STQ-UA: A dataset of synthetic and translated search queries for the Ukrainian language

Danylo Boiko[1,*,†], Nazar Kohut[2,†], Viktoriia Mishkurova[3,†] and Oleh Basystiuk[2,†]

[1]Innoloft Inc., 701 Brazos Street, Austin, TX 78701, USA
[2]Lviv Polytechnic National University, Stepana Bandery Street, 12, Lviv, 79013, Ukraine
[3]Bogomolets National Medical University, Beresteiskyi Avenue, 34, Kyiv, 03057, Ukraine

Abstract

This paper introduces a novel dataset of 100,000 search queries specifically compiled for the Ukrainian language. Given the scarcity of such resources, the dataset was created using a dual approach: synthetic generation and machine translation. To generate authentic-sounding queries, we used zero-shot and three-shot prompting techniques with eight distinct state-of-the-art closed-source large language models (LLMs) from five leading providers: OpenAI, Google, Cohere, Anthropic, and Mistral AI. These providers have headquarters in the USA, Canada, and France, which are located on two continents, thereby adding a layer of geographical and potentially cultural diversity to the dataset. To accurately reflect realistic search intent and phrasing, we also used the same suite of models to translate a substantial set of anonymized real-world English search queries taken from two major search engines: Google and Bing. The resulting dataset provides a high-quality resource essential for training, evaluating, and fine-tuning models in a wide range of tasks, including information retrieval, query understanding, relevance ranking, and related search challenges within the Ukrainian context.

Keywords

Search queries, synthetic generation, machine translation, large language models

1. Introduction

The efficiency of modern search engines and information retrieval systems heavily depends on their ability to accurately understand and process user queries. To achieve this, advanced algorithms analyze linguistic patterns and semantic structures to capture the essence. Training, evaluating, and fine-tuning the underlying models require extensive, high-quality datasets that reflect real-world search behaviors. These resources allow models to learn the correlation between query intent and relevant content, ensuring that results meet user expectations.

Nowadays, models for widely spoken languages like English demonstrate the best performance and dominate on the global stage, while many other languages face a significant data gap, which leads to a spread of low-quality models [1]. In particular, the Ukrainian language has historically faced a longstanding scarcity of resources, including educational materials, linguistic research, digital tools, and cultural initiatives.

Unfortunately, all generative models, including LLMs, have their biases due to the data they are trained on, which can lead to outputs that are systematically prejudiced, unfair, or skewed against certain groups or viewpoints [2]. A strategy that emphasizes complementary diversity of models can address this fundamental problem and help us achieve more balanced and less biased outcomes in the dataset.

We carefully selected a suite of eight state-of-the-art models from five leading providers, including OpenAI's GPT-4o and GPT-4o Mini, Google's Gemini 1.5 Flash and Gemini 2.0 Flash, Cohere's Command

WDA'26: International Workshop on Data Analytics, January 26, 2026, Kyiv, Ukraine
*Corresponding author.
†These authors contributed equally.
danielboyko02@gmail.com (D. Boiko); nazar.kohut.mknssh.2024@lpnu.ua (N. Kohut); michkourovaviktoire@gmail.com (V. Mishkurova); oleh.a.basystiuk@lpnu.ua (O. Basystiuk)
0009-0005-6341-0095 (D. Boiko); 0009-0003-0529-7210 (N. Kohut); 0009-0004-9304-0825 (V. Mishkurova); 0000-0003-0064-6584 (O. Basystiuk)

A and Command R+, Anthropic's Claude 3.5 Haiku, and Mistral AI's Mistral Large. These providers have headquarters located in three countries (USA, Canada, and France) spread across two continents (North America and Europe), adding a layer of geographical and potentially cultural diversity to the dataset. In addition to improving robustness, it enables a thorough investigation of how various models interpret and react to different prompts, which eventually promotes a deeper comprehension of behavior in different cultural contexts [3].

2. Related work

Historically, large-scale search query datasets have relied either on logs released by major search engine providers or on data collected through specialized academic or commercial efforts. Here are the most notable and widely used English datasets, representing a wealth of real-world queries and search interactions collected from well-known search engines:

- MS MARCO [4], a large-scale dataset designed by Microsoft for machine reading comprehension and information retrieval tasks. It comprises 1,010,916 anonymous questions extracted directly from Bing's search logs, offering a valuable collection of concise, real-world, natural language queries.
- Natural Questions [5], a question-answering dataset developed by Google Research, consists of real, anonymized, and aggregated queries issued to the Google search engine and corresponding answers. The public release includes 307,373 training samples with single annotations, 7,830 samples with 5-way annotations for development data, and a further 7,842 samples with 5-way annotations as test data.
- MIMICS [6], a collection of search clarification datasets created by Microsoft for research on conversational information seeking systems. It was built from real search queries sampled from Bing's query logs, where each data sample includes a clarifying question and up to five candidate answers intended to refine the original query. The total collection includes 3 datasets, comprising more than 450,000 unique queries.

To meet diverse needs, there are a few datasets available for a range of Ukrainian natural language processing (NLP) tasks. For example, large corpora such as UberText 2.0 [7] and CC-100 [8] derived from web crawls serve as the basis for pre-training LLMs. The BRUK corpus [9] offers genre-balanced samples that can be used for more structured linguistic analysis or model training on different text styles. Furthermore, there are enough datasets for less common tasks: Djinni Recruitment [10] focuses on IT recruitment, UA-GEC [11] provides annotated text for grammatical error correction, ParaRook||DE-UK [12] serves as a parallel German-Ukrainian corpus for machine translation, etc.

In turn, the landscape of publicly available search query datasets for the Ukrainian language is significantly more limited than for other Ukrainian NLP tasks and pales in comparison to the millions of real-world queries available for English. With partial success, we can only use questions from the UA-SQuAD dataset [13], which is a translation of part of the original SQuAD 2.0 [14] and consists of 13,859 samples.

3. Synthetic generation

Synthetic data generation [15, 16] is a widely used approach for creating artificial data that mimics the statistical properties and patterns of real-world resources. This technique is especially valuable in our case because getting real data is impossible without access to search providers.

To control the randomness and diversity of the content produced by LLMs, it is crucial to use the temperature and top-p parameters [17]. The temperature parameter affects the probability distribution of the model's predictions. It essentially controls how "creative", or "conservative" the model's outputs will be. Top-p sampling, also known as nucleus sampling, limits the selection to a subset of words whose cumulative probability exceeds a given threshold instead of selecting from the entire vocabulary.

Balancing these settings effectively allows for a tailored interaction with models, whether for creative writing or providing informative content.

To generate a subset of synthetic search queries, we used a temperature of 0.85 and nucleus sampling of 0.8 to balance creativity and variance, while kernel sampling was used to maintain relevance and consistency. Beyond these direct parameter adjustments, we also indirectly influenced the models using both zero-shot and three-shot techniques. This allowed us to explore different hint strategies to control the characteristics of the generated queries.

We used zero-shot prompting, which involves giving the model common generation instruction without providing any examples, to directly create 25,000 search queries. This approach allowed us to consistently guide the generation process based solely on the knowledge embedded in the parameters of the models.

By providing a few proper examples, models can better understand the desired style of content, resulting in more accurate and contextually relevant synthetic data. To generate another batch of 25,000 queries using the three-shot prompting, we combined the common generation instruction with three examples in the Ukrainian language.

4. Neural machine translation

A valuable alternative to the synthetic generation described earlier is machine translation using LLMs [18], which, being trained on massive corpora, can produce remarkably fluent and contextually appropriate outcomes across a wide range of topics and query styles.

To create a subset with machine-translated queries, we used anonymized real-world samples from two major search engines (Google and Bing). The previously described English datasets, Natural Questions and MIMICS, served as the data source.

For machine translation, it is appropriate to use low values for parameters responsible for the randomness during model configuration. A temperature of 0 reduces variability, enforcing determinism by sequentially selecting the most probable tokens. Meanwhile, a nucleus sampling of 0.05 restricts the set of tokens to the most confident predictions, balancing the accuracy and fluency.

We provided original queries to the models in batches of size 10 and overrode the system prompt. For some samples, the models produced incorrectly formatted outputs [19]. Comparing the number of failed batches to the total number determines the failure rate for each model (Table 1).

Table 1
Failure Rate of Models Across Search Engines (%)

	Google	**Bing**
GPT-4o	1.06	0
GPT-4o Mini	0.53	0
Gemini 1.5 Flash	0	0.25
Gemini 2.0 Flash	0.27	0.50
Command A	0	0
Command R+	2.85	0.74
Claude 3.5 Haiku	0.27	0.25
Mistral Large	6.48	6.10

These failure rates highlight the importance of selecting the appropriate models based on the specific task to unleash their potential. We translated 25,000 queries each from Google and Bing logs, distributing them evenly among models regardless of the obstacles confronting some of them.

During machine translation, some abbreviations, names, digits, etc., may retain their original spelling in English. To compare the content similarity between the translated and source queries, we used the Ratcliff-Obershelp algorithm and computed a score ranging from 0.0 to 1.0 for each pair. Presenting the average value and standard deviation helps maintain simplicity in displaying the outcomes (Table 2).

Table 2
Average Similarity Scores and Standard Deviations of the Translated and Source Queries

	Google		Bing	
	Average	SD	Average	SD
GPT-4o	0.10	0.18	0.12	0.19
GPT-4o Mini	0.06	0.14	0.10	0.18
Gemini 1.5 Flash	0.06	0.14	0.10	0.18
Gemini 2.0 Flash	0.09	0.18	0.11	0.22
Command A	0.05	0.14	0.10	0.18
Command R+	0.07	0.16	0.11	0.21
Claude 3.5 Haiku	0.07	0.15	0.11	0.18
Mistral Large	0.07	0.16	0.09	0.18

The Ratcliff-Obershelp algorithm compares two strings by finding the largest common substrings between them. It recursively identifies the largest common fragment in two strings and then repeats this process for the remaining strings to the left and right. The similarity score reflects how alike the two strings are in terms of content and overall structure.

All models demonstrated relatively low average similarity scores between the translated and original queries. Given the differences in language structures, it is not surprising that the Ukrainian queries significantly differ from the English ones. The standard deviations are quite small, which indicates the consistency of the dataset.

At first glance, it may seem that datasets based on search engine logs have no disadvantages, but not everything is so unequivocally. Bias exists everywhere, and search engines are no exception. If the claim that the query “download .net 8” is more likely to appear in Bing than in Google logs may be open to debate, the fact that the query “how to upload images on google drive” is more expected to be found in Google than in Bing logs is impossible to dispute.

It is important to notice that Bing queries have slightly higher average similarity scores and standard deviations compared to Google. This bias can be explained by the specific nature of the queries that users enter in different search engines.

5. Data overview

The final dataset comprises 100,000 real-world-like queries, evenly split between machine-translated and synthetically generated samples. The translated queries are divided equally between Bing and Google subgroups. Similarly, the generated queries are split based on zero-shot and three-shot techniques. Each of these four subgroups is further divided into eight parts based on utilized models, resulting in 32 subsegments of 3,125 queries each.

This well-organized composition enables in-depth analysis according to source, generation method, or particular model performance. To support this, each query is accompanied by relevant metadata and follows a consistent schema (Table 3), which outlines the fields provided for each sample.

Considering the prevalence of semantic text processing in today’s digital world [20], we focused on queries that provide enough context. That is why STQ-UA doesn’t include queries containing less than 3 words, as they are too short and lack the necessary semantic information (Figure 1).

The overall trend shows a positive correlation between the number of words and characters. The graph’s lower part concentrates the majority of points, suggesting that most queries range from 3 to 14 words and 7 to 80 characters. However, there are several points in the upper right corner, which indicate the presence of extreme cases with long queries.

To reflect the semantic diversity of the dataset, it is appropriate to use clustering. We used the LaBSE model [21] to construct high-dimensional vector representations of search queries. The HDBSCAN algorithm [22] combined these embeddings, identified clusters of varying density, and separated noise.

Table 3
Structured Dataset Fields with Descriptions

Field	Description
`query`	The unique search query obtained from the `model_name`.
`model_provider`	The provider responsible for the `model_name` (*openai*, *google*, *cohere*, *anthropic* or *mistral*).
`model_name`	The model used to obtain the `query` (*gpt-4o*, *gpt-4o-mini*, *gemini-1.5-flash*, *gemini-2.0-flash*, *command-a*, *command-r-plus*, *claude-3.5-haiku* or *mistral-large*).
`approach`	The approach used to obtain the `query` (*zero-shot*, *three-shot* or *translation*).
`search_engine`	The search engine from which the `source_query` was taken. If the `approach` is *translation*, this field will contain either *google* or *bing*. In all other cases, this field will be empty.
`source_query`	The real-world user query taken from `search_engine` logs.

For visualization, we used the UMAP method [23], which projected vectors into a two-dimensional space while preserving their structure (Figure 2).

Two-dimensional projection demonstrates a high fragmentation of the feature space with hundreds of compact clusters located unevenly and with varying densities. In the center of the space, there are areas with an increased concentration of points corresponding to the most frequent types of queries, while the peripheral areas are represented by small groups and isolated points, potentially anomalous or rare. Elongated and curved structures reveal gradual transitions between semantically related groups.

Since we retain both source and translated queries, the dataset could be valuable for NLP tasks that utilize data in multiple languages. For example, it can be applied to knowledge distillation [24], where knowledge from a more complex model is transferred to a smaller one [25], as well as adapting

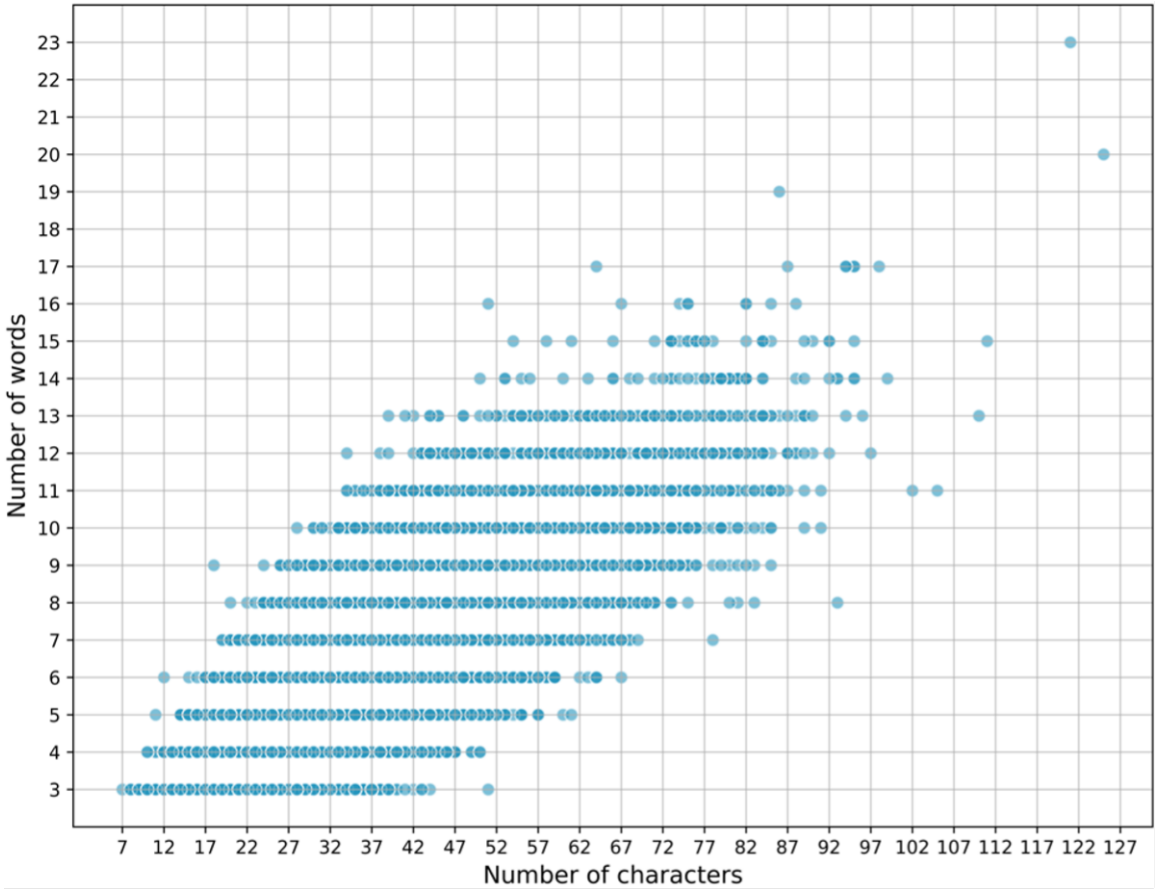

Figure 1: Correlation between the number of words and characters in search queries.

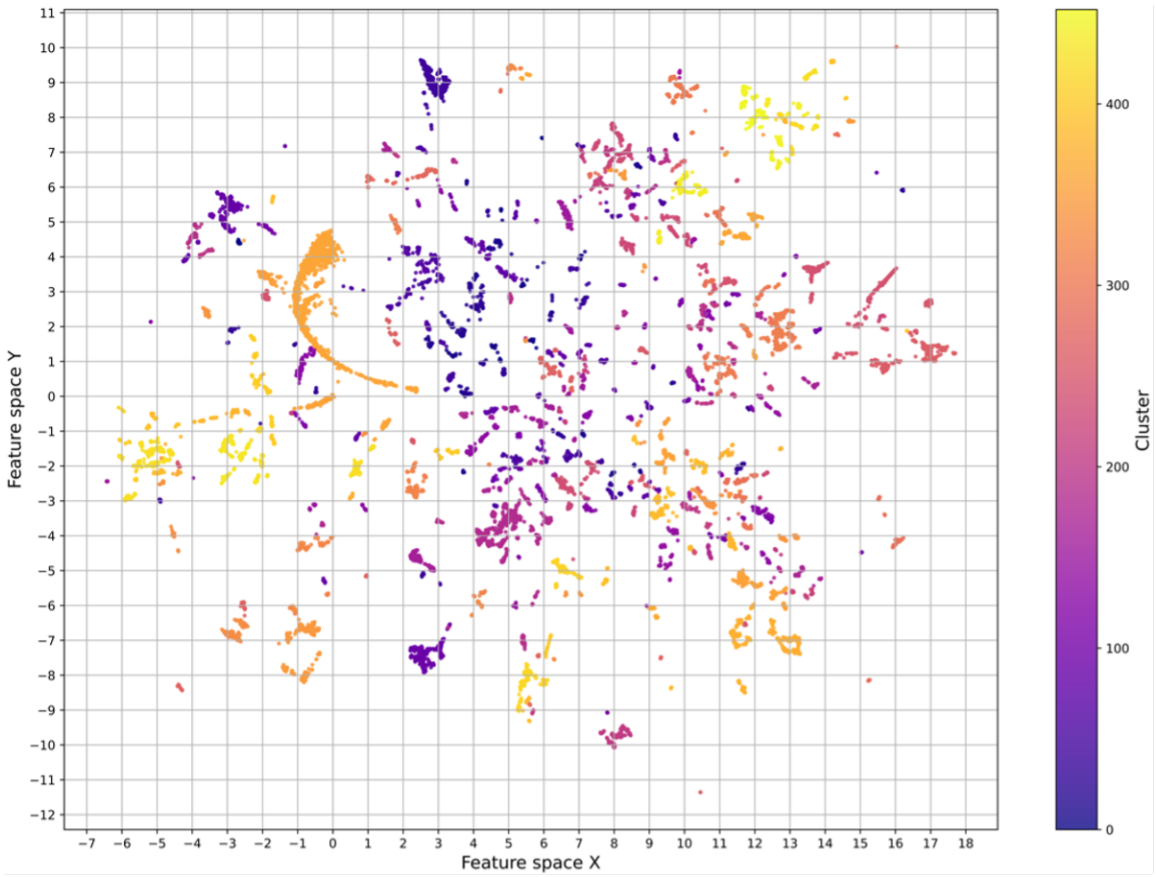

Figure 2: Clustered representation of search queries.

monolingual models for multilingual capabilities [26, 27].

6. Practical application

One of the useful applications of the final dataset is training a model for autocorrection in search queries, focusing on the most common types of errors, such as typos. At the initial stage of this field, systems mostly relied on rule-based approaches. Later, they were gradually replaced by statistical methods, which analyzed large corpora of texts to learn the probabilities of word sequences (for example, using n-gram models). The modern approaches use machine learning, in particular deep learning established on the sequence-to-sequence architecture [28].

In April 2024, Grammarly introduced the spivavtor-large [29], a model for the Ukrainian language based on the mt0-large multilingual transformer [30] with approximately 1.2 billion parameters, designed for efficient text editing and solving complex linguistic tasks. However, despite its advantages, the model demonstrates limited performance for typo correction in search queries, particularly when handling short, highly informal user inputs, which emphasizes the importance of fine-tuning on the STQ-UA to better capture domain-specific patterns.

Using a script for typo generation, we created a dataset with search queries containing one of three predefined errors: adding an extra letter, replacing one with another, or omitting one. Based on the original search queries and their versions with synthetically generated typos, spivavtor-large was fine-tuned on an NVIDIA A100 GPU (Figure 3). The main configuration parameters of the pipeline were a learning rate of 5e-5, a batch size of 8, and 5 training epochs, with the sequence length limited to 128 tokens.

The training loss indicates that the model is learning and improving its fit to the data. However, the validation loss reveals overfitting after 3 epochs, which may impair the performance on new, unseen

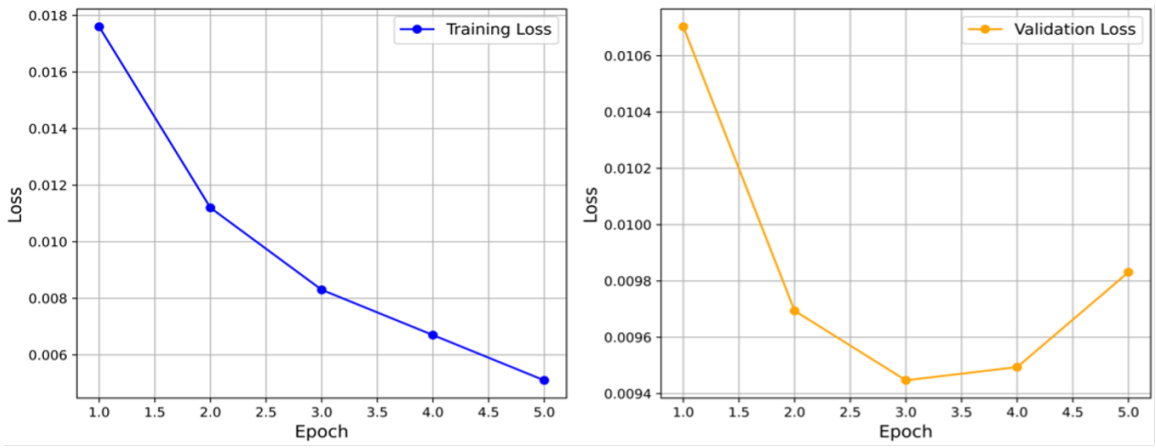

Figure 3: Cross-entropy loss on the training and validation subsets.

data. For models like spivavtor-large, effective evaluation involves using metrics such as precision, recall, and the combined $F\beta$ score. In the context of typo correction, it is reasonable to use β at 0.5, as the value allows us to consider the balance between identifying relevant and avoiding irrelevant predictions. We compared the evaluation metrics of the baseline model with variations after 3 and 5 epochs of fine-tuning (Table 4).

Table 4
Performance of the Base and Fine-Tuned Models

Model	Precision	Recall	$\mathbf{F_{0.5}}$
spivavtor-large	0.5936	0.6029	0.5944
spivavtor-large-stq-3rd-epoch	0.9149	0.9157	0.9149
spivavtor-large-stq-5th-epoch	0.9085	0.9073	0.9082

Comparison of the base and fine-tuned models reveals a significant improvement in efficiency. In particular, the best results were achieved after 3 epochs, which was expected based on the analysis of the validation loss. This suggests that the fine-tuning process effectively captured the task-specific patterns.

Using high-quality datasets is the key factor that enables faster adaptation and reduces overall computational requirements. Such optimization is especially important because modern models require significant computing resources, including sufficient memory and powerful GPUs, not only for training but also for inference.

7. Conclusions

This paper presents STQ-UA, a new large-scale dataset comprising 100,000 search queries for the Ukrainian language. Following the significant lack of such publicly available resources, we applied a dual strategy combining synthetic generation and machine translation to maintain linguistic diversity and consistency.

To ensure less biased outcomes, we used a diverse set of eight state-of-the-art LLMs incorporating varying architectures from five leading providers (OpenAI, Google, Cohere, Anthropic, and Mistral AI) headquartered in three countries (USA, Canada, and France) and spread across two continents (North America and Europe).

We applied both zero-shot and three-shot prompting techniques for synthetic generation, producing 50,000 queries that were intended to mimic real-world user search intent. To ensure the inclusion of authentic search patterns, we translated 50,000 real-world English search queries taken from Google

and Bing logs.

The analysis involved evaluating the performance of the models during translation, revealing varying failure rates. We also assessed the content similarity between translated and source queries using the Ratcliff-Obershelp algorithm, finding generally low average scores, indicating a significant transformation while retaining some original elements such as abbreviations and digits.

The resulting dataset was manually verified and offers a previously scarce resource to the Ukrainian NLP community, making another step toward bridging the global data gap for under-resourced languages. It can be used for training, evaluating, and fine-tuning models for various search-related tasks, including information retrieval, query autocompletion, and relevance ranking. Future work would involve building a dataset with a larger number of used models and search engines, as well as attempting to find real-world Ukrainian search queries in common crawls.

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] H. H. Nigatu, A. L. Tonja, B. Rosman, T. Solorio, M. Choudhury, The zeno's paradox of 'low-resource' languages, in: Y. Al-Onaizan, M. Bansal, Y.-N. Chen (Eds.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 17753–17774. doi:10.18653/v1/2024.emnlp-main.983.

[2] I. O. Gallegos, R. A. Rossi, J. Barrow, M. M. Tanjim, S. Kim, F. Dernoncourt, T. Yu, R. Zhang, N. K. Ahmed, Bias and fairness in large language models: A survey, Computational Linguistics 50 (2024) 1097–1179. doi:10.1162/coli_a_00524.

[3] M. F. Adilazuarda, S. Mukherjee, P. Lavania, S. S. Singh, A. F. Aji, J. O'Neill, A. Modi, M. Choudhury, Towards measuring and modeling "culture" in LLMs: A survey, in: Y. Al-Onaizan, M. Bansal, Y.-N. Chen (Eds.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 15763–15784. doi:10.18653/v1/2024.emnlp-main.882.

[4] P. Bajaj, D. Campos, N. Craswell, L. Deng, J. Gao, X. Liu, R. Majumder, A. McNamara, B. Mitra, T. Nguyen, M. Rosenberg, X. Song, A. Stoica, S. Tiwary, T. Wang, MS MARCO: A human generated machine reading comprehension dataset, 2018. URL: https://arxiv.org/abs/1611.09268.

[5] T. Kwiatkowski, J. Palomaki, O. Redfield, M. Collins, A. Parikh, C. Alberti, D. Epstein, I. Polosukhin, J. Devlin, K. Lee, K. Toutanova, L. Jones, M. Kelcey, M.-W. Chang, A. M. Dai, J. Uszkoreit, Q. Le, S. Petrov, Natural Questions: A benchmark for question answering research, Transactions of the Association for Computational Linguistics 7 (2019) 452–466. doi:10.1162/tacl_a_00276.

[6] H. Zamani, G. Lueck, E. Chen, R. Quispe, F. Luu, N. Craswell, MIMICS: A large-scale data collection for search clarification, 2020. URL: https://arxiv.org/abs/2006.10174.

[7] D. Chaplynskyi, Introducing UberText 2.0: A corpus of Modern Ukrainian at scale, in: M. Romanyshyn (Ed.), Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP), Association for Computational Linguistics, Dubrovnik, Croatia, 2023, pp. 1–10. doi:10.18653/v1/2023.unlp-1.1.

[8] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, in: D. Jurafsky, J. Chai, N. Schluter, J. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2020, pp. 8440–8451. doi:10.18653/v1/2020.acl-main.747.

[9] V. Starko, A. Rysin, Creating a POS gold standard corpus of Modern Ukrainian, in: M. Romanyshyn (Ed.), Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP),

Association for Computational Linguistics, Dubrovnik, Croatia, 2023, pp. 91–95. doi:10.18653/v1/2023.unlp-1.11.

[10] N. Drushchak, M. Romanyshyn, Introducing the djinni recruitment dataset: A corpus of anonymized CVs and job postings, in: M. Romanyshyn, N. Romanyshyn, A. Hlybovets, O. Ignatenko (Eds.), Proceedings of the Third Ukrainian Natural Language Processing Workshop (UNLP) @ LREC-COLING 2024, ELRA and ICCL, Torino, Italia, 2024, pp. 8–13. URL: https://aclanthology.org/2024.unlp-1.2.

[11] O. Syvokon, O. Nahorna, P. Kuchmiichuk, N. Osidach, UA-GEC: Grammatical error correction and fluency corpus for the Ukrainian language, in: M. Romanyshyn (Ed.), Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP), Association for Computational Linguistics, Dubrovnik, Croatia, 2023, pp. 96–102. doi:10.18653/v1/2023.unlp-1.12.

[12] M. Shvedova, A. Lukashevskyi, Creating parallel corpora for Ukrainian: A German-Ukrainian parallel corpus (ParaRook||DE-UK), in: M. Romanyshyn, N. Romanyshyn, A. Hlybovets, O. Ignatenko (Eds.), Proceedings of the Third Ukrainian Natural Language Processing Workshop (UNLP) @ LREC-COLING 2024, ELRA and ICCL, Torino, Italia, 2024, pp. 14–22. URL: https://aclanthology.org/2024.unlp-1.3.

[13] FIdo AI, UA-SQuAD, 2022. URL: https://huggingface.co/datasets/FIdo-AI/ua-squad.

[14] P. Rajpurkar, R. Jia, P. Liang, Know what you don't know: Unanswerable questions for SQuAD, in: I. Gurevych, Y. Miyao (Eds.), Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 784–789. doi:10.18653/v1/P18-2124.

[15] A. Bauer, S. Trapp, M. Stenger, R. Leppich, S. Kounev, M. Leznik, K. Chard, I. Foster, Comprehensive exploration of synthetic data generation: A survey, 2024. URL: https://arxiv.org/abs/2401.02524.

[16] Y. Lu, L. Chen, Y. Zhang, M. Shen, H. Wang, X. Wang, C. van Rechem, T. Fu, W. Wei, Machine learning for synthetic data generation: A review, 2025. URL: https://arxiv.org/abs/2302.04062.

[17] M. Renze, The effect of sampling temperature on problem solving in large language models, in: Y. Al-Onaizan, M. Bansal, Y.-N. Chen (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2024, Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 7346–7356. doi:10.18653/v1/2024.findings-emnlp.432.

[18] W. Zhu, H. Liu, Q. Dong, J. Xu, S. Huang, L. Kong, J. Chen, L. Li, Multilingual machine translation with large language models: Empirical results and analysis, in: K. Duh, H. Gomez, S. Bethard (Eds.), Findings of the Association for Computational Linguistics: NAACL 2024, Association for Computational Linguistics, Mexico City, Mexico, 2024, pp. 2765–2781. doi:10.18653/v1/2024.findings-naacl.176.

[19] D. X. Long, N.-H. Nguyen, T. Sim, H. Dao, S. Joty, K. Kawaguchi, N. F. Chen, M.-Y. Kan, LLMs are biased towards output formats! systematically evaluating and mitigating output format bias of LLMs, in: L. Chiruzzo, A. Ritter, L. Wang (Eds.), Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), Association for Computational Linguistics, Albuquerque, New Mexico, 2025, pp. 299–330. doi:10.18653/v1/2025.naacl-long.15.

[20] D. Maulud, S. Zeebaree, K. Jacksi, M. M.Sadeeq, K. Hussein, State of art for semantic analysis of natural language processing, Qubahan Academic Journal 1 (2021) 21–28. doi:10.48161/qaj.v1n2a40.

[21] W. Wang, G. Chen, H. Wang, Y. Han, Y. Chen, Multilingual sentence transformer as a multilingual word aligner, in: Y. Goldberg, Z. Kozareva, Y. Zhang (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2022, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022, pp. 2952–2963. doi:10.18653/v1/2022.findings-emnlp.215.

[22] C. Malzer, M. Baum, A hybrid approach to hierarchical density-based cluster selection, in: 2020 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI), 2020, pp. 223–228. doi:10.1109/MFI49285.2020.9235263.

[23] B. Ghojogh, A. Ghodsi, F. Karray, M. Crowley, Uniform manifold approximation and projection (umap) and its variants: Tutorial and survey, 2021. URL: https://arxiv.org/abs/2109.02508.

[24] S. Hahn, H. Choi, Self-knowledge distillation in natural language processing, in: R. Mitkov, G. Angelova (Eds.), Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019), INCOMA Ltd., Varna, Bulgaria, 2019, pp. 423–430. doi:10.26615/978-954-452-056-4_050.

[25] P. Liu, X. Wang, L. Wang, W. Ye, X. Xi, S. Zhang, Distilling knowledge from bert into simple fully connected neural networks for efficient vertical retrieval, in: Proceedings of the 30th ACM International Conference on Information & Knowledge Management, CIKM '21, Association for Computing Machinery, New York, NY, USA, 2021, p. 3965–3975. doi:10.1145/3459637.3481909.

[26] I. Yurchuk, D. Boiko, Extending monolingual asymmetric semantic search models for multilingual query processing using knowledge distillation, in: V. Snytyuk, V. Morozov, I. Javorskyj, V. G. Levashenko (Eds.), Proceedings of the Information Technology and Implementation (IT&I) Workshop: Intelligent Systems and Security (IT&I-WS 2024: ISS), Kyiv, Ukraine, November 20 - 21, 2024, volume 3933 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2024, pp. 1–10. URL: https://ceur-ws.org/Vol-3933/Paper_1.pdf.

[27] N. Reimers, I. Gurevych, Making monolingual sentence embeddings multilingual using knowledge distillation, in: B. Webber, T. Cohn, Y. He, Y. Liu (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, 2020, pp. 4512–4525. doi:10.18653/v1/2020.emnlp-main.365.

[28] T. Ge, X. Zhang, F. Wei, M. Zhou, Automatic grammatical error correction for sequence-to-sequence text generation: An empirical study, in: A. Korhonen, D. Traum, L. Màrquez (Eds.), Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 6059–6064. doi:10.18653/v1/P19-1609.

[29] A. Saini, A. Chernodub, V. Raheja, V. Kulkarni, Spivavtor: An instruction tuned Ukrainian text editing model, in: M. Romanyshyn, N. Romanyshyn, A. Hlybovets, O. Ignatenko (Eds.), Proceedings of the Third Ukrainian Natural Language Processing Workshop (UNLP) @ LREC-COLING 2024, ELRA and ICCL, Torino, Italia, 2024, pp. 95–108. URL: https://aclanthology.org/2024.unlp-1.12.

[30] N. Muennighoff, T. Wang, L. Sutawika, A. Roberts, S. Biderman, T. Le Scao, M. S. Bari, S. Shen, Z. X. Yong, H. Schoelkopf, X. Tang, D. Radev, A. F. Aji, K. Almubarak, S. Albanie, Z. Alyafeai, A. Webson, E. Raff, C. Raffel, Crosslingual generalization through multitask finetuning, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 15991–16111. doi:10.18653/v1/2023.acl-long.891.