UDC 004.94
DOI 10.56525/NMWD4742

ADAPTATION OF BIG DATA TO LOCAL INFORMATION LANGUAGE MODELS: DEVELOPMENT OF THE BIGTOR CHATBOT SYSTEM

F.R. Adgozalov
Azerbaijan State Oil and Industry University, Baku, Azerbaijan
e-mail: feridadgozelov.0@gmail.com

Abstract: This paper presents the development of BigTor, a domain-specific chatbot system designed to address cultural, administrative, and informational gaps in Azerbaijan through the adaptation of large language models to localized contexts. Existing global AI systems often demonstrate limited effectiveness when handling low-resource languages such as Azerbaijani, particularly in areas requiring cultural understanding, legal terminology, and context-sensitive reasoning. To overcome these limitations, the DeepSeek-R1-Distill-Llama-8B model was selected as the foundational architecture due to its balance between computational efficiency and advanced reasoning capabilities. The model was fine-tuned using Parameter-Efficient Fine-Tuning (PEFT) techniques and Low-Rank Adaptation (LoRA), combined with 4-bit quantization and bfloat16 precision to optimize hardware usage and reduce memory consumption.

A large synthetic dataset based on the Alpaca instruction-tuning format was generated through the AICTIA AI Studio platform, covering more than 50 thematic categories related to Azerbaijani culture, administration, history, politics, and public services. Experimental evaluation demonstrated that BigTorV1 significantly outperformed baseline models in localized tasks, achieving 92% accuracy in national music-related queries and 88% accuracy in historical knowledge tasks while maintaining low response latency. The findings confirm that synthetic data generation and efficient fine-tuning approaches can successfully enhance the performance of localized AI systems for low-resource languages. Furthermore, the project highlights the importance of culturally aware artificial intelligence in supporting digital sovereignty, preserving national heritage, and improving access to localized information services.

Keywords: Large Language Models, Fine-tuning, Synthetic Data, Specialized Chatbots, Cultural Preservation.

Introduction

Artificial Intelligence (AI) and Large Language Models (LLMs) have become central components of the modern technological ecosystem. However, globally deployed commercial models often fail to adequately address localized knowledge gaps in low-resource languages such as Azerbaijani, which possess distinct linguistic and cultural characteristics [14]. This limitation is particularly evident in domains such as legal terminology, national traditions, and public service systems.

The BigTor project aims to bridge this gap by introducing a specialized model enriched with localized data and tailored to the specific needs of Azerbaijan. As highlighted by Bafghi (2025), the fine-tuning process plays a critical role in adapting pre-trained models to domain-specific applications [1]. Furthermore, Weng (2024) emphasizes that Parameter-Efficient Fine-Tuning (PEFT) techniques enhance memory and energy efficiency, enabling faster and more scalable model adaptation [14].

By leveraging these methodologies, BigTor represents a strategic AI initiative designed to support Azerbaijan's digital sovereignty while preserving and promoting its cultural heritage.

Research objective and problem statement

The primary objective of this study is to design and develop a localized large language model capable of delivering accurate, context-aware, and comprehensive information about Azerbaijan's

administrative, cultural, and social structures. The model is intended to address national-level information demands while ensuring alignment with local linguistic nuances, institutional frameworks, and socio-cultural dynamics. In this context, the research also aims to contribute to the development of AI systems that are not only technically robust but also culturally and contextually relevant.

A key challenge underlying this objective is the limited availability of high-quality, structured Azerbaijani-language training data. As a low-resource language, Azerbaijani lacks sufficiently large and diverse annotated datasets, particularly in specialized domains such as legal documentation, public administration, and cultural heritage. This data scarcity directly impacts model generalization, factual accuracy, and domain-specific performance, creating a critical bottleneck in developing effective localized AI solutions.

