Scientific African 28 (2025) e02649

Contents lists available at ScienceDirect

Scientific African

journal homepage: www.elsevier.com/locate/sciaf

Assessing the effectiveness of large language models for intent detection in tourism chatbots: A comparative analysis and performance evaluation

Charaf Ouaddi [a,*], Lamya Benaddi [a], El mahi Bouziane [a], Lahbib Naimi [a], Mohamed Rahouti [b], Abdeslam Jakimi [a], Rachid Saadane [c]

[a] *Software Engineering and Information Systems Engineering Team, Department of Computer Sciences, Faculty of Sciences and Techniques Errachidia, Moulay Ismail University, Morocco*
[b] *CIS Dept. Fordham University New York, NY 10023 USA*
[c] *Electrical Engineering Department, Hassania School of Public Works, Casablanca, Morocco*

ARTICLE INFO

Editor DR B Gyampoh

Keywords:
Intent Detection
Tourism
Chatbot
Deep learning
Generative AI
LLM

ABSTRACT

In recent years, the tourism industry has observed a significant transformation by integrating chatbots, which enable tourists to interact with various services using natural language. At the heart of each chatbot is a Natural Language Understanding (NLU) component, which processes natural language inputs through intent classification. This paper evaluates the performance of Large Language Models (LLMs) such as GPT, BERT, LLaMA, and RoBERTa in the intent classification task for tourism chatbots. Our study conducts a comparative analysis of various LLMs to determine their effectiveness in classifying user intents in tourism related interactions. We assess the models' capabilities using a tourism-specific dataset labeled according to the "Six A" criteria for tourist destination analysis. The models are evaluated using performance metrics such as accuracy, precision, recall, and F1-score. The findings provide practical insights into developing efficient NLU components for tourism chatbots, enhancing their ability to understand and assist users effectively. This paper contributes to the field by offering a comprehensive performance evaluation of LLMs for NLU in tourism, guiding researchers and practitioners in building more responsive and accurate chatbots for the tourism industry.

Introduction

In years, the tourism industry has observed a significant transformation by integrating chatbots, which enable tourists to interact with various services using natural language [1]. At the heart of each chatbot lies the Natural Language Understanding (NLU) component, which processes natural language inputs by performing tasks such as intent detection. However, the challenge lies in selecting the optimal Large Language Model (LLM) for building an effective NLU system for tourism chatbots.

This paper aims to address this challenge by evaluating the performance of several LLMs in classifying intents. The LLMs considered include GPT, BERT, LLaMA, and RoBERTa. The motivation behind this study stems from the increasing demand for more responsive, accurate, and intelligent tourism chatbots capable of understanding user intents and providing relevant and contextual responses [2,

[*] Corresponding author.
E-mail address: c.ouaddi@edu.umi.ac.ma (C. Ouaddi).

https://doi.org/10.1016/j.sciaf.2025.e02649
Received 18 September 2024; Received in revised form 7 February 2025; Accepted 14 March 2025
Available online 15 March 2025

3],. As tourists increasingly rely on digital tools for planning and navigating their journeys, the ability of chatbots to comprehend and interpret diverse and often complex natural language queries becomes crucial [6]. While there are several LLMs available that excel in various natural language processing tasks, their effectiveness in intent detection for domain-specific applications like tourism is not well understood. This necessitates a comprehensive evaluation to identify which models are most suitable for intent detection tasks, ensuring that tourism chatbots can deliver a high-quality user experience.

Current state-of-the-art LLMs such as GPT, BERT, LLaMA, and RoBERTa have demonstrated impressive performance in general-purpose natural language understanding tasks [4,5],; however, there are notable gaps when these models are applied to domain-specific tasks such as intent detection in tourism [6]. Existing research often focuses on general benchmarks like SQuAD or GLUE, which may not fully capture the intricacies involved in tourism-related interactions. Furthermore, the need for models that balance accuracy with computational efficiency remains a significant challenge, particularly in real-world applications where resources are limited [7,8],. This paper aims to bridge these gaps by providing a comparative analysis of these models within the context of tourism chatbots, identifying their strengths and weaknesses, and recommending the most effective LLMs for practical implementation.

Our study conducts a comprehensive comparative analysis using a tourism-specific dataset labeled according to the “Six A” criteria [9]: Attractions, Activities, Accessibility, Available packages, Amenities, and Ancillary Services. We employ accuracy, precision, recall, and F1-score metrics to evaluate the models’ performance. The key contributions of this work are outlined as follows:

- Curate a diverse dataset specific to the tourism sector, ensuring robust training, validation, and testing processes: We utilize a tourism-specific dataset labeled according to the “Six A” criteria. This dataset provides a focused and relevant context for assessing the performance of the algorithms and LLMs.
- Assessment of LLMs: Our study extends the evaluation to include LLMs such as GPT, BERT, LLaMA, and RoBERTa, analyzing their effectiveness in classifying user intents within the context of tourism services.
- Recommendation of the most effective LLMs and insights for practical implementation: We identify the most effective LLM for intent detection in tourism chatbots based on the comparative analysis.

The rest of this paper is organized as follows. Section 2 provides a comprehensive background on intent detection in chatbots and reviews related work, discussing linguistic rule-based, machine learning, and deep learning approaches. Section 3 presents the adopted methodology, describing the LLMs evaluated, the tourism-specific dataset used, and the evaluation metrics employed. Section 4 discusses the experimental results and provides a detailed performance evaluation of the BERT, GPT-2, RoBERTa, and LLaMA models on the intent classification task. Section 5 provides a discussion of the results, including interpretations, time complexity considerations, and practical applications and limitations of the models in tourism chatbots. Finally, Section 6 concludes the study by summarizing the main findings and suggesting directions for future research.

Background and state-of-the-art

Intent detection in chatbots

Chatbots are intelligent computer programs designed to simulate human conversation and interact with users [10]. They can be broadly categorized into rule-based and AI-based chatbots [2,3],. Rule-based chatbots operate on predefined rules and scripts, limiting them to specific responses. In contrast, AI-based chatbots use advanced algorithms and machine learning to understand and respond to a broader range of queries. The core of AI-based chatbots is the NLU component [11], which classifies user intents to generate appropriate responses. This classification task is crucial, as it determines how accurately the chatbot understands and processes user input, making it essential for effective communication and user satisfaction (Fig. 1).

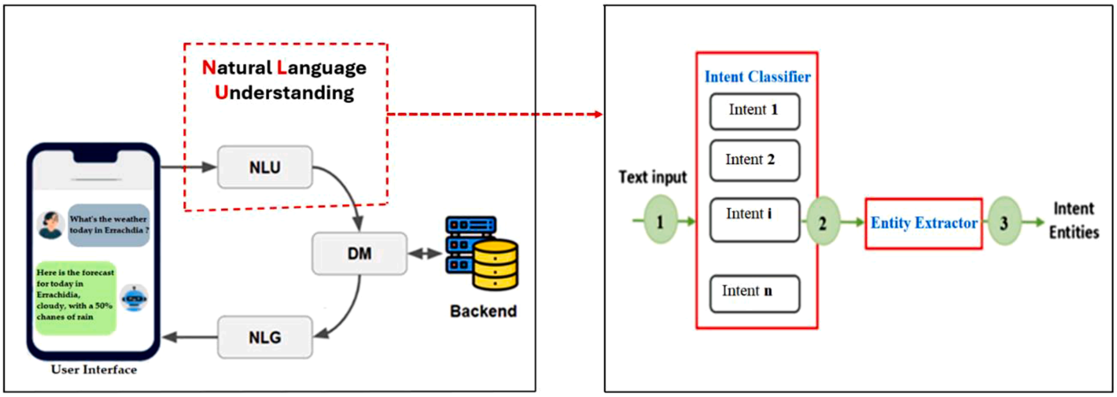
Natural Language Understanding
NLU
DM
NLG
Backend
What's the weather today in Errachdia ?
Here is the forecast for today in Errachidia, cloudy, with a 50% chanes of rain
User Interface
Intent Classifier
Intent 1
Intent 2
Intent i
Intent n
Text input
1
2
3
Entity Extractor
Intent
Entities

Fig. 1. Overview of NLU component in Chatbot Systems (adopted from [12] and [13]).

Overview of LLMs

LLMs are advanced AI models designed to comprehend and generate human language. These models, predominantly based on the Transformer architecture [4], are trained on extensive text corpora, which enables them to execute a wide array of natural language processing (NLP) tasks such as text generation, translation, summarization, and question answering. LLMs are pivotal in the progression of AI and NLP as they empower machines to understand and produce human language with remarkable precision. Their diverse applications range from chatbots and virtual assistants to automated content creation and language translation. Furthermore, the capability to fine-tune these models for specific tasks enhances their versatility, rendering them indispensable tools in various industry applications [7,8],.

Related work

In the literature, there are three main approaches to realizing an NLU component: the first is the linguistic rule-based approach [14, 15], the second is the machine learning approach [14,16], and the third is the deep learning approach.

Linguistic rule-based approach

Historically, rule-based systems have been employed for NLU. This method depends on expertise in a specific field to recognize important terms or words in the user's statement, which helps determine their intention [17,18],.

In addition, this approach also relies on predefined grammatical and syntactical rules to parse and understand NL. It was widely used in the early stages of NLU development due to its simplicity and the availability of linguistic expertise. Systems like ELIZA, one of the first chatbots, used rule-based methods to simulate conversation by pattern matching [19]. Another example is the LFG (Lexical Functional Grammar) Parser, which employs deep linguistic analysis to interpret user inputs [20]. Conditional Random Fields (CRFs) are commonly employed for the task of entity extraction or name-entity recognition (NER) and are widely utilized in many applications [21]. The approach is limited in adaptability and less scalable due to the requirement of a laborious feature extraction process.

However, despite their early success, rule-based systems have inherent limitations [22]. They are often brittle, struggling to adapt to the variability of NL as it is used in different contexts. For instance, as new linguistic expressions and colloquial phrases emerge, rule-based systems require extensive manual updates to remain effective. This lack of flexibility has been noted as a significant drawback in various studies [23], which ultimately led to the exploration of more adaptive approaches (discussed in the next sections 2.3.2 and 2.3.3), such as ML and deep learning, for both intent classification and entity extraction tasks.

Machine learning approaches for intent classification and named entity extraction in NLU

This approach, based on machining learning techniques, defines NLU through two tasks: intent classification and named entity extraction [14,16],. Intent classification is typically treated as a classification task, where the model identifies the intent from a predefined list based on a given user message. Named entity extraction involves identifying key elements within a sentence to utilize them for subsequent actions. Both tasks involve training models on large datasets to learn patterns and make predictions. This approach improves the system's ability to generalize from examples rather than relying on manually crafted rules.

Most intent classification techniques have used supervised machine learning algorithms, including SVM [24] and decision trees [25]. Furthermore, Naive Bayes and KNN are among the prevalent algorithms employed in this methodology. For example, [26] employs a Support Vector Machine (SVM) approach for intent categorization, attaining a 78.9% success rate. This method demonstrates substantial enhancements compared to rule-based methods. Furthermore, Pang et al. [27] have successfully utilized Naive Bayes in sentiment analysis applications.

Deep learning techniques for NLU

The deep learning approach utilizes artificial neural networks (ANN) to represent complex patterns in language data. Methods such as Recurrent Neural Networks (RNNs) [28], Long Short-Term Memory (LSTM) networks [29], Gated Recurrent Unit (GRU) [23], bidirectional LSTM with an attention mechanism [30], and LLMs [6] based on the Transformers have significantly transformed NLU by allowing the efficient analysis of extensive amounts of unorganized textual data with exceptional precision.

The Transformer model developed by Vaswani et al. [4], which incorporates the attention mechanism, is the foundation for numerous cutting-edge NLU systems, such as BERT (Bidirectional Encoder Representations from Transformers). The BERT model, created by Devlin et al. [5], significantly improves NLU by training on extensive text collections and then fine-tuning for particular tasks. This approach leads to exceptional performance in many NLU evaluations.

Methodology

To achieve the primary objective of this paper, which is to evaluate the performance of various LLMs in the tourism sector, we need to carefully select the candidate LLMs for examination and the dataset for training and testing these models. This section provides a detailed account of our selection process for the different LLMs, the tourism functionalities used in the evaluation, and the design of our experiments.

Experimental methodology

This study evaluates the performance of diverse ML techniques for intent classification (IC) in tourism chatbots. The methodology comprises steps 1, 2, 3, 4, and 5 depicted in Fig. 2.

Overview of evaluated LLMs

Several widely used LLMs can be easily fine-tuned for specific tasks. To make our study comprehensive, we examined the performance of five LLMs: GPT-2, LLaMA 3.1, BERT, Flacon, and RoBERTa. These LLMs were selected because they are popular and extensively used by researchers. In the following sections, we describe these LLMs.

GPT: The GPT (Generative Pre-trained Transformer) model follows the evolutionary path of language models. Like its predecessors, GPT is trained unsupervised on many documents to produce a general language model upon which more specific natural language processing tasks can be conducted. The encoding architecture also adopts the “transformers” solution introduced by Vaswani et al. [4], featuring a multi-layer structure of attention modules. The model has been developed in six successive versions: GPT-1 (Radford et al., 2018 in [31]), GPT-2 (Radford et al., 2019 in [32]), and GPT-3 (Brown et al., 2020 in [33]), each progressively improving the model’s performance by increasing both the size of the training datasets and, more importantly, the model’s complexity (117 million parameters for GPT-1, 1.5 billion for GPT-2, 175 billion for GPT-3, GPT-3.5, GPT-4, and GPT-4o).

Like other language models, GPT has been made available in a library allowing word encoding from a text.[1]

BERT: The BERT model (Bidirectional Encoder Representations from Transformers) offers a pre-trained network layer (itself composed of several sub-layers) that dynamically produces word embeddings adapted to the context of the document being analyzed. The model exhibits several similarities with embeddings. First, it translates a set of linguistic information into a vector form, allowing the word to be integrated into the processing chain of neural networks. Second, it is trained on huge document corpora (800 million words and 2.5 billion words [5]), ensuring good representativeness of the contexts in which words are used. Third, the model comes in multiple versions, varying the number of parameters used. As with embeddings, the larger the dimensions, the more accurate the model and the more demanding it is in terms of computational time [5]. Fourth, the trained model is made available in libraries that can be directly used as a starting point for solving natural language processing tasks.[2]

Additionally, the model adopts the encoder-decoder architecture used in unsupervised approaches or sequence generation tasks (for example, translation tasks where an input sequence must produce an output sequence). Encoder-decoder networks operate in two stages: the encoder part generates an encoding of the input sequence in the form of a vector representation, which is then used by the decoder part to generate the output sequence. Both parts are composed of neural networks consisting of recursive, recurrent, or convolutional modules for processing sequential data, with auxiliary modules, such as attention modules, attached for data synchronization. An attention module can also be used at the junction between the encoder and decoder parts.

The BERT model is unique because it relies on an encoder-decoder architecture composed exclusively of attention modules, known as the “Transformer” [4]. It is trained unsupervised on two natural language processing tasks: predicting a word in a sequence and detecting the continuity of two sequences [5]. The decoding module is specifically adapted to solve a particular task and is only helpful for the BERT model within the context of its training. Therefore, only the encoder part is utilized in the context of transfer learning and is made available in public libraries. BERT serves as the encoding layer of a network, positioned upstream of a decoding layer tailored to the task at hand, which can vary from one study to another.

RoBERTa: RoBERTa, aka Robustly Optimized BERT Approach, is an advanced variant of the BERT model, developed by Facebook AI, designed for NLP tasks. It enhances BERT by training with more data, larger mini-batches, and removing the next-sentence prediction objective, improving performance on various NLP benchmarks.

LLaMA: LLaMA is an advanced language model developed by Meta AI, designed to push the boundaries of NLP. It offers capabilities that can be fine-tuned for a wide range of tasks, from text generation and summarization to sentiment analysis, translation, and others. In our study, we will focus on fine-tuning LLaMA 3.1, the latest iteration of this model, which has been further optimized for performance and efficiency. LLaMA 3.1 boasts improved architecture, making it more effective in understanding and generating human-like text. This version of LLaMA is particularly suited for tasks that require a deep understanding of context.

Algorithm 1 outlines a comprehensive procedure for evaluating the performance of various LLMs in intent detection for tourism chatbots. The process begins by taking a labeled dataset, a set of LLMs, and a range of learning rates as input. The dataset is split into training, validation, and test sets for each LLM, and the model is initialized with each specified learning rate. The model is then trained on the training set and evaluated on the validation and test sets to compute performance metrics, including Precision, Recall, and F1-score. These metrics are stored for each model and learning rate combination to determine the optimal configuration for intent detection tasks.

Description of dataset

To prepare the dataset used in intent classification, we adopted an approach based on the 6A or the “Six A” Framework for tourist destination analysis [9]: attractions, accessibility, amenities, activities, available packages, and ancillary services (Fig. 3). The key

[1] https://huggingface.co/docs/transformers/index

[2] https://huggingface.co/docs/transformers/model doc/bert

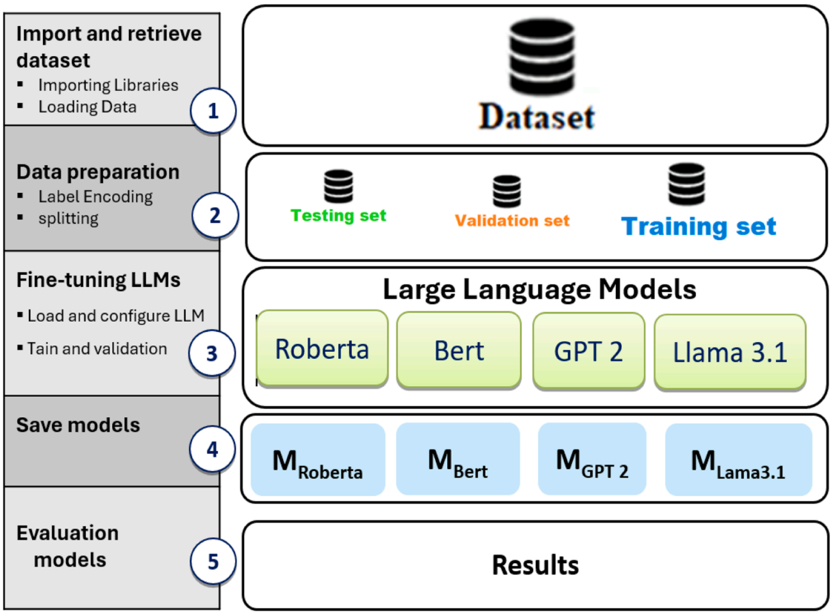

Fig. 2. An overview of the methodology.

Algorithm 1
The proposed LLMs model for intent detection in tourism chatbots procedure.

```
Input: Dataset $D$ with labeled intents, Learning rates $LR = \{lr_1, lr_2, ..., lr_k\}$, LLMs $M = \{m_1, m_2, ..., m_l\}$.
Output: Performance metrics: Precision, Recall, F1-score for each LLM.
procedure TRAIN AND EVALUATE LLMS
    for each model $m \in M$ do
        for each learning rate $lr \in LR$ do
            // Split dataset $D$ into training set $D_{train}$, validation set $D_{val}$, and test set $D_{test}$
            $D_{train}, D_{val}, D_{test} \leftarrow \text{Split}(D)$
            // Initialize model $m$ with learning rate $lr$
            $m_{lr} \leftarrow \text{Initialize}(m, lr)$
            // Train model on $D_{train}$
            $m_{lr} \leftarrow \text{Train}(m_{lr}, D_{train})$
            // Evaluate model on $D_{val}$ and $D_{test}$
            $\text{metrics}_{val} \leftarrow \text{Evaluate}(m_{lr}, D_{val})$
            $\text{metrics}_{test} \leftarrow \text{Evaluate}(m_{lr}, D_{test})$
            // Store performance metrics
            $\text{Results}[m][lr] \leftarrow (\text{metrics}_{val}, \text{metrics}_{test})$
        end for
    end for
end procedure
procedure COMPUTE METRICS
    for each model $m \in M$ do
        for each learning rate $lr \in LR$ do
            Precision $= \frac{TP}{TP+FP}$
            Recall $= \frac{TP}{TP+FN}$
            F1-score $= 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision}+\text{Recall}}$
        end for
    end for
end procedure
return Performance metrics for each model and learning rate combination.
```

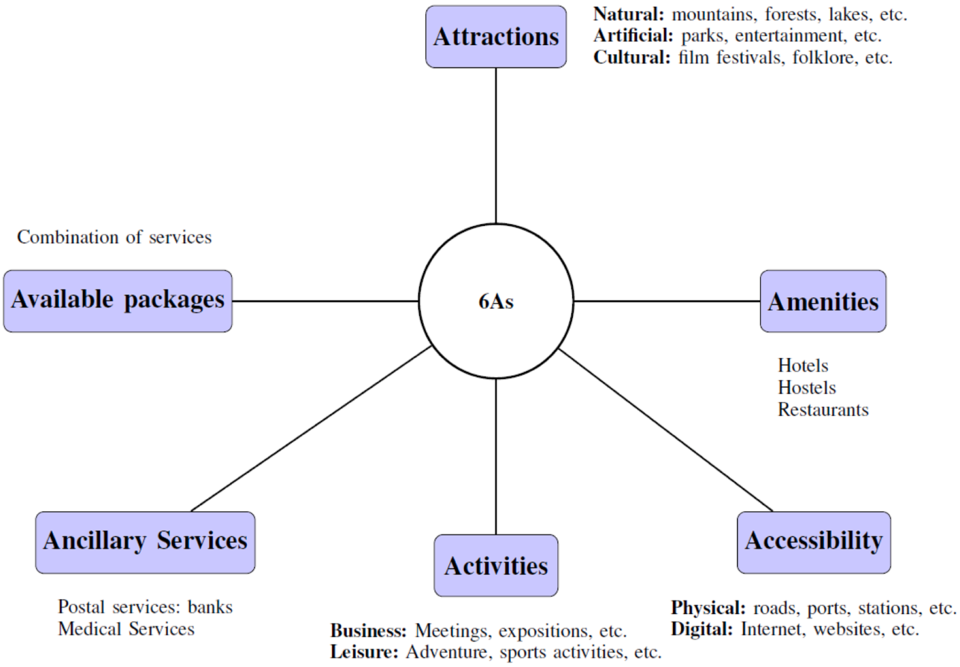
Attractions
Natural: mountains, forests, lakes, etc.
Artificial: parks, entertainment, etc.
Cultural: film festivals, folklore, etc.
Combination of services
Available packages
6As
Amenities
Hotels
Hostels
Restaurants
Ancillary Services
Postal services: banks
Medical Services
Activities
Business: Meetings, expositions, etc.
Leisure: Adventure, sports activities, etc.
Accessibility
Physical: roads, ports, stations, etc.
Digital: Internet, websites, etc.

Fig. 3. The “Six A” Framework for tourist destination analysis (adopted from [1,34],).

features of the datasets are summarized in Table 1 and detailed as follows.

- **Attractions**: Information on tourist attractions, historical sites, and natural wonders was meticulously gathered through extensive research from tourism websites, travel guides, and official destination sources. This effort resulted in 1200 training examples comprising 12 intents: GetArtificialAttractions, GetNaturalAttractions, MountainAttractions, ForestAttractions, CaveAttractions, ReligiousAttractions, ArchitecturalAttractions, Museum, Monument, BeachAttractions, Waterfalls, and HillAttractions to cover a diverse range of attractions.
- **Amenities**: We identified available amenities and services for tourists, such as accommodations, dining options, and hotels. Five hundred training examples, distributed across five intents: GetHotels, GetCafe, GetRestaurant, GetCamping, and PublicTransportationOptions, were collected to provide comprehensive information about amenities in the Draa-Tafilalet region.
- **Accessibility**: Information on transportation options, routes, and accessibility features was compiled to facilitate tourists’ travel plans. This included details about airports, train stations, bus services, and car rentals. We prepared 700 training examples with seven intents: PublicTransportationOptions, AskForTaxiTransportation, BusTransportation, AirplaneTransportation, MiniTaxiTransportation, GetInformationAccessibilityFeatures, and FaresTransportation to address various accessibility-related inquiries.
- **Activities**: A list of leisure activities available to tourists, including adventures, cultural experiences, and entertainment options, was assembled. This resulted in 400 training examples, categorized into four intents: GetLocationActivities, GetLocationEvents, GetSportActivities, and AskAdventureSports, covering various activities and experiences offered in the destinations.
- **Ancillary services**: Ancillary services that complement tourists’ experiences, such as tour guides, translators, travel insurance, and local assistance services, were identified. We prepared 300 training examples, distributed over three intents: GetLocationServices, GetTravelAgencies, and AskBankLocation to address inquiries related to ancillary services and support tourists’ needs during their travels.

Table 1
Summary of dataset features for Draa-Tafilalet tourism.

Feature	Description	Training examples
Attractions	Tourist attractions, historical landmarks, and natural wonders.	1200
Amenities	Amenities for tourists, including accommodations, dining options, and hotels	500
Accessibility	Transportation options, routes, and accessibility features	700
Activities	Leisure activities for tourists, including adventures, cultural experiences, and entertainment options	400
Ancillary services	Tour guides, translators, travel insurance, and local assistance services	300
Available packages	Travel deals, itineraries, pricing, inclusions, bookings	400
Total		3500

- **Available packages**: We collected detailed information about travel deals, including itineraries, pricing, inclusions, and booking details. Four hundred training examples, each intent comprising 100 utterances, were collected to provide insights into the available packages and assist users in making informed decisions. The four intents covered are: GetAvailablePackages, FindSeasonalPackages, FindSportsPackages, and FindFamilyFriendlyPackages.

Overall, the dataset comprises 3500 utterances, including 35 intents, ensuring a comprehensive and diverse representation of the tourism-related information for the Draa-Tafilalet region. The dataset is available at this link: https://github.com/charaf83/Dataset.

Evaluation metrics

The effectiveness of the models is assessed using several evaluation metrics derived from the test dataset, specifically accuracy, precision, recall, and F1-score [35]. Accuracy, which quantifies the ratio of correctly identified instances, is calculated as (TP+TN)/ (TP+TN+FP+FN), where TP (true positives), TN (true negatives), FP (false positives), and FN (false negatives) denote the respective counts of each outcome. Precision, given by TP/ (TP+FP), indicates the proportion of true positive predictions among all positive predictions, while recall, defined as TP/(TP+FN), reflects the model's capability to detect all actual positive instances. The F1-score, formulated as (2 × Precision × Recall)/ (Precision+Recall), serves as the harmonic mean of precision and recall, providing a balanced measure of both metrics.

Although we evaluate each model's accuracy, precision, and recall, our focus in this paper is on presenting the weighted F1-score, as it provides a balanced measure of model performance across different classes without detailing each metric for individual intents.

Experimental results

In this section, a comparative analysis of the performance of various LLMs is presented. It is intended to classify one task intent from training examples. A dataset focused on tourism intents specific to the Draa-Tafilalet region in Morocco is used to implement this task.

Performance evaluation of BERT

Table 2 above presents the BERT model's precision, recall, and F1-score metrics across different learning rates, evaluated on the training, validation, and test sets. The comparison between the two learning rates (2e-5 and 5e-6) allows us to assess the impact of learning rate adjustments on the model's performance. As demonstrated, the model maintains high precision, recall, and F1-score across all sets, with slight validation and test performance variations. These results highlight the effectiveness of fine-tuning BERT for the intent classification task within the tourism domain, explicitly focusing on intents related to the Draa-Tafilalet region in Morocco.

Fig. 4 illustrates the training and validation accuracy of the BERT model for the learning rate 5e-6 during the intent classification task.

Performance evaluation of GPT 2

Table 3 presents the results of the performance evaluation of the GPT-2 model in the intent classification task across different learning rates. The metrics reported include precision, recall, and F1-score for the training, validation, and test datasets. Additionally, Fig. 5 illustrates the training and validation accuracy for the GPT-2 model with learning rates of 1e-4 and 1e-5 over 30 epochs. These results provide a comprehensive view of the model's performance and ability to generalize across different data splits.

Performance evaluation of RoBERTa

Table 4 provides a detailed summary of the performance evaluation of the RoBERTa model in the intent classification task, measured across different learning rates. The metrics reported include precision, recall, and F1-score for the training, validation, and test datasets. Additionally, Fig. 6 presents the training and validation accuracy of the RoBERTa model when fine-tuned with a learning rate of 5e-6, illustrating the model's learning progression over 30 epochs. These results are essential for understanding how well the RoBERTa model generalizes and performs under different training conditions in the context of intent classification.

Table 2
Results of performance evaluation of BERT model in intent classification Task.

Learning rate	BERT								
	Precision			Recall			F1-score		
	Training	Validation	Testing	Training	Validation	Testing	Training	Validation	Testing
2e-5	1.00	0.97	0.99	1.00	0.95	0.98	1.00	0.95	0.98
4e-5	1.00	0.95	0.99	1.00	0.94	0.99	1.00	0.94	0.99
3e-6	1.00	0.96	0.99	1.00	0.94	0.98	1.00	0.94	0.98
5e-6	**1.00**	**0.97**	**0.99**	**1.00**	**0.95**	**0.98**	**1.00**	**0.95**	**0.98**

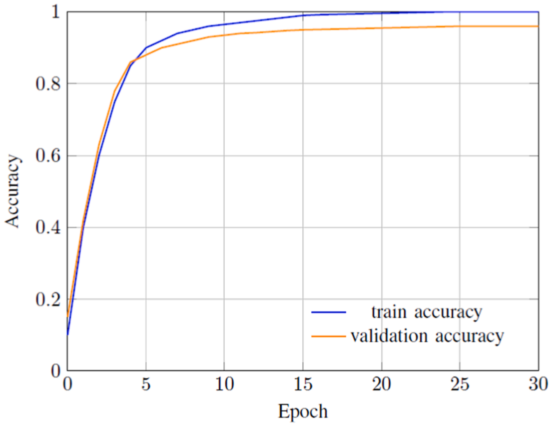

Fig. 4. Training and validation accuracy for BERT model with LR 5e-6.

Table 3
Results of performance evaluation of GPT 2 model in intent classification Task.

Learning rate	**GPT 2**								
	Precision			**Recall**			**F1-score**		
	Training	Validation	Testing	Training	validation	Testing	Training	validation	Testing
1e-2	0.73	0.74	0.75	0.73	0.69	0.71	0.73	0.68	0.71
1e-3	0.82	0.84	0.77	0.82	0.78	0.75	0.82	0.79	0.75
1e-4	**1.00**	**0.97**	**0.99**	**1.00**	**0.96**	**0.99**	**1.00**	**0.96**	**0.99**
1e-5	**1.00**	**0.97**	**0.99**	**1.00**	**0.96**	**0.99**	**1.00**	**0.96**	**0.99**

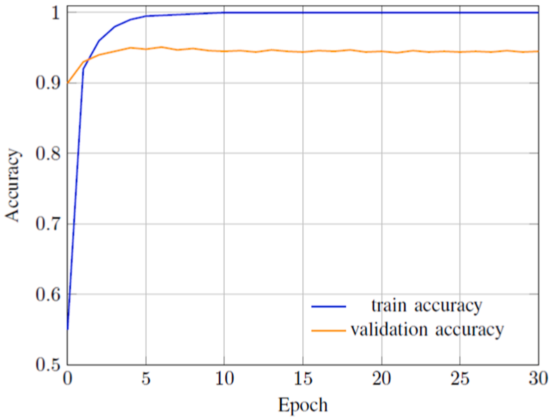

Fig. 5. Training and validation accuracy for GPT 2 model with LR 1e-4 and 1e-5.

Table 4
Results of performance evaluation of Roberta model in intent classification Task.

Learning rate	**ROBERTA**								
	Precision			**Recall**			**F1-score**		
	Training	Validation	Testing	Training	Validation	Testing	Training	Validation	Testing
2e-5	1.00	0.98	0.99	1.00	0.96	0.98	1.00	0.96	0.98
4e-5	1.00	0.94	0.99	1.00	0.93	0.98	1.00	0.93	0.98
3e-6	1.00	0.97	0.99	1.00	0.95	0.99	1.00	0.95	0.99
5e-6	**1.00**	**0.98**	**0.99**	**1.00**	**0.96**	**0.99**	**1.00**	**0.96**	**0.99**

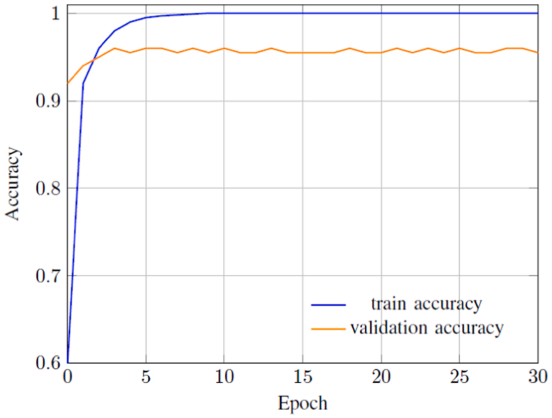

Fig. 6. Training and validation accuracy for Roberta model with LR 5e-6.

Performance evaluation of LLaMA

The following section presents the performance evaluation of the LLaMA 3.1 model in the intent classification task. The model's precision, recall, and F1-score were evaluated across different learning rates, as shown in Table 5. Additionally, Fig. 7 visually represents the training and validation accuracy trends for the LLaMA model with a learning rate of 2e-4.

Discussion

Interpretation of results

The results presented in Tables 2, 3, 4, and 5, alongside Figs. 4, 5, 6, and 7, offer valuable insights into the performance of BERT, GPT-2, RoBERTa, and LLaMA 3.1 models in the context of intent classification within the tourism domain. The analysis of these results allows us to draw several key interpretations.

First, all four models (BERT, GPT-2, RoBERTa, and LLaMA) present strong performance across the evaluated metrics (precision, recall, and F1-score) and datasets (training, validation, and test). This indicates that these models can capture the input data and correctly classify intents.

RoBERTa notably shows consistently high performance across different learning rates, particularly with the learning rate 5e-6, as seen in Table 4. The validation accuracy remains stable throughout the training process, as depicted in Fig. 6.

As shown in Table 2, the BERT model also performs exceptionally well, with near-perfect precision, recall, and F1-scores, especially at lower learning rates. The results demonstrate that BERT is particularly well-suited for intent classification tasks, with minimal differences between the training and validation performances, highlighting its robustness in this domain. Similarly, GPT-2, as presented in Table 3, achieves high performance, particularly at learning rates of 1e-4 and 1e-5. The model's precision, recall, and F1-scores indicate that it can effectively handle intent classification tasks, and Fig. 5 further supports this with consistent validation accuracy over 30 epochs.

The LLaMA 3.1 model, as outlined in Table 5, also demonstrates outstanding performance, particularly at the learning rate of 2e-4. The model maintains near-perfect precision, recall, and F1-scores across training, validation, and test sets, highlighting its ability to handle complex intent classification tasks accurately. The training and validation accuracy trends, shown in Fig. 7, confirm that LLaMA generalizes effectively across different data splits, making it a robust choice for such applications.

The comparison across the models indicates that while all four models perform well, RoBERTa and LLaMA 3.1 slightly outperform BERT and GPT-2 regarding validation metrics. This suggests that RoBERTa and LLaMA may have a slight edge in generalization for this specific task, remarkably when fine-tuned with optimal learning rates.

Computational complexity and resource overhead

The time complexity and computational overhead of training and fine-tuning LLMs such as BERT, GPT-2, RoBERTa, and LLaMA 3.1 are critical considerations, especially when deploying these models in resource-constrained environments like real-time tourism chatbots. The complexity primarily depends on each model's architecture, the dataset's size, and the specific hyperparameters used during training.

For instance, the BERT and RoBERTa models, built on the Transformer architecture, involve time complexities that scale approximately as $O(n^2d)$, where n is the sequence length and d is the dimensionality of the embeddings. This quadratic complexity arises from the self-attention mechanism that requires each token to attend to every other token in the input sequence. The computational overhead for BERT and RoBERTa can be substantial, particularly with longer input sequences, resulting in higher training times and memory consumption. The fine-tuning process for these models, even on modern GPUs, can take several hours to days

Table 5
Results of performance evaluation of Llama 3.1 model in intent classification Task.

Learning rate	Llama 3.1								
	Precision			Recall			F1-score		
	Training	Validation	Testing	Training	Validation	Testing	Training	Validation	Testing
2e-4	1.00	0.99	0.99	1.00	0.98	0.99	1.00	0.98	0.99
1e-5	0.82	0.82	0.80	0.82	0.76	0.78	0.82	0.76	0.78
1e-6	0.87	0.83	0.90	0.87	0.79	0.83	0.87	0.80	0.84

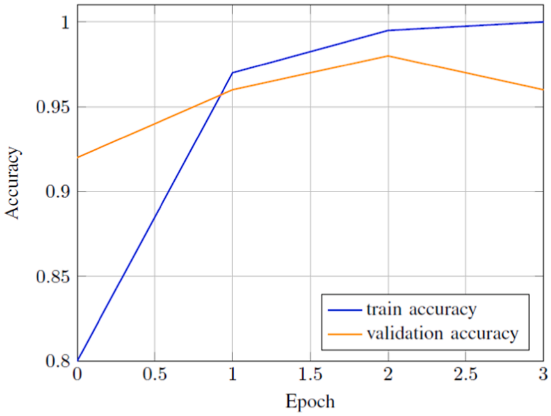

Fig. 7. Training and validation accuracy for Llama 3.1 model with LR 2e-4.

depending on the dataset size and model configuration, leading to significant overhead when frequent updates or retraining are required.

GPT-2, also based on the Transformer architecture, shares similar time complexity concerns. However, its autoregressive nature, where predictions are generated one token at a time, can introduce additional latency during inference. This characteristic can impact the responsiveness of real-time chatbots, making GPT-2 less ideal for applications requiring instantaneous responses.

On the other hand, LLaMA 3.1, a more recent model optimized for efficiency, presents a relatively lower computational overhead than its predecessors. The architectural improvements in LLaMA reduce training and inference time, making it more suitable for deployment in environments where computational resources are limited, or low latency is crucial. However, like all LLMs, fine-tuning still requires considerable GPU memory and computational power.

Overall, the choice of LLM for a tourism chatbot should consider the trade-off between model performance and computational overhead. While models like RoBERTa and LLaMA 3.1 offer robust performance, their time complexity and resource requirements necessitate careful planning regarding the deployment infrastructure and potential optimizations, such as model pruning, distillation, or quantization, to reduce overhead and improve efficiency.

Practical applications and limitations

The findings and models discussed in this study have several practical applications, particularly in smart tourism and AI-based chatbots [Ouaddi, 2024, DSL-Driven Approaches and Metamodels for Chatbot Development: A Systematic Literature Review]. Using advanced language models like BERT, GPT-2, and RoBERTa can significantly enhance the performance and capabilities of chatbots. Below are some key practical applications:

- Tourism chatbots: The models can be deployed in intelligent tourism applications to improve user interaction. Chatbots powered by these models can assist tourists by providing information and answering queries about attractions, amenities, activities, etc.
- Intent classification: These models are beneficial in intent classification tasks, enabling businesses to understand better and respond to customer needs. For instance, e-commerce platforms can use these models to identify purchase intents and offer relevant product information.

In addition to the study's positive results, it is important to know several limitations. Below are two key limitations of the study:

- Domain-specific dataset: The study is based on a dataset specific to the tourism industry, mainly focusing on intents related to the Draa-Tafilalet region in Morocco. While the models perform strongly in this context, the findings may not be generalizable to other domains or areas.

- Computational resource requirements: The fine-tuning of LLMs like BERT, GPT-2, and RoBERTa demands significant computational resources. This limitation may restrict the accessibility of these models for researchers or organizations with limited computational infrastructure.

Conclusion

This study evaluated the performance of several Large Language Models (LLMs) — BERT, GPT-2, RoBERTa, and LLaMA 3.1 — for intent detection in tourism chatbots using a dataset specific to the tourism domain, categorized by the "Six A" criteria. The results show that all models perform well, with RoBERTa and LLaMA 3.1 slightly outperforming BERT and GPT-2 in validation accuracy and F1-score, demonstrating their robustness and generalization capabilities. BERT also performed strongly with near-perfect precision and recall, while GPT-2 showed solid results at optimal learning rates. While the findings indicate that these LLMs are effective for intent classification in tourism chatbots, the study is limited by its use of a domain-specific dataset and the significant computational resources required for fine-tuning. Future research could explore more diverse datasets, hybrid model approaches, and alternative evaluation metrics to enhance further the performance and applicability of LLMs in real-world scenarios, such as incorporating multimodal inputs for more personalized and context-aware chatbot experiences.

CRediT authorship contribution statement

Charaf Ouaddi: Conceptualization, Methodology, Investigation, Software, Data curation, Formal analysis, Resources, Writing – review & editing, Writing – original draft. **Lamya Benaddi:** Conceptualization, Methodology, Investigation, Data curation, Formal analysis, Resources. **El mahi Bouziane:** Conceptualization, Resources. **Lahbib Naimi:** Conceptualization, Resources. **Mohamed Rahouti:** Validation, Writing – review & editing. **Abdeslam Jakimi:** Methodology, Validation, Supervision, Writing – review & editing. **Rachid Saadane:** Validation, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work was supported by the Ministry of Higher Education, Scientific Research and Innovation, the Digital Development Agency (DDA), and the National Center for Scientific and Technical Research (CNRST) of Morocco (Alkhawarizmi/2020/32).

References

[1] L. Benaddi, C. Ouaddi, A. Jakimi, B. Ouchao, A systematic review of chatbots: classification, development, and their impact on tourism, IEEE Access 12 (2024) 78 799. –78 810.

[2] C. Ouaddi, L. Benaddi, A. Souha, A. Jakimi, B. Ouchao, R. Saadane, Exploring and analyzing the impact of chatbots in tourism industry, in: Proceedings of the 7th International Conference on Networking, Intelligent Systems and Security, 2024, pp. 1–6.

[3] H. El Alaoui, Z. El Aouene, V. Cavalli-Sforza, Building intelligent chatbots: tools, technologies, and approaches, in: 2023 3rd International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET), IEEE, 2023, pp. 1–12.

[4] A. Vaswani, Attention is all you need. Advances in Neural Information Processing Systems, 2017.

[5] J. Devlin, Bert: pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[6] A. Souha, C. Ouaddi, L. Benaddi, A. Jakimi, Pre-trained models for intent classification in chatbot: comparative study and critical analysis, in: 2023 6th International Conference on Advanced Communication Technologies and Networking (CommNet), IEEE, 2023, pp. 1–6.

[7] J. Sánchez Cuadrado, S. Pérez-Soler, E. Guerra, J. De Lara, Automating the development of task-oriented llm-based chatbots, in: Proceedings of the 6th ACM Conference on Conversational User Interfaces, 2024, pp. 1–10.

[8] W.X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, et al., A survey of large language models, arXiv preprint arXiv: 2303.18223 (2023).

[9] D. Buhalis, Marketing the competitive destination of the future, Tourism Manage. 21 (1) (2000) 97–116.

[10] R. Sarikaya, The technology behind personal digital assistants: an overview of the system architecture and key components, IEEE Signal Process. Mag. 34 (1) (2017) 67–81.

[11] C. Ouaddi, L. Benaddi, A. Jakimi, Architecture, tools, and dsls for developing conversational agents: an overview, Procedia Comput. Sci. 231 (2024) 293–298.

[12] C. Ouaddi, L. Benaddi, A. Souha, A. Jakimi, A comparative and analysis study for recommending a chatbot development tool, in: 2024 International Conference on Global Aeronautical Engineering and Satellite Technology (GAST), IEEE, 2024, pp. 1–6.

[13] S. Perez-Soler, E. Guerra, J. De Lara, Model-driven chatbot development, in: International Conference on Conceptual Modeling, Springer, 2020, pp. 207–222.

[14] A. Deoras, R. Sarikaya, G. Tür, and D. Hakkani-Tür, "Joint decoding for speech recognition and semantic tagging." in INTERSPEECH, 2012, pp. 1067–1070.

[15] G. Tur, R. De Mori, Spoken Language understanding: Systems for Extracting Semantic Information from Speech, John Wiley & Sons, 2011.

[16] E. Levin, R. Pieraccini, W. Eckert, A stochastic model of human-machine interaction for learning dialog strategies, IEEE Trans. Speech Audio Process. 8 (1) (2000) 11–23.

[17] A. Ashkan, C.L. Clarke, Term-based commercial intent analysis, in: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval, 2009, pp. 800–801.

[18] B. Hollerit, M. Kröll, M. Strohmaier, Towards linking buyers and sellers: detecting commercial intent on twitter, in: Proceedings of the 22nd international conference on world wide web, 2013, pp. 629–632.

[19] J. Weizenbaum, Eliza—A computer program for the study of natural language communication between man and machine, Commun. ACM 9 (1) (1966) 36–45.

[20] R.M. Kaplan, J. Bresnan, et al., Lexical-functional grammar: A formal System For Grammatical Representation, Massachusetts Institute Of Technology, Center For Cognitive Science, 1981.

[21] D. Nadeau, S. Sekine, A survey of named entity recognition and classification, Lingvisticae Investigationes 30 (1) (2007) 3–26.

[22] R. Grishman, Information extraction: techniques and challenges. Information Extraction A Multidisciplinary Approach to an Emerging Information Technology: International Summer School, SCIE-97, Springer, Frascati, Italy, 1997, pp. 10–27. July 14–181997.

[23] J.-C. Na, W.Y.M. Kyaing, C.S. Khoo, S. Foo, Y.-K. Chang, Y.-L. Theng, Sentiment classification of drug reviews using a rule-based linguistic approach, in: The Outreach of Digital Libraries: A Globalized Resource Network: 14th International Conference on Asia-Pacific Digital Libraries, ICADL 2012 14, Springer, Taipei, Taiwan, 2012, pp. 189–198. November 12-15, 2012Proceedings.

[24] T. Zhang, J.H. Cho, C. Zhai, Understanding user intents in online health forums, in: Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics, 2014, pp. 220–229.

[25] M. Mendoza, J. Zamora, Building decision trees to identify the intent of a user query, in: International Conference on Knowledge-Based and Intelligent Information and Engineering Systems, Springer, 2009, pp. 285–292.

[26] K. Li, X. Zhang, Y. Du, A svm based classification of eeg for predicting the movement intent of human body, in: 2013 10th International Conference on Ubiquitous Robots and Ambient Intelligence (URAI), IEEE, 2013, pp. 402–406.

[27] B. Pang, L. Lee, and S. Vaithyanathan, “Thumbs up? Sentiment classification using machine learning techniques,” arXiv preprint cs/0205070, 2002.

[28] M. Mensio, G. Rizzo, M. Morisio, Multi-turn qa: A rnn contextual approach to intent classification for goal-oriented systems, in: Companion Proceedings of the The Web Conference 2018, 2018, pp. 1075–1080.

[29] L. Meng, M. Huang, Dialogue intent classification with long short-term memory networks, in: Natural Language Processing and Chinese Computing: 6th CCF International Conference, NLPCC 2017 6, Springer, Dalian, China, 2018, pp. 42–50. November 8–12, 2017Proceedings.

[30] X. Zhang, H. Wang, A joint model of intent determination and slot filling for spoken language understanding, IJCAI 16 (2016) (2016) 2993–2999.

[31] A. Radford, “Improving language understanding by generative pre-training,” 2018.

[32] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al., Language models are unsupervised multitask learners, OpenAI blog 1 (8) (2019) 9.

[33] B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, et al., Language models are few-shot learners, arXiv preprint arXiv:2005.14165 1 (2020).

[34] L. Benaddi, C. Ouaddi, A. Souha, A. Jakimi, B. Ouchao, Chatbots in tourism sector: classification, evolution, and functionalities, Proc. IEEE (2024).

[35] D.M. Powers, Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation, arXiv preprint (2020).