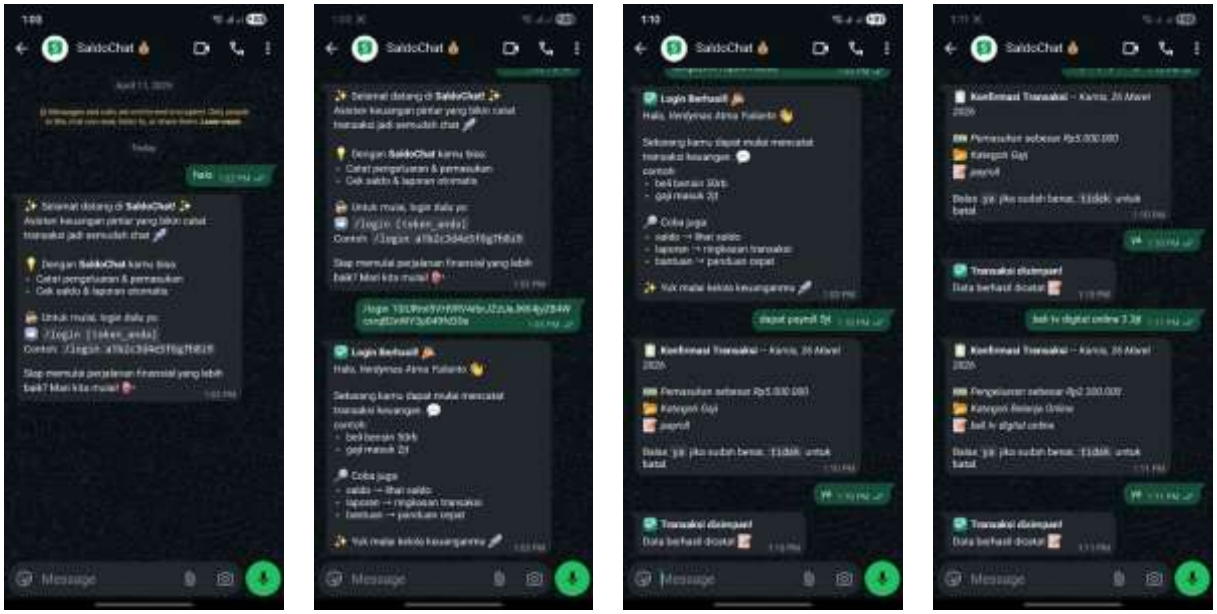

Fig. 4 Example of conversational chatbot interaction

User Acceptance Testing (UAT)

Functional testing was conducted with 30 participants using five predefined scenarios. A scenario is considered successful when the chatbot produces the expected response and stores or retrieves the correct data.

Table 3
UAT Test Scenarios

Scenario	Test Scenario	Expected Result
S1	User login using /login <token>	System authenticates the user and activates chatbot session
S2	Record expense transaction	Chatbot extracts expense information and stores it in Database
S3	Record income transaction	Chatbot extracts income information and stores it correctly
S4	Retrieve transaction history	System returns previously stored transaction records
S5	Request financial summary	System generates a summary of recorded financial transactions

A test case is considered successful when the chatbot correctly interprets the user message, produces the expected response, and the corresponding transaction data is accurately stored or retrieved from the database without requiring additional user correction.

Table 4 UAT Result (n = 30)

Scenario	Success Count	Success Rate
S1	30	100%
S2	29	96.7%
S3	30	100%
S4	28	93.3%
S5	29	96.7%
Average	**29.2**	**97.3%**

The chatbot achieved an average success rate of 97.3% across the tested scenarios. Minor failures occurred mainly when user messages were ambiguous, incomplete, or contained multiple financial details in a single sentence.

System Usability Scale (SUS)

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 2, April 2026
DOI : https://doi.org/10.33395/sinkron.v10i2.16020
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Usability was measured with the System Usability Scale using responses from 30 participants. The SUS questionnaire consists of ten statements that measure different aspects of perceived system usability, including ease of use, system complexity, learnability, and user confidence when interacting with the chatbot.

Table 5 Question System Usability Scale (SUS)

No	Question
Q1	I would use the system frequently
Q2	The system is unnecessarily complex
Q3	The system is easy to use
Q4	I need technical support to use it
Q5	System functions are well integrated
Q6	The system is inconsistent
Q7	Most people would learn it quickly
Q8	The system is cumbersome to use
Q9	I feel confident using the system
Q10	I need to learn many things first

The responses collected from the participants were converted into adjusted SUS scores following the standard scoring procedure.