To address this issue, the study employs the AICTIA AI Studio platform to construct a scalable synthetic data generation pipeline based on the Alpaca instruction-tuning format [11]. This approach enables the systematic creation of high-quality instruction–response pairs that reflect real-world scenarios and user queries relevant to Azerbaijan. The generated synthetic data is carefully designed to preserve linguistic authenticity while incorporating domain-specific knowledge, thereby enhancing both the depth and breadth of the training corpus.

Moreover, synthetic data generation serves not only as a substitute for limited real-world data but also as a strategic mechanism for improving model robustness, consistency, and adaptability. By diversifying the training distribution and introducing controlled variations, the model becomes better equipped to handle a wide range of queries, including those related to governance, cultural practices, and everyday informational needs. Prior research indicates that such approaches significantly enhance model accuracy and task-specific performance when properly aligned with real-world contexts [4].

Overall, this research frames the development of a localized LLM as both a technical and strategic initiative, addressing data scarcity challenges while advancing the broader goal of building inclusive, efficient, and nationally relevant artificial intelligence systems.

Methods for problem solving and validation

Model Selection and Reasoning Mechanism:

As the core architecture of the BigTor system, the DeepSeek-R1-Distill-Llama-8B model was selected. With 8 billion parameters, this model provides an optimal trade-off between computational efficiency and language understanding capability. A key advantage of the distilled variant lies in its internal “thinking tag” mechanism, which enables structured, step-by-step reasoning prior to generating responses. This capability is particularly critical when addressing complex administrative and cultural queries, ensuring logical coherence and contextual accuracy within the Azerbaijani domain.

PEFT and LoRA Implementation:

The fine-tuning process was conducted using the Unsloth framework, applying the Low-Rank Adaptation (LoRA) method as part of a Parameter-Efficient Fine-Tuning (PEFT) strategy [7]. To further optimize resource utilization, 4-bit quantization and bfloat16 data types were employed. The hyperparameters were configured as follows:

- Rank (r): 16
- Alpha: 32
- Dropout: 0.05
- Optimizer: AdamW (learning rate: 2e-4)

This configuration resulted in approximately a 70% reduction in memory consumption compared to conventional fine-tuning approaches, with peak GPU usage reduced to around 14GB. Additionally, distributed training was implemented across three local servers, reducing synchronization latency to below 5% and improving training throughput by approximately 2.8× compared to a single-node setup. The entire training process was monitored using the Weights & Biases platform [13].

Synthetic Data Generation and Dataset Composition:

The synthetic dataset, generated via the AICTIA AI Studio platform, follows the Alpaca instruction-tuning format and spans over 50 thematic categories. To ensure domain depth and contextual relevance, several specialized categories were incorporated:

- General information about Azerbaijan: 570 entries
- RIIN (Real Estate Information Infrastructure): 183 entries
- Numerology and symbolic patterns: 107 entries
- Space industry: 84 entries
- Historical memory (January 20 martyrs): 66 entries
- Politics and international sanctions: 92 entries

This diversified dataset design enhances the model's ability to generalize across both formal institutional knowledge and culturally embedded concepts.

System Architecture

The overall system architecture follows a client–server paradigm and consists of four main layers:

1. Model Service Layer: The BigTorV1 model is deployed on GPU servers using Ollama [8].
2. API Layer: A RESTful API built with Python-based frameworks such as Flask and FastAPI.
3. Frontend Layer: An interactive user interface developed using Vue.js.
4. Database Layer: Integration of MySQL and PostgreSQL for managing user sessions and system logs.

Validation and Testing (Aprobation):

The model was evaluated through a combination of qualitative and quantitative validation procedures. Domain-specific test queries were designed to assess accuracy, contextual relevance, and reasoning consistency. Comparative analysis with baseline models demonstrated improved performance in Azerbaijani-language comprehension and domain-specific response generation. Additionally, user-based testing scenarios confirmed the system's practical applicability in real-world informational contexts, particularly in public services and culturally sensitive domains**.**

Application of the obtained results

Quantitative Analysis

The performance of the BigTorV1 model was evaluated through a comparative analysis against the benchmark model Mistral-7B-Instruct. The evaluation focused on key performance indicators, including accuracy, contextual relevance, response coherence, and domain-specific understanding within Azerbaijani-language queries.

The results indicate that BigTorV1 demonstrates a measurable improvement in handling localized and culturally contextualized tasks, particularly in areas such as administrative procedures, legal terminology, and nationally specific knowledge domains. This performance gain can be attributed to the integration of synthetic datasets and domain-focused fine-tuning strategies, which enhance the model's ability to generate precise and context-aware responses.

In contrast, the benchmark model, while strong in general-purpose language understanding, exhibits limitations when addressing queries that require deep localization or familiarity with Azerbaijan-specific structures and terminology. These findings highlight the effectiveness of the proposed approach in bridging the gap between global language models and localized information needs.

A detailed comparison of the two models is presented in Table 1, where the performance differences are clearly illustrated across multiple evaluation metrics.

Table 1. Comparison of BigTorV1 and Benchmark Model Mistral-7B-Instruct

Metrics	BigTorV1	Mistral-7B-Instruct
Cultural Prompt Score	5.0 / 5	3.0 / 5
National Music Accuracy	92%	45%

Historical Knowledge Accuracy	88%	51%
Avg. Response Naturalness	4.7 / 5	3.2 / 5
Avg. Response Time	1.2 sec	1.1 sec

The model demonstrates a strong understanding of the grammatical and idiomatic features of the Azerbaijani language, particularly its agglutinative morphological structure. It is capable of accurately interpreting and contextualizing culturally specific terms such as “elçilik,” “xınayaxdı,” and “yallı,” providing explanations that are consistent with their sociocultural and traditional meanings.

In addition, the model shows high precision in responding to queries related to Azerbaijani musical heritage, particularly Muğam. It correctly identifies the seven principal dastgahs—Rast, Şur, Segah, Çahargah, Bayatı-Şiraz, Şüştər, and Humayun—and provides appropriate descriptions of their associated emotional and expressive characteristics [15].

Empirical evaluation indicates that the model achieves 92% accuracy in national music-related tasks and 88% accuracy in historical knowledge domains, demonstrating strong domain adaptation capabilities. Furthermore, the system maintains an average response latency of 1.2 seconds, indicating efficient inference performance suitable for real-time applications.

Overall, these results confirm that the model effectively integrates linguistic competence with culturally grounded knowledge representation, particularly in music, history, and cultural heritage domains.

Conclusion

This project demonstrates that effective artificial intelligence systems for low-resource languages can be successfully developed through the combined use of synthetic data generation and Parameter-Efficient Fine-Tuning (PEFT) techniques. The results indicate that these approaches significantly reduce data dependency while maintaining high model performance in domain-specific tasks.

By the 40th training step, the model had already reached approximately 90% of its final performance level, which clearly validates the efficiency and convergence speed of the adopted methodology. This rapid performance gain highlights the effectiveness of the training strategy and the quality of the constructed dataset.

The results further show that BigTorV1, even when trained and deployed on limited hardware resources (14GB GPU), is capable of outperforming large-scale global models in tasks requiring localized understanding and contextual awareness. This demonstrates that computational efficiency and high task performance are not mutually exclusive when appropriate optimization techniques are applied.

Future development directions include extending the system’s capabilities toward multimodal learning, implementing containerized deployment using Kubernetes, and preparing the platform for full public release. These improvements aim to enhance scalability, robustness, and accessibility, further strengthening the system’s role in supporting localized AI applications.

REFERENCES

1. Bafghi, R. A., et al. (2025). Fine-tuning without catastrophic forgetting via selective low-rank adaptation.
2. Cubed. (2024). Evaluation strategies for fine-tuned chatbots.
3. Devlin, J., et al. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding.
4. Gartner. (2022). Is synthetic data the future of AI?
5. Hermansson, L. L., & Parvanian, S. (2022). Synthetic health data at a glance.
6. Howard, J., & Ruder, S. (2018). Universal language model fine-tuning for text classification.
7. Hu, E. J., et al. (2021). LoRA: Low-rank adaptation of large language models.

8. Ollama. (2025). Ollama Model Serving Platform.
9. Touvron, H., et al. (2023). LLaMA: Open and efficient foundation language models.
10. Vaswani, A., et al. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*.
11. Wang, L., et al. (2023). Synthetic data generation for large language models.
12. Wang, Y. Q., et al. (2024). Model-in-the-Loop (MILO): Accelerating Multimodal AI Data Annotation.
13. Weights & Biases. (2025). Weights & Biases Documentation.
14. Weng, B. (2024). Navigating the landscape of large language models.
15. Орынбасар, М., Жумадилова, М. Б., & Абдыкеримова, Э. А. (2024). Қазақстанның білім беру жүйесіндегі жасанды интеллект: талдау және перспективалар. Том №48, 48(3).

ҮЛКЕН ДЕРЕКТЕРДІ ЖЕРГІЛІКТІ АҚПАРАТТЫҚ ТІЛДІК МОДЕЛЬДЕРГЕ БЕЙІМДЕУ: BIGTOR ЧАТ-БОТ ЖҮЙЕСІН ӘЗІРЛЕУ

Адгозалов Фарид
Әзірбайжан мемлекеттік мұнай және өнеркәсіп университеті, Баку қ., Әзірбайжан
e-mail: feridadgozelov.0@gmail.com

Аңдатпа. Бұл мақалада BigTor атты мамандандырылған чат-бот жүйесінің әзірленуі ұсынылады. Жоба Әзербайжандағы мәдени, әкімшілік және ақпараттық олқылықтарды жою мақсатында ірі тілдік модельдерді жергілікті контекске бейімдеу арқылы жүзеге асырылған. Қолданыстағы жаһандық жасанды интеллект жүйелері әзербайжан тілі сияқты цифрлық ресурсы аз тілдермен жұмыс істеуде, әсіресе мәдени түсінік, құқықтық терминология және контекстке тәуелді пайымдау қажет болатын салаларда, шектеулі тиімділік көрсетеді. Осы мәселелерді шешу үшін есептеу тиімділігі мен кеңейтілген логикалық мүмкіндіктер арасындағы тепе-теңдікті қамтамасыз ететін DeepSeek-R1-Distill-Llama-8B моделі негізгі архитектура ретінде таңдалды. Модель Parameter-Efficient Fine-Tuning (PEFT) және Low-Rank Adaptation (LoRA) әдістері арқылы қайта оқытылып, аппараттық ресурстарды оңтайландыру және жад тұтынуын азайту мақсатында 4-биттік кванттау мен bfloat16 дәлдігі қолданылды.

AICTIA AI Studio платформасының көмегімен Alpaca instruction-tuning форматына негізделген ірі синтетикалық деректер жиынтығы жасалып, оған әзербайжан мәдениеті, мемлекеттік басқару, тарих, саясат және қоғамдық қызметтерге қатысты 50-ден астам тақырыптық санат енгізілді. Эксперименттік бағалау нәтижелері BigTorV1 моделінің локализацияланған тапсырмаларда базалық модельдерден айтарлықтай жоғары нәтиже көрсеткенін дәлелдеді: ұлттық музыкаға қатысты сұраныстарда 92% дәлдікке және тарихи білім тапсырмаларында 88% дәлдікке қол жеткізілді, сонымен қатар жауап беру уақыты төмен деңгейде сақталды. Зерттеу нәтижелері синтетикалық деректерді генерациялау мен тиімді қайта оқыту тәсілдері ресурсы шектеулі тілдерге арналған локализацияланған AI жүйелерінің өнімділігін айтарлықтай арттыра алатынын көрсетеді. Сонымен қатар, жоба мәдