UDC 681.518
DOI: 10.15587/1729-4061.2025.340492

DETERMINING THE EFFECTIVENESS OF GPT-4.1-MINI FOR MULTICLASS TEXT CATEGORIZATION IN ENGLISH AND UKRAINIAN

Yurii Voloshchuk
Corresponding author
PhD Student*
E-mail: yurii.voloshchuk@uzhnu.edu.ua
Oleksandr Mitsa
Doctor of Technical Sciences, Professor,
Head of Department*
*Department of Informative and
Operating Systems and Technologies
Uzhhorod National University
Narodna sq., 3, Uzhhorod, Ukraine, 88000

The object of this study is the process of multiclass automatic categorization of user queries using large language models under the conditions of a language transition from English to Ukrainian.

The scientific task relates to the fact that most modern large language models (LLMs) are optimized for English while their effectiveness for morphologically complex and low-resource languages, particularly Ukrainian, remains insufficiently studied.

In this work, an experimental approach was devised and implemented to evaluate the transferability of the GPT-4.1-mini model from English to Ukrainian in the task to categorize 11,047 user queries spanning nine applied domains. The analysis employed conventional metrics (Recall, Precision, Weighted-F1, Macro-F1) alongside a novel indicator, the Uncertainty/Error Rate (U/E), which captures the proportion of model refusals and "hallucinations."

The findings demonstrate that the highest quality was achieved on the English dataset (Macro-F1 = 69.78%, U/E = 0.05%). When Ukrainian prompts were applied, Macro-F1 decreased to 63.73%; however, the U/E equaled 0%, indicating higher reliability of responses. Using English prompts with Ukrainian-language data preserved nearly the same level of accuracy (Macro-F1 = 69.66%), thereby revealing strong internal translation and generalization mechanisms.

The novelty of this study is attributed to the use of a large multi-domain parallel corpus, the systematic comparison of prompts in two languages, the application of the state-of-the-art model GPT-4.1-mini, and the introduction of the U/E metric as a reliability criterion. The proposed approach demonstrates the feasibility of applying GPT-4.1-mini to Ukrainian-language information services without additional training, particularly for automatic query routing in financial, medical, legal, and other domains

Keywords: data analysis, large language model, GPT-4.1-mini, text categorization, multilingual evaluation, the Ukrainian language

Received 01.07.2025
Received in revised form 02.09.2025
Accepted date 10.09.2025
Published date 31.10.2025

How to Cite: *Voloshchuk, Y., Mitsa, O. (2025). Determining the effectiveness of GPT-4.1-mini for multiclass text categorization in English and Ukrainian. Eastern-European Journal of Enterprise Technologies, 5 (2 (137)), 98–106. https://doi.org/10.15587/1729-4061.2025.340492*

1. Introduction

Over the past decade, large-scale language models have evolved from laboratory prototypes to large-scale commercial applications, including in support services, content moderation systems, and business analytics [1]. Data analysis based on multilingual assessment methods is increasingly using large-scale language models as a universal text processing tool. However, most academic research and corporate benchmarks are still based on English-language corpora, which causes a significant quality gap for morphologically complex and resource-constrained languages [2].

The Ukrainian language, in particular, is characterized by the presence of seven cases, productive prefixation, and developed synonym series, which complicates the direct transfer of models trained on the English-language material [3]. At the same time, the military-political context highlights the need to set up national information services, in which high-quality automatic classification of citizens' requests is crucial for the effective functioning of the systems.

The need for research under current conditions is justified by the following factors:

– legislative imperative. With the entry into force of the Law of Ukraine "On Ensuring the Functioning of the Ukrainian Language as the State Language", digital platforms are obliged to provide full-fledged Ukrainian-language interaction [4];

– business need for scalable automation. Conventional machine learning approaches to categorizing requests require laborious additional training and significant time resources while large language models promise "zero-shot" or "few-shot" performance without additional training [5];

– high cost of inference and reliability factor. Commercial APIs (e.g., GPT-4 series) remain expensive; therefore, it is important not only to assess accuracy [6, 7] but also to take into account the probability and cost of failures or generation of incorrect answers ("hallucinations") [8, 9]. The combination of Macro-F_1 and U/E metrics makes it possible to quantitatively measure the total proportion of uncertain or erroneous responses [10];

– lack of representative studies. Earlier works focus mainly on multilingual models of the XLM-R type; modern architectures based on GPT are not sufficiently studied in the context of the Ukrainian language [11, 12].

The results of investigating the possibilities of text categorization using the zero-shot approach involving GPT-4.1-mini have considerable practical value:

– optimization of the "cost/quality" ratio. Empirical results will contribute to the selection of a model that provides the optimal balance between accuracy and cost of inference in scalable services;

– reduction of the load on operators. Improving the accuracy of automatic query routing systems in the Ukrainian language will reduce the amount of manual processing and decrease the response time to customers;

– universal metric framework. The proposed combination of Macro-F_1 and U/E metrics is suitable for assessing the reliability of LLMs in regulated areas (in particular, finance and healthcare) where false or fictitious model responses ("hallucinations") are unacceptable;

– an incentive for the development of low-resource NLP. The presence of an open benchmark and generalized recommendations could be an impetus for further research into Ukrainian and other low-resource languages, which is especially relevant in the current military-political context.

Therefore, a systematic evaluation of GPT-4.1-mini in the text categorization task using multilingual assessment is an important task for modern applied linguistics and business analytics. Hence, it is a relevant task to carry out studies on the capabilities and limitations of GPT-4.1-mini in Ukrainian text categorization.

2. Literature review and problem statement

Over the past decade, the task of multilingual text categorization has evolved from classical machine learning methods to approaches based on large language models capable of working under a zero-shot mode. The principal achievements and existing research gaps are summarized below.

In study [13], a zero-shot methodology for text categorization integrated with knowledge bases is reported, which works even when there is no training data for a specific task at all, and all labels of the test set are unknown. The proposed approach does not require additional training for a specific task and is model-agnostic (any sentence transformer can be substituted). Testing on six open sets (general-purpose: DBPedia, AG's News, Yahoo; domain-specific: Law Stack Exchange, BugClassify, Medical Abstract) showed that the system outperforms previous state-of-the-art zero-shot methods and sets new benchmarks without using large generative LLMs. However, the effectiveness of this method for low-resource languages has not been investigated.

This gap is partially closed in work [14]. The authors carried out a systematic comparison of ELMo and BERT family models for nine low-resource European languages. The evaluation was carried out in 14 tasks: named entity recognition, part-of-speech identification, dependency parsing, lexical analogies, contextual similarity, terminology alignment, and six SuperGLUE subtasks. The study offers practical recommendations for the selection and tuning of compact encoders when the use of very large generative LLMs is impossible. However, neither Ukrainian language nor the latest LLM models were included in the experiments.

In [15], data analysis of four open LLMs (Mistral-7b, Llama3-8b, Mixtral-8x7b, Llama3-70b) and the XLM-RoBERTa encoder model on two multilingual corpora – MultiEURLEX (legal texts, multilabel classification) and AmazonReview (consumer reviews) for English, French, and Spanish is reported. The authors investigated three cross-lingual transfer strategies: zero-shot direct-test, direct-test with fine-tuning on English data, and translate-test with further fine-tuning. The experiments were conducted in generative and embedded classification configurations. It is shown that fine-tuning significantly improves the quality of the results while translate-test provides the best performance. At the same time, LLMs show a slight decrease in accuracy on non-English texts, but they mostly outperform XLM-RoBERTa, which remains a competitive and less resource-intensive solution for categorization tasks on MultiEURLEX. However, the study did not cover resource-constrained languages and did not include models from OpenAI.

The GPT-series of LLMs was investigated in [16]. The authors tested GPT-3 on seven typical NLP tasks: part-of-speech detection, named entity recognition, relation detection, logical inference, question-answer, common sense logic, and generalization. However, the multi-domain categorization of customer questions was not included in this list as it requires practical semantic understanding of the text, not just processing linguistic markers. The Ukrainian language was represented only in the context of Relation Extraction, so its effectiveness on other tasks remained unassessed. In addition, the study dates back to 2023 and does not cover the GPT-4 series, which significantly improved the quality of multilingual modeling.

Interestingly, the authors compared non-English datasets using both English and non-English prompts, which is not found in most similar works. However, the U/E Rate metric was not analyzed in the work, so the safety of the responses was left out of consideration.

In [17], fine-tuning of four pre-trained transformers – BERT-base, DistilBERT, XLM-RoBERTa and UkrRoBERTa – was performed on a corpus of Ukrainian responses, with a focus exclusively on the task of sentiment analysis. However, the proposed solution is not suitable for use in environments with a lack of annotation resources or high cost of their preparation. Using the modern LLM model GPT-4.1-mini under a zero-shot mode makes it possible to categorize queries without additional training.

A recent paper [18] analyzed the performance of two models – the open-source Mistral and the multilingual XLM-RoBERTa – on three tasks: toxicity detection, formality assessment, and natural language inference (NLI). However, the performance of the models was not evaluated on semantically more complex queries that require a deeper understanding of the context in different application domains. In addition, in [18], only English prompts were used, and the impact of the prompt language (English and Ukrainian) on the quality of the results was not evaluated.

In [19], three new Ukrainian corpora are presented: UA-CBT (filling in the gaps in the text), UP-Titles (choosing a title from ten options), and LMES (elementary linguistic tests). Similar to [18], these corpora do not make it possible to evaluate the performance of the models in scenarios that require deeper semantic analysis of the data. In [19], an early version of the LLM model from Open AI GPT-4 (2024) was tested. However, due to a limited budget, the experiments covered only a few hundred examples, which reduces their statistical significance. In addition, only Ukrainian-language prompts were used, and the question of the model's performance on Ukrainian and English prompts remains open.

The Uncertainty/Error Rate (U/E) methodology was proposed in [10] for quantitative assessment of failures and "hallucinations" of large language models but it has not been applied to cross-lingual text categorization tasks or for low-resource languages.

Despite the variety of topics, all the above studies have a number of common limitations:

– first, none of the papers explores the GPT-4.1-mini model or other recent modifications of GPT-4 on a large Ukrainian-language corpus. Thus, the capabilities of modern models in the context of the Ukrainian language remain poorly studied. Most existing research either focuses on open-source models or is based on previous generations of GPT;

– second, in most cases, either exclusively English-language or only Ukrainian-language prompts are used, which does not make it possible to assess the impact of the query language on the quality of the result. Researchers either rely on external machine translation systems or do not compare results for prompts in different languages because of the doubled computational costs and complexity of the research design;

– third, the Uncertainty/Error Rate metric is almost not used in the reviewed papers. The U/E methodology was proposed only in 2024, so the studies of 2020–2023 focused mainly on F_1 indicators and did not take into account the proportion of failures or false responses (hallucinations);

– fourth, the datasets used are often domain-specific (e.g., tonality analysis, toxicity detection) or small in size, which does not make it possible to reflect the real variability of queries in application scenarios. The question of the role of morphological complexity and token inflation as key factors in reducing accuracy in individual domains remains open within the framework of modern research. The domain classes Home Improvement, Tax, Legal often contain a large number of word forms and specialized terms, which complicates annotation. In addition, the lack of agreed multilingual corpora with parallel markup does not make it possible to clearly separate the influence of the language factor and thematic complexity.

Thus, a complex issue related to the functioning of the state-of-the-art LLM model GPT-4.1-mini on a large parallel English-Ukrainian corpus of multi-domain categorization remains open if we simultaneously evaluate its accuracy (Macro-F_1) and reliability (Uncertainty/Error Rate, U/E). It is also important to study the effectiveness of the impact of the prompt language (English and Ukrainian) on the results.

The lack of such an assessment significantly hinders the implementation of Ukrainian-language LLM solutions in business analytics, the public sector, and user services. Our study is designed to fill the scientific and applied vacuum in this area.

3. The aim and objectives of the study

The purpose of our study is to assess the effectiveness of using the large language model GPT-4.1-mini for multi-class categorization of user questions in Ukrainian and English. This will make it possible to determine the practical conditions for the effective use of the model in Ukrainian-language services, taking into account the ratio of accuracy, computational cost, as well as the risk of generating "uncertain" or irrelevant answers.

To achieve the goal, the following tasks were set:

– to prepare a parallel dataset: generate an English-language set of questions with nine classes and obtain a Ukrainian equivalent through machine translation of GPT-4.1-mini with minimal post-editing;

– to categorize the English-language dataset using the GPT-4.1-mini model and evaluate progress compared to previous versions;

– to categorize the Ukrainian-language dataset using both English and Ukrainian prompts, and compare the results in terms of accuracy (Macro-F_1) and reliability (U/E);

– to identify domain-specific patterns: to analyze the changes in F_1 for each of the nine classes to determine which subject areas are most sensitive to language transition;

4. The study materials and methods

4. 1. The object, hypothesis, and acceptable assumptions

The object of our study is the process of multi-class automatic categorization of user queries using the large language model GPT-4.1-mini under the conditions of the language transition "English → Ukrainian".

The hypothesis of the study assumes that without additional training, the GPT-4.1-mini model is able to provide acceptable accuracy and a low level of uncertainty (Uncertainty/Error Rate, U/E) when categorizing Ukrainian-language queries, while maintaining at least 90% of the Macro-F_1 obtained on English-language data.

Acceptable assumptions:

– machine translation performed by GPT-4.1-mini preserves semantic correspondence at a reasonable level so that the impact of translation errors on categorization is less than the impact of morphology;

– a single run of inference at temperature = 0 gives representative results. This was confirmed in a previous study [20] for the GPT-4-turbo model. The standard deviation for the F1 score was 0.19%, which is a good indicator of the repeatability of results.

Simplifications:

– no fine-tuning is performed; all experiments were performed under a zero-shot mode;

– the categorization task is implemented as single-label: the model must select only one category from the available list.

4. 2. Input data

The English corpus contains 11,047 unique user questions, distributed among nine categories: Car, Veterinary, Computer, Consumer Electronics, Tax, Home Improvement, Homework, Legal, Medical. The distribution is uneven, which was taken into account when choosing metrics.

Fig. 1 shows the schematic of translating the English corpus into Ukrainian.

Translation into Ukrainian was performed using the GPT-4.1-mini API (USA). Results were validated using the TRANSLATE() function in MS Excel (USA) by comparing the number of characters in the translations from both sources. Cases where the deviation exceeded 10% were reviewed and corrected manually.

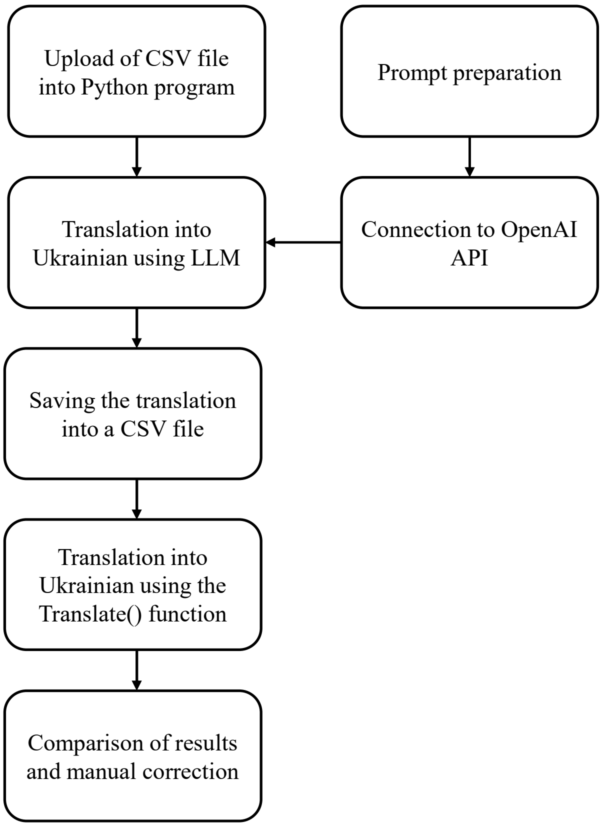
Upload of CSV file into Python program
Prompt preparation
Translation into Ukrainian using LLM
Connection to OpenAI API
Saving the translation into a CSV file
Translation into Ukrainian using the Translate() function
Comparison of results and manual correction

Fig. 1. Schematic of translating the English corpus into Ukrainian

4. 3. Choosing a large language model

The MMLU (Measuring Massive Multitask Language Understanding) and Multilingual MMLU metrics for current Open AI models [21] were analyzed, as well as the cost of 1 million input and output tokens [22]. To find the optimal solution from a price/quality perspective, the model was rated for each of these metrics and the overall rating was summarized. The comparison results are given in Table 1.

The best rating is achieved by the GPT-4.1-mini model, which has relatively high MMLU = 87.5% and Multilingual MMLU = 78.5%, at an acceptable input/output token price of USD 0.4/USD 1.6. The latest version of the model at the time of the research was used – gpt-4.1-mini-2025-04-14.

4. 4. Prompts and inference mode

To ensure comparability with previous studies of GPT-3.5-turbo and GPT-4-turbo models, an identical prompt was used, which demonstrated the best results [23]: "You are category moderator. Your job is to classify the customer's problem based on their question text. You have to choose ONLY one of predefined categories and don't provide any description: {list of categories}". Ukrainian version of the prompt: "Ти – модератор категорій. Твоє завдання – класифікувати проблему клієнта на основі тексту його запитання. Тобі потрібно вибрати ЛИШЕ одну з попередньо визначених категорій і не надавати жодного опису: {список категорій}". API operating parameters: temperature = 0, max_tokens = 65.

4. 5. Evaluation methodology and metrics

The experimental data processing procedure is shown in Fig. 2 in the form of a functional diagram.

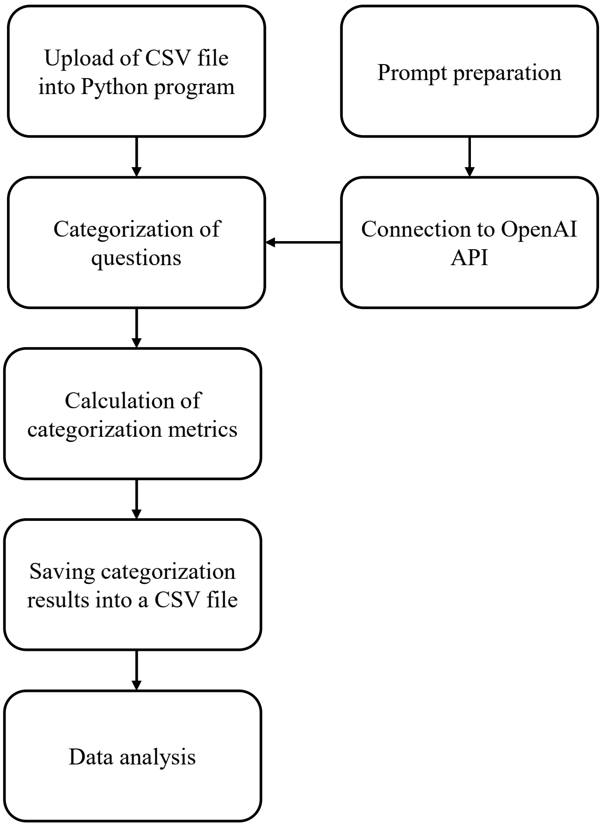
Upload of CSV file into Python program
Prompt preparation
Categorization of questions
Connection to OpenAI API
Calculation of categorization metrics
Saving categorization results into a CSV file
Data analysis

Fig. 2. Query categorization scheme using GPT-4.1-mini

Table 1

LLM comparison based on MMLU, Multilingual MMLU, and price

Model	MMLU, %	Multilingual MMLU, %	Token price (1M), USD		Score				
			Input	Output	MMLU	Multilingual MMLU	Input price	Output price	Total
GPT-4.1 mini	87.5	78.5	0.4	1.6	4	6	2	2	14
GPT-4.1	90.2	87.3	2	8	3	2	5	5	15
OpenAI o1	91.8	87.7	15	60	1	1	7	7	16
GPT-4.1 nano	80.1	66.9	0.1	0.4	8	8	1	1	18
OpenAI o3-mini	86.9	80.7	1.1	4.4	5	5	4	4	18
GPT-4o mini	82.0	70.5	0.6	2.4	7	7	3	3	20
GPT-4.5	90.8	85.1	75	150	2	3	8	8	21
GPT-4o	85.7	81.4	5	20	6	4	6	6	22

Input data processing in CSV format was carried out using the Python software: each query was sent as a separate API request, and the result was written to a new CSV file. Text categorization in English and Ukrainian was performed by one program. Only input and output files were changed and prompts in English and Ukrainian were used for the Ukrainian-language corpus.

The standard categorization metrics Recall, Precision, Weighted-F_1, Macro-F_1, calculated by the open-source scikit-learn library, were used for evaluation. Uncertainty/Error Rate (U/E) was manually calculated: U – answers of the type "I'm not sure...", "Cannot categorize ..."; E – non-existent categories or completely irrelevant text were received. This approach ensures that data analysis considers accuracy and reliability, which is necessary for implementation in practical systems. The assessment was conducted in a multilingual assessment format where the English-language and Ukrainian-language prompts were compared.

4. 6. Ethical considerations

All queries were previously anonymized and do not contain personal data. Use of the API complies with the terms of the OpenAI policy.

5. Results of investigating the effectiveness of GPT-4.1-mini for categorization of queries in English and Ukrainian

5. 1. Preparation of a parallel dataset

Table 2 gives the final class distribution in the original English and translated Ukrainian datasets.

After machine translation, the number of lines and domain balance were preserved without loss. At the same time, the number of characters in the Ukrainian-language set increased by 12% compared to the English-language set.

Table 2

Comparison of English and Ukrainian datasets

Category	English-language data		Ukrainian-language data		Difference	
	Number of queries	Number of characters	Number of queries	Number of characters	Number of queries	Number of characters
Car	2530	419986	2530	483070	0	+15%
Computer	1720	261707	1720	291383	0	+11%
Consumer Electronics	758	117050	758	131126	0	+12%
Home Improvement	1494	265839	1494	299964	0	+13%
Homework	27	3786	27	4209	0	+11%
Legal	1606	272233	1606	303492	0	+11%
Medical	705	119176	705	132805	0	+11%
Tax	439	66867	439	78577	0	+18%
Veterinary	1768	315876	1768	338897	0	+7%
Total	11047	1842520	11047	2063523	0	+12%

5. 2. Categorization of the English-language dataset using the GPT-4.1-mini model

Table 3 gives the results from categorizing the English-language dataset using the GPT-3.5-turbo, GPT-4-turbo, and GPT-4.1-mini models.

Table 3

Comparing the results of categorizing an English-language dataset by different OpenAI models

Metric	GPT-3.5-turbo	GPT-4-turbo	GPT-4.1-mini
Recall	70.38%	71.50%	75.70%
Precision	79.84%	85.43%	83.65%
Weighted-F_1	71.97%	72.48%	75.42%
Macro-F_1 (unweighted)	66.52%	68.11%	69.78%
U/E rate	0.82%	0.18%	0.05%

As we can see, the new model has better weighted/unweighted F1 performance and a significantly lower U/E rate.

5. 3. Categorization of the Ukrainian-language dataset using the GPT-4.1-mini model and applying English and Ukrainian prompts

Table 4 gives the results from categorizing the Ukrainian-language dataset.

Table 4

Comparison of categorization results of the English-language and Ukrainian-language datasets using prompts in both languages

Metric	GPT-4.1-mini English dataset	GPT-4.1-mini Ukrainian dataset English prompt	GPT-4.1-mini Ukrainian dataset Ukrainian prompt
Recall	75.70%	76.18%	71.18%
Precision	83.65%	82.63%	78.01%
Weighted-F_1	75.42%	75.70%	70.38%
Macro-F_1 (unweighted)	69.78%	69.66%	63.73%
U/E rate	0.05%	0.08%	0.00%

The results that involve the English prompt are almost identical to the results for the English set. When using the Ukrainian prompt, there is a decrease in Macro-F_1 by about 6%, although the U/E indicator is zero.

5. 4. Detecting domain-specific patterns

Fig. 3 shows F_1 values for different classes.

The decrease in the overall result for the configuration "Ukrainian data – Ukrainian prompt" is mainly due to lower indicators in the Home Improvement, Consumer Electronics, and Tax classes.

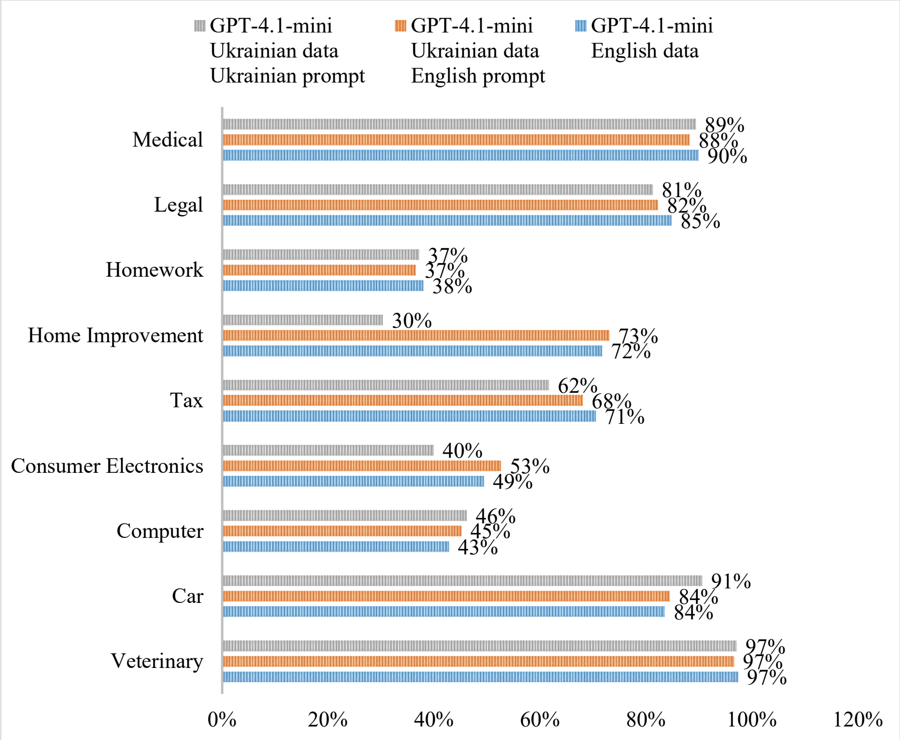

Fig. 3. Comparison of F_1 scores for different classes in three experiment configurations

6. Discussion of results related to investigating the GPT-4.1-mini efficiency for categorization of queries in English and Ukrainian

Our results, specifically Macro-F_1 = 69.78% and Weighted-F_1 = 75.42% for the GPT-4.1-mini model, confirm the increase in the efficiency of applying zero-shot prompting methods for categorization tasks of short texts, in particular user-defined questions. Consideration of the results in the context of each task makes it possible to explain the reasons for the observed effects, as well as to determine their advantages and limitations in comparison with existing approaches.

The increase in the length of the input text after translation by 12% (Table 2) is due to the morphological richness of the Ukrainian language. It has 7 cases, productive prefixation, and a developed suffix system, which leads to an increase in the number of word forms compared to English. In addition, Ukrainian words are often longer than their English counterparts ("homework" → "домашнє завдання"). This causes "token inflation" for Slavic languages, which is reflected in study [24]. Thus, the increase in input length is an expected linguistic effect, and not a consequence of translation error.

As given in Table 3, the GPT-4.1-mini model outperformed GPT-3.5-turbo and GPT-4-turbo in terms of Weighted-/Macro-F_1 metrics (75.42% and 69.78%, respectively) at an uncertain response rate of U/E = 0.05%. The increased accuracy is associated with the updated architecture (context per 1 million tokens) and increased transformer depth. It is important to note the absence of cases where the model responded empathetically or tried to answer the question instead of categorization – behavior typical of previous models, according to [6, 7]. The residual U/E is explained by only six cases of incorrect category assignment outside the list (Travel – 3, Financial – 2, Banking – 1).

The data in Table 4 demonstrate that applying the English prompt to the Ukrainian text provides almost equivalent Macro-F_1 (69.66%) compared to the English set (69.78%). In contrast, the Ukrainian prompt reduces the accuracy to 63.73% but completely eliminates ambiguous responses (U/E = 0). This is explained by the fact that when using the English template, the model is oriented to a well-trained corpus of instructions while the Ukrainian prompt has a lower correspondence to the pre-train distribution but reduces language collisions, increasing the confidence of the model.

As shown in Fig. 3, the Home Improvement, Consumer Electronics, and Tax domains experienced the largest decrease in F_1 due to terminological complexity and morphological variability. The largest number of misclassifications – 1106 cases – was inherent in Home Improvement and Consumer Electronics. The reasons include the following:

– lexical overlap: queries about repairs contain words like "thermostat", "sensor", "LED", which tend to be related to electronics;

– morphological polysemy of Ukrainian words – the Ukrainian terms "розетка", "вимикач", "провід" are used equally as elements of household electrical networks (Home Improvement) and as accessories to gadgets (Consumer Electronics). GPT-4.1-mini without the context "repair-or-gadget" tends to the more frequent pattern. The term "плата" may mean "circuit board" in the Consumer Electronics domain, or "fee/payment" in the Tax or Legal domain. Another example is "ключ", which is used as "key" for a lock (Home Improvement/Car), as "keyword" in the Computer domain, or as "solution key" in Homework. Within the framework of zero-shot text categorization without additional training, such polysemous words cause errors in determining the category, especially when there is a lack of extended context;

– translation artifacts and tokenization bias – some Home Improvement queries could have received electronically colored words during translation ("controller", "smart device"), increasing semantic drift. Some English-language technical concepts do not have a stable translation into Ukrainian ("router", "sensor"), so the corpus contains both translations ("маршрутизатор") and transcribed borrowings. This increases the likelihood of semantic shifts. In addition, the Ukrainian compound words ("шестигранник", "шліфмашина") are broken into 3–4 tokens, reducing their "weight" in the context, while borrowings such as "router", "sensor" remain one token and have a stronger impact on the attention mechanism.

Thus, the Ukrainian dataset, although it retains full correspondence to the English-language one in terms of class structure and number of examples, is "heavier" in terms of tokenization and morphological variability. These factors explain the increase in the cost of inference and possible differences in the accuracy of text categorization. Based on our results, the following recommendations can be formulated for the application of LLM in Ukrainian-language services:

1. The GPT-4.1-mini model is reasonable in terms of price-performance ratio. It outperforms previous versions in terms of accuracy at a significantly lower cost.

2. Ukrainian-language queries have a larger volume of characters, which causes an increase in processing costs. If possible, it is advisable to use English-language data as input.

3. Using an English prompt for Ukrainian-language queries provides high quality categorization, close to the English-language configuration.

4. When using a Ukrainian-language prompt, it is advisable to use additional glossaries in sensitive domains (especially Home Improvement, Consumer Electronics, Tax) to improve accuracy.

The proposed practical recommendations will help businesses better understand the capabilities of large language models and their use in real scenarios.

Unlike work [18], in which the Ukrainian language is considered only in Relation Extraction and without U/E, the task to categorize user queries is a more complex practical task. The result at Macro-F_1 = 63.73% and U/E = 0% for nine domains proves the suitability of GPT-4-level for semantically heterogeneous service. The combination of accuracy and reliability metrics allows for a comprehensive assessment of categorization results.

In [19], an early version of GPT-4 was tested on several hundred examples. The analysis of 11,047 queries of various topics increases statistical significance and makes it possible to detail domain failures. Such results allow us to resolve the model gap – GPT-4.1-mini was tested on a large Ukrainian corpus. It is shown that the English prompt retains its effectiveness on Ukrainian text. At the same time, the Ukrainian prompt reduces accuracy by ≈6%. Separate domains were identified that require more complex prompting techniques to obtain acceptable results with Ukrainian prompts.

Unlike study [17], which uses fine-tuning of BERT models, the zero-shot approach opens up wider possibilities for application. Such a solution is immediately scalable to new domains and is suitable for conditions of limited or no resources for training models. However, it should be noted that there is a potential for improving the results in certain classes by building glossaries for problem domains or using knowledge transfer frameworks [25, 26].

It is necessary to take into account the practical limitations of this study. First, the lack of an "external" gold translation does not make it possible to separate translation error from categorization error. The second limitation is the lack of fine-tuning. The aim of our study was to test the zero-shot approach, but it is possible that the categorization result for the Home Improvement class would have improved after several demonstrations (few-shot).

The shortcomings of this work include performing only one run per configuration. Although the σ calculated in [20] is minimal, 5–10 runs are required for more stringent confidence intervals in each configuration. To offset the cost of a large number of runs, future studies should consider using a cascade of models (e.g., nano → mini) to optimize costs without compromising quality [27]. Another drawback is the imbalance of classes (Homework has only 27 examples), which potentially underestimates Macro-F_1.

Future research into this area will focus on reverse experiments comparing the categorization of the original Ukrainian data with their translation into English and the use of prompts in both languages. To complicate the task and approach real conditions, it is worth considering the use of a dataset generated using a special platform – the Ukrainian dialect map [28]. To improve the performance of the model, a multi-threshold approach can be used, based on the use of both discrete activation functions [29, 30] and their continuous smoothed modifications [31]. It is advisable to compare the results of the study with approaches based on classical machine learning algorithms [32].

Thus, our results have been logically explained; they demonstrate the advantage of GPT-4.1-mini over previous models and outline the boundaries and prospects of its practical application in Ukrainian-language services.

7. Conclusions

1. A high-quality parallel corpus with controlled language effects has been formed. The English and Ukrainian samples retained complete coincidence in 11,047 queries and nine classes. After translation, the number of characters increased by 12%, which indicates the absence of thematic losses but captures the effect of "tokenization inflation" – an important factor in increasing the cost of inference for Ukrainian-language text.

2. The GPT-4.1-mini model demonstrates improved accuracy and reliability for English-language categorization. Compared to previous versions, the model achieved better indicators (Macro-F_1=69.78%, U/E=0.05%), which indicates a decrease in the number of uncertain answers and increased stability due to a deeper architecture and 1M-context.

3. Using an English prompt for Ukrainian-language queries preserves the quality of categorization. Translating queries into Ukrainian did not result in significant losses: Macro-F_1 was 69.66% (a difference of −0.12% compared to the English version). However, when using the Ukrainian prompt, the Macro-F_1 index decreased to 63.73% while the level of uncertainty (U/E) dropped to zero. This indicates that the main factor in the loss of accuracy is the language of the prompt, not the data.

4. The largest losses in accuracy were recorded in domains with similar vocabulary. In particular, the Home Improvement category lost 41.35% F_1 when switching to the Ukrainian prompt, due to terminological overlap with the Consumer Electronics domain. In contrast, the Computer category showed even better results on the Ukrainian set. This proves that morphology and semantic overlap determine qualitative losses.

Conflicts of interest

The authors declare that they have no conflicts of interest in relation to the current study, including financial, personal, authorship, or any other, that could affect the study, as well as the results reported in this paper.

Funding

The study was conducted without financial support.

Data availability

The data will be provided upon reasonable request.

Use of artificial intelligence

The authors used artificial intelligence technologies within the permissible limits to conduct research (this study analyzes capabilities of artificial intelligence) and improve readability and grammar when writing the paper.

References

1. Huang, J., Xu, Y., Wang, Q., Wang, Q. (Cheems), Liang, X., Wang, F. et al. (2025). Foundation models and intelligent decision-making: Progress, challenges, and perspectives. The Innovation, 6 (6), 100948. https://doi.org/10.1016/j.xinn.2025.100948
2. Doddapaneni, S., Ramesh, G., Khapra, M., Kunchukuttan, A., Kumar, P. (2025). A Primer on Pretrained Multilingual Language Models. ACM Computing Surveys, 57 (9), 1–39. https://doi.org/10.1145/3727339
3. Yermolenko, S. (2019). From the history of Ukrainian stylistics: from stylistics of languages to integrative stylistics. Ukrainska Mova, 1, 3–17. https://doi.org/10.15407/ukrmova2019.01.003
4. Zakon Ukrainy «Pro zabezpechennia funktsionuvannia ukrainskoi movy yak derzhavnoi» No. 2704-VIII. Verkhovna Rada Ukrainy. Available at: https://zakon.rada.gov.ua/laws/show/2704-19
5. Syromiatnikov, M. V., Ruvinskaya, V. M., Troynina, A. S. (2024). ZNO-Eval: Benchmarking reasoning capabilities of large language models in Ukrainian. Informatics. Culture. Technology, 1 (1), 186–191. https://doi.org/10.15276/ict.01.2024.27
6. Mitsa, O., Voloshchuk, Y., Levchuk, O., Petsko, V. (2025). A Comparative Study of Machine Learning Algorithms and the Prompting Approach Using GPT-3.5 Turbo for Text Categorization. Advances in Computer Science for Engineering and Education VII, 156–167. https://doi.org/10.1007/978-3-031-84228-3_13
7. Voloshchuk, Y. O., Mitsa, O. V. (2024). Comparison of text categorization efficiency using the prompting approach with GPT-3.5-turbo and GPT-4-turbo. Science and Technology Today, 6 (34), 768–777. https://doi.org/10.52058/2786-6025-2024-6(34)-768-777
8. Garrachón Ruiz, A., de La Rosa, T., Borrajo, D. (2025). TRIM: Token Reduction and Inference Modeling for Cost-Effective Language Generation. arXiv. https://doi.org/10.48550/arXiv.2412.07682
9. Chen, L., Zaharia, M., Zou, J. (2024). FrugalGPT: How to use large language models while reducing cost and improving performance. Transactions on Machine Learning Research. Available at: https://openreview.net/forum?id=cSimKw5p6R
10. Wang, Z., Pang, Y., Lin, Y., Zhu, X. (2024). Adaptable and Reliable Text Classification using Large Language Models. 2024 IEEE International Conference on Data Mining Workshops (ICDMW), 67–74. https://doi.org/10.1109/icdmw65004.2024.00015
11. Panchenko, D., Maksymenko, D., Turuta, O., Yerokhin, A., Daniiel, Y., Turuta, O. (2022). Evaluation and Analysis of the NLP Model Zoo for Ukrainian Text Classification. Information and Communication Technologies in Education, Research, and Industrial Applications, 109–123. https://doi.org/10.1007/978-3-031-20834-8_6
12. Panchenko, D., Maksymenko, D., Turuta, O., Luzan, M., Tytarenko, S., Turuta, O. (2022). Ukrainian News Corpus as Text Classification Benchmark. ICTERI 2021 Workshops, 550–559. https://doi.org/10.1007/978-3-031-14841-5_37
13. Wang, Y., Wang, W., Chen, Q., Huang, K., Nguyen, A., De, S. (2024). Zero-shot text classification with knowledge resources under label-fully-unseen setting. Neurocomputing, 610, 128580. https://doi.org/10.1016/j.neucom.2024.128580
14. Ulčar, M., Žagar, A., Armendariz, C. S., Repar, A., Pollak, S., Purver, M., Robnik-Šikonja, M. (2026). Mono- and cross-lingual evaluation of representation language models on less-resourced languages. Computer Speech & Language, 95, 101852. https://doi.org/10.1016/j.csl.2025.101852
15. Han, B., Yang, S. T., LuVogt, C. (2025). Cross-Lingual Text Classification with Large Language Models. Companion Proceedings of the ACM on Web Conference 2025, 1005–1008. https://doi.org/10.1145/3701716.3715567
16. Lai, V., Ngo, N., Pouran Ben Veyseh, A., Man, H., Dernoncourt, F., Bui, T., Nguyen, T. (2023). ChatGPT Beyond English: Towards a Comprehensive Evaluation of Large Language Models in Multilingual Learning. Findings of the Association for Computational Linguistics: EMNLP 2023. https://doi.org/10.18653/v1/2023.findings-emnlp.878
17. Prytula, M. (2024). Fine-tuning BERT, DistilBERT, XLM-RoBERTa and Ukr-RoBERTa models for sentiment analysis of Ukrainian language reviews. Artificial Intelligence, 2, 85–97. https://doi.org/10.15407/jai2024.02.085
18. Dementieva, D., Khylenko, V., Groh, G. (2025). Cross-lingual text classification transfer: The case of Ukrainian. arXiv. https://doi.org/10.48550/arXiv.2404.02043
19. Hamotskyi, I., Levbarg, A., Hänig, C. (2024). Eval-UA-tion 1.0: Benchmark for Evaluating Ukrainian (Large) Language Models. UNLP 2024. Available at: https://hal.science/hal-04534651v2
20. Voloshchuk, Yu., Mitsa, O. (2024). Otsinka stabilnosti rezultativ katehoryzatsiyi tekstu z vykorystanniam prompting pidkhodu z velykymy movnymy modeliamy. Materialy konferentsii MTsND. https://doi.org/10.62731/mcnd-21.06.2024.002
21. GPT-4.1. OpenAI. Available at: https://openai.com/index/gpt-4-1/
22. Pricing. OpenAI. Available at: https://platform.openai.com/docs/pricing

23. Voloshchuk, Yu., Mitsa, O. (2024). Porivniannia prompting pidkhodiv z vykorystanniam gpt 4-turbo dlia tekstovoi katehoryzatsiyi. Materialy konferentsii MTsND. https://doi.org/10.62731/mcnd-14.06.2024.006
24. Ahia, O., Kumar, S., Gonen, H., Kasai, J., Mortensen, D., Smith, N., Tsvetkov, Y. (2023). Do All Languages Cost the Same? Tokenization in the Era of Commercial Language Models. Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 9904–9923. https://doi.org/10.18653/v1/2023.emnlp-main.614
25. Li, X., Zhang, K. (2025). Heterogeneous Graph Neural Network with Multi-View Contrastive Learning for Cross-Lingual Text Classification. Applied Sciences, 15 (7), 3454. https://doi.org/10.3390/app15073454
26. Gui, A., Xiao, H. (2024). Multi-level multilingual semantic alignment for zero-shot cross-lingual transfer learning. Neural Networks, 173, 106217. https://doi.org/10.1016/j.neunet.2024.106217
27. Huang, K., Shi, Y., Ding, D., Li, Y., Fei, Y., Lakshmanan, L., Xiao, X. (2025). ThriftLLM: On Cost-Effective Selection of Large Language Models for Classification Queries. Proceedings of the VLDB Endowment, 18 (11), 4410–4423. https://doi.org/10.14778/3749646.3749702
28. Mitsa, O., Sharkan, V., Maksymchuk, V., Varha, S., Shkurko, H. (2023). Ethnocultural, Educational and Scientific Potential of the Interactive Dialects Map. 2023 IEEE International Conference on Smart Information Systems and Technologies (SIST), 226–231. https://doi.org/10.1109/sist58284.2023.10223544
29. Kotsovsky, V. (2024). Learning of Multi-valued Multithreshold Neural Units. Proceedings of the 8th International Conference on Computational Linguistics and Intelligent Systems. Volume III: Intelligent Systems Workshop. https://doi.org/10.31110/colins/2024-3/004
30. Kotsovsky, V., Batyuk, A. (2024). Towards the Design of Bithreshold ANN Regressor. 2024 IEEE 19th International Conference on Computer Science and Information Technologies (CSIT), 1–4. https://doi.org/10.1109/csit65290.2024.10982560
31. Kotsovsky, V. (2025). Multithreshold neurons with smoothed activation functions. Proceedings of the Intelligent Systems Workshop at 9th International Conference on Computational Linguistics and Intelligent Systems (CoLInS-2025). https://doi.org/10.31110/colins/2025-2/007
32. Lupei, M., Mitsa, O., Sharkan, V., Vargha, S., Gorbachuk, V. (2022). The Identification of Mass Media by Text Based on the Analysis of Vocabulary Peculiarities Using Support Vector Machines. 2022 International Conference on Smart Information Systems and Technologies (SIST), 1–6. https://doi.org/10.1109/sist54437.2022.9945774