Table 6 SUS Scoring Data (n = 30)

Respondent	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Amount	Score (×2.5)
R1	4	2	4	2	4	2	4	2	4	2	33	82.5
R2	5	2	4	1	4	2	4	2	4	2	34	85.0
R3	4	2	5	2	4	2	4	2	4	2	33	82.5
…	…	…	…	…	…	…	…	…	…	…	…	…
R30	4	2	4	2	4	2	4	2	4	2	33	82.5
										Total	**990**	**2475**
										Average	**33.0**	**82.5**

To interpret the resulting usability score, the SUS value is compared with the standard SUS rating scale shown in Table 7.

Table 7 SUS Score Range and Interpretation

SUS Score Range	Usability Rating
> 80	Excellent
68 – 80	Good / Acceptable
50 – 67	Marginal
< 50	Poor

The SUS score of 82.5 indicates that the chatbot was well accepted and easy to use for daily financial transaction recording.

Intent Classification Performance

Table 8 Intent Classification Results

Method	Correct Prediction	Accuracy
Rule-based	137	0.457
Llama3.1	133	0.443
Qwen2.5	293	0.977

Qwen2.5 achieved the highest accuracy at 0.977. The rule-based method and Llama 3.1 performed much lower, with accuracies of 0.457 and 0.443, respectively. This indicates that Qwen2.5 is more reliable for intent classification in conversational financial messages.

Entity Extraction Performance (Amount)

Table 9 Amount Extraction Evaluation

Method	TP	FP	FN	Precision	Recall	F1-score
Rule-based	262	38	0	0.873	1	0.932
Llama3.1	157	6	137	0.963	0.534	0.687
Qwen2.5	285	13	2	0.956	0.993	0.974

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 2, April 2026
DOI : https://doi.org/10.33395/sinkron.v10i2.16020
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Qwen2.5 achieved the best amount extraction performance with an F1-score of 0.974. The rule-based method also performed strongly with 0.932, while Llama 3.1 was lower at 0.687. This suggests that Qwen2.5 provides the most balanced extraction of monetary values.

Entity Extraction Performance (Category)

Table 10
Category Extraction Evaluation

Method	TP	FP	FN	Precision	Recall	F1-score
Rule-based	156	27	117	0.852	0.571	0.684
Llama3.1	110	54	136	0.671	0.447	0.537
Qwen2.5	232	66	2	0.779	0.991	0.872

Qwen2.5 produced the strongest category extraction result with an F1-score of 0.872 and very high recall. The rule-based method reached 0.684, while Llama 3.1 obtained 0.537. These results show that Qwen2.5 is better at inferring category information from context.

JSON Validity

Table 11
JSON Validity Evaluation

Method	Valid Outputs	Valid Rate
Llama3.1	199	66%
Qwen2.5	300	100%

The evaluation shows a clear difference in structured output reliability between the evaluated LLMs. Qwen2.5 produced valid JSON in all responses, achieving a 100% validity rate, which indicates strong compliance with the predefined output schema. In contrast, Llama 3.1 generated valid JSON in 66% of the responses, meaning that a portion of the outputs required correction before they could be parsed by the backend system. These results highlight the importance of schema-consistent generation when LLMs are used for structured data extraction.

Error Analysis

A mixed-method evaluation combining quantitative trends and qualitative error analysis was applied to assess the extraction models. The results reveal a clear shift in failure modes from the lexical rigidity of the rule-based system to the generative characteristics of large language models.

Table 12
Error Analysis Rule-Based System Extraction

Text	Prediction	Ground Truth	Error Type
cair bonus target kuartal ini 1.5jt	+ 15000000 Listrik & Air	+ 1500000 Bonus	Substring Overlap and Decimal Parsing
beli kuota telkomsel 30gb 75k	- 30 Pulsa & Internet	- 75000 Pulsa & Internet	Multiple Number Entity Trap
jajan cilok depan kampus 5k	unknown 5000 Pendidikan	- 5000 Makanan & Minuman	Spatial Entity vs. Primary Action

The rule-based approach struggles significantly with implicit information, impairing its intent classification and category extraction. The system is highly vulnerable to substring overlaps (erroneously matching "air" within "cair") and spatial entity traps. Although amount extraction achieved perfect recall, precision suffered noticeably due to multiple number entity traps where the system extracted product specifications instead of actual prices. These failures prove that deterministic systems are fundamentally ill-equipped for unstructured conversational language.

Table 13
Error Analysis LLM Llama 3.1 Extraction

Text	Prediction	Ground Truth	Error Type
gaji pokok bulan april cair 6.500.000	unknown NaN Other	+ 6500000 Gaji	Prompt Adherence Failure
isi nyawa hp paket internet 50rb	- 50000 Harga Paket Internet	- 50000 Pulsa & Internet	Category Hallucination
sewa lapangan futsal 2 jam 150rb	- 150000 Transportasi	- 150000 Hobi	Contextual Hallucination

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 2, April 2026
DOI : https://doi.org/10.33395/sinkron.v10i2.16020
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Llama 3.1 yielded the lowest overall performance in intent recognition and category extraction, primarily due to poor structured output reliability. The model frequently exhibited prompt adherence failures by wrapping the required JSON output within conversational text. This behavior breaks the application programmatic parsing logic and severely degrades recall. Additionally, the model demonstrated category hallucination by inventing out-of-vocabulary categories instead of strictly adhering to predefined lists.

Table 14 Error Analysis LLM Qwen2.5 Extraction

Text	Prediction	Ground Truth	Error Type
beli kuota telkomsel 30gb 75k	- 75 Pulsa & Internet	- 75000 Pulsa & Internet	Slang Suffix Blindness
pemasukan dari jualan pulsa 300rb	+ 300000 Pulsa & Internet	+ 300000 Gaji	Entity Over-Reliance
isi emoney buat naik tol 100rb	- 100000 Pulsa & Internet	- 100000 Transportasi	Contextual Blindness

Qwen2.5 emerged as the superior model, achieving the highest accuracy and a flawless structural validity rate that successfully eliminated previous formatting errors. However, the qualitative analysis in Table 3 reveals minor morphological gaps and contextual misalignments. The model struggled with slang suffix blindness by failing to comprehend localized numerical slang (interpreting "75k" as 75 instead of 75000). It also exhibited entity over-reliance, such as classifying a toll road electronic money top-up as an internet expense by over-focusing on the digital product rather than the explicit transportation intent.

The evaluation demonstrates a clear technological progression and its associated trade-offs. The rule-based method fails at contextual understanding due to rigid syntax. Llama 3.1 grasps general intents but fails significantly at following strict structural constraints. Qwen2.5 masters both structural compliance and contextual reasoning, yet exhibits minor localized morphological gaps. Consequently, developing a highly accurate financial extraction system requires deploying a compliant generative model augmented by localized prompt engineering to address specific cultural nuances.

DISCUSSIONS

The findings indicate that the proposed chatbot is not only usable, but also technically capable of converting conversational financial messages into structured records. The strongest result came from Qwen2.5, which outperformed both the rule-based baseline and Llama 3.1 in intent detection, amount extraction, category extraction, and JSON validity. This pattern suggests that prompt-based reasoning is more effective than deterministic parsing when user messages contain informal wording, abbreviated amounts, or context-dependent expressions. The high JSON validity rate also shows that Qwen2.5 follows the output schema more consistently, which is important for backend parsing and direct database storage.

The rule-based approach produced acceptable results for amount extraction because it is well suited to detecting explicit numeric patterns. Its recall was especially strong when the input contained clear monetary expressions. However, the method performed much worse in intent detection and category extraction because rule-based matching depends on exact lexical cues and predefined mappings. As a result, it is vulnerable to substring overlap, implicit phrasing, and category ambiguity. This confirms that deterministic rules remain useful as a transparent baseline, but they are not sufficient for conversational financial text that is short, informal, and highly variable.

Llama 3.1 showed a different failure pattern. Although it achieved high precision for amount extraction, its recall and JSON validity were notably weaker than Qwen2.5. This indicates that the model often produced correct values when it detected them, but failed to identify many relevant entities and frequently deviated from the required output format. In other words, the model was less reliable in following the structured extraction template. This result is consistent with previous findings that open LLMs may vary in reproducibility and schema compliance when used in controlled extraction settings (Dietrich & Hollstein, 2025). It also shows that strong generative ability alone does not guarantee reliable structured output.

Compared with earlier chatbot and NLP studies, the present results support the view that conversational systems become more effective when they combine semantic understanding with structured output constraints. Prior work on conversational AI and intent detection shows that intent identification remains a bottleneck when data are noisy or domain-specific, while LLM-based approaches can improve coverage and flexibility . In financial NLP, structured extraction from text has also been shown to benefit from explicit output formatting and model-guided prompting, rather than unconstrained generation (Bai et al., 2025; Dagdelen et al., 2024; Yan et al., 2025). The present study extends these findings to conversational personal finance by showing that a prompt-constrained open model can outperform a rule-based baseline in both intent and entity extraction.

*name of corresponding author

e-ISSN : 2541-2019
p-ISSN : 2541-044X

The usability results also reinforce the practical value of the system. The SUS score of 82.5 indicates that users found the chatbot easy to use and suitable for daily financial recording. This aligns with earlier chatbot studies that emphasize iterative design and user-centered evaluation as key factors in acceptance and practical deployment (Sasmita et al., 2025; Sezgin et al., 2024). The UAT success rate of 97.3% further suggests that the system functions reliably in standard interaction scenarios. Together, these results show that technical accuracy and user acceptance can be achieved at the same time when the interaction flow is simple and the output format is constrained.

Despite the positive findings, the study has several limitations. It uses only 300 expense and income samples, so the results may not generalize to more complex financial tasks such as transfer, savings, or mixed transactions. The evaluation relies on a single prompt template and two open-source models, meaning performance could change with different prompts or fine-tuned models. Moreover, the latency analysis isn't optimized for deployment-scale benchmarking, so the runtime comparison is only a functional estimate. Consequently, future work should employ larger, more diverse datasets, cover broader transaction categories, explore hybrid rule-plus-LLM pipelines, and conduct rigorous latency testing. The small dataset size especially limits the generalizability of the findings to wider financial-transaction and conversational scenarios.

Overall, the study shows that conversational financial recording benefits from a hybrid research direction. Rule-based extraction remains useful for highly regular patterns, but Qwen2.5 demonstrates that a well-constrained open LLM can better handle the ambiguity of natural language financial input. The main insight is that success in this task does not depend on generation alone, but on combining contextual understanding with strict schema control. This makes prompt design, structured validation, and domain-limited evaluation central to reliable financial information extraction in chat-based systems.

CONCLUSION

This study contributes to the evaluation of conversational financial transaction extraction by comparing a rule-based baseline with two open large language models, Qwen2.5 and Llama 3.1. The main scientific contribution is the demonstration that structured financial information can be extracted more reliably when prompt-constrained LLMs are used instead of deterministic rules, especially in conversational messages that contain informal wording, abbreviated amounts, and context-dependent expressions. The findings also show that schema control is a critical factor in structured NLP tasks, because extraction quality alone is not sufficient if the generated output cannot be parsed consistently by the backend system.

From a practical perspective, the developed chatbot provides a usable and lightweight interface for personal financial recording through natural language interaction. The system supports token-based login, transaction recording, transaction history retrieval, and financial summaries, making it suitable for everyday use without requiring users to fill out conventional form-based inputs. The UAT and SUS results indicate that the chatbot is functionally reliable and well accepted by users, which suggests that conversational interfaces can reduce the burden of daily financial logging while maintaining usability.

Based on the experimental findings, Qwen2.5 is the most effective model in this study because it combines strong intent classification, accurate entity extraction, and complete JSON validity. This makes it the most suitable approach for the proposed financial extraction workflow among the tested methods. The rule-based approach remains useful as a transparent baseline for regular patterns, while Llama 3.1 shows that generative capability alone does not guarantee stable structured output. These results support the claim that reliable conversational financial systems require both contextual understanding and strict output control.

Future work may extend the dataset to broader transaction types and larger sample sizes. In addition, more complex financial interactions such as transfers, savings tracking, subscription payments, and multi-intent conversational transactions can be explored to further evaluate the robustness of conversational financial extraction systems. Further research may also examine deployment-scale latency and robustness in real-world usage scenarios to strengthen the generalizability of the proposed approach.

REFERENCES

Alda, M. (2023). Pengembangan Aplikasi Pengolahan Data Siswa Berbasis Android Menggunakan Metode Prototyping. *Jurnal Manajemen Informatika (JAMIKA)*, *13*(1), 11–23. https://doi.org/10.34010/jamika.v13i1.8216

Alfath, M. F., Fanani, L., & Kharisma, A. P. (2024). Pengembangan Aplikasi Berlatih Membaca Cepat Berbahasa Inggris Berbasis Progressive Web App dengan Metode Prototyping. *Jurnal Teknologi Informasi Dan Ilmu Komputer*, *11*(5), 1001–1008. https://doi.org/10.25126/jtiik.2024117982

Annisa, Z. A., Perdana, R. S., & Adikara, P. P. (2024). Kombinasi Intent Classification dan Named Entity Recognition pada Data Berbahasa Indonesia dengan Metode Dual Intent and Entity Transformer.

*name of corresponding author

Jurnal Teknologi Informasi Dan Ilmu Komputer, *11*(5), 1017–1024. https://doi.org/10.25126/jtiik.2024117985

Bai, Y., Gong, J., Han, M., & Yang, J. (2025). The Financial Institution Text Data Mining and Value Analysis Model Based on Big Data and Natural Language Processing. *Journal of Organizational and End User Computing*, *37*(1). https://doi.org/10.4018/JOEUC.374213

Batura, T., Yerimbetova, A., Mukazhanov, N., Shvarts, N., Sakenov, B., & Turdalyuly, M. (2025). Information Extraction from Multi-Domain Scientific Documents: Methods and Insights. *Applied Sciences (Switzerland)*, *15*(16). https://doi.org/10.3390/app15169086

Benayas, A., Miguel-Ángel, S., & Mora-Cantallops, M. (2024). Enhancing Intent Classifier Training with Large Language Model-generated Data. *Applied Artificial Intelligence*, *38*(1). https://doi.org/10.1080/08839514.2024.2414483

Chandrakala, C. B., Bhardwaj, R., & Pujari, C. (2024). An intent recognition pipeline for conversational AI. *International Journal of Information Technology (Singapore)*, *16*(2), 731–743. https://doi.org/10.1007/s41870-023-01642-8

Chen, Z., Ma, D., Li, H., Chen, L., Ji, J., Liu, Y., Chen, B., Wu, M., Zhu, S., Dong, X., Ge, F., Miao, Q., Lou, J. G., Fan, S., & Yu, K. (2025). DFM: Dialogue foundation model for universal large-scale dialogue-oriented task learning. *AI Open*, *6*, 108–117. https://doi.org/10.1016/j.aiopen.2025.04.001

Dagdelen, J., Dunn, A., Lee, S., Walker, N., Rosen, A. S., Ceder, G., Persson, K. A., & Jain, A. (2024). Structured information extraction from scientific text with large language models. *Nature Communications*, *15*(1). https://doi.org/10.1038/s41467-024-45563-x

Dave, E., & Chowanda, A. (2024). IPerFEX-2023: Indonesian personal financial entity extraction using indoBERT-BiGRU-CRF model. *Journal of Big Data*, *11*(1). https://doi.org/10.1186/s40537-024-00987-6

Dietrich, J., & Hollstein, A. (2025). Performance and Reproducibility of Large Language Models in Named Entity Recognition: Considerations for the Use in Controlled Environments. *Drug Safety*, *48*(3), 287–303. https://doi.org/10.1007/s40264-024-01499-1

Dwi Astuti, M., & Soleha, E. (2023). Pengaruh Literasi Keuangan, Inklusi Keuangan Dan Locus of Control Terhadap Pengelolaan Keuangan UMKM di Kecamatan Bojongmangu. *JURNAL EKONOMI PENDIDIKAN DAN KEWIRAUSAHAAN*, *11*(1), 51–64. https://doi.org/10.26740/jepk.v11n1.p51-64

Eriana, E. S., & Subariah, R. (2025). Implementation of Natural Language Processing Based Chatbot as a Virtual Assistant in Science Learning. *Jurnal Penelitian Pendidikan IPA*, *11*(10), 633–640. https://doi.org/10.29303/jppipa.v11i10.12747

Ferrera, A., Mezzotero, G., & Ursino, D. (2025). A linguistics-based approach to refining automatic intent detection in conversational agent design. *Information Sciences*, *689*. https://doi.org/10.1016/j.ins.2024.121493

Mahastanti, L., & Utoyo, D. R. R. (2022). Pengaruh Payment Gateway (GO-PAY) Terhadap Kinerja Finansial UMKM di Kota Salatiga. *JURNAL EKONOMI PENDIDIKAN DAN KEWIRAUSAHAAN*, *10*(2), 105–116. https://doi.org/10.26740/jepk.v10n2.p105-116

Otoritas Jasa Keuangan, & Badan Pusat Statistik. (2025). *Survei Nasional Literasi dan Inklusi Keuangan (SNLIK) 2025*. Otoritas Jasa Keuangan (OJK) dan Badan Pusat Statistik (BPS).

Ouaddi, C., Benaddi, L., Bouziane, E. mahi, Naimi, L., Rahouti, M., Jakimi, A., & Saadane, R. (2025). Assessing the effectiveness of large language models for intent detection in tourism chatbots: A comparative analysis and performance evaluation. *Scientific African*, *28*. https://doi.org/10.1016/j.sciaf.2025.e02649

Pressman, R. S. (2010). *Software Engineering: A Practitioner's Approach* (7th ed.). McGraw-Hill Higher Education.

Puspitasari, A., Paradhita, A. N., Tineka, Y. W., Sulistyowati, V., Noriska, N. K. S., & Haryanto. (2024). Natural Language Processing (NLP) Technology for Chatbot Website. *Jurnal Penelitian Pendidikan IPA*, *10*(SpecialIssue), 319–324. https://doi.org/10.29303/jppipa.v10ispecialissue.8241

Putri Oktavianita, R., & Andreas Sutanto, F. (2024). Rekomendasi Pemilihan Hotel Berbasis Chatbot dengan Framework Rasa Dengan Metode Natural Language Processing (NLP). *Jurnal Riset Sistem Informasi Dan Teknik Informatika (JURASIK)*, *9*(2), 634–641. https://doi.org/10.30645/jurasik.v9i2.795.g770

Rachmawati, F. F., Sudarno, S., & Sabandi, M. (2023). Pengaruh Literasi Keuangan dan Lingkungan Sosial Dimoderasi Tingkat Pendidikan Terhadap Penggunaan QRIS Pada Pelaku UMKM di Kota Surakarta. *JURNAL EKONOMI PENDIDIKAN DAN KEWIRAUSAHAAN*, *11*(1), 21–36. https://doi.org/10.26740/jepk.v11n1.p21-36

*name of corresponding author

Ricaldi, L. C., Martin, T. K., & Huston, S. J. (2022). Financial literacy and its impact on the credit card debt puzzle. *Financial Services Review*, *30*(2), 107–124. https://doi.org/10.61190/fsr.v30i2.3477

Sasmita, W. M. H., Sumpeno, S., & Rachmadi, R. F. (2025). Improving Government Helpdesk Service With an AI-Powered Chatbot Built on the Rasa Framework. *Jurnal RESTI*, *9*(2), 393–403. https://doi.org/10.29207/resti.v9i2.6293

Sezgin, E., Kocaballi, A. B., Dolce, M., Skeens, M., Militello, L., Huang, Y., Stevens, J., & Kemper, A. R. (2024). Chatbot for Social Need Screening and Resource Sharing With Vulnerable Families: Iterative Design and Evaluation Study. *JMIR Human Factors*, *11*. https://doi.org/10.2196/57114

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing and Management*, *45*(4), 427–437. https://doi.org/10.1016/j.ipm.2009.03.002

Tandrio, F., & Fianty, M. I. (2026). Web-Based Payroll System Development Using The Prototyping Method and Structured Database Design. *JITK (Jurnal Ilmu Pengetahuan Dan Teknologi Komputer)*, *11*(3), 851–863. https://doi.org/10.33480/jitk.v11i3.7044

Widiyanti, N. F., Sukmana, H. T., Hulliyah, K., Khairani, D., & Oh, L. K. (2023). Improving Indonesian Named Entity Recognition for Domain Zakat Using Conditional Random Fields. *Jurnal Online Informatika*, *8*(2), 131–138. https://doi.org/10.15575/join.v8i2.898

Wildannissa Pinasti, & Lya Hulliyyatus Suadaa. (2024). Named Entity Recognition in Statistical Dataset Search Queries. *Jurnal Nasional Teknik Elektro Dan Teknologi Informasi*, *13*(3), 171–177. https://doi.org/10.22146/jnteti.v13i3.11580

Wobst, J., Röttger, P., & Lueg, R. (2025). Measuring value-based management using natural language processing. *Management Accounting Research*, *67*. https://doi.org/10.1016/j.mar.2025.100946

Yan, Z., Ye, Z., Ge, J., Qin, J., Liu, J., Cheng, Y., & Gurrin, C. (2025). DocExtractNet: A novel framework for enhanced information extraction from business documents. *Information Processing and Management*, *62*(3). https://doi.org/10.1016/j.ipm.2024.104046

Zahra, A., Hidayatullah, A. F., & Rani, S. (2022). Bidirectional long-short term memory and conditional random field for tourism named entity recognition. *IAES International Journal of Artificial Intelligence*, *11*(4), 1270–1277. https://doi.org/10.11591/ijai.v11.i4.pp1270-1277

*name of corresponding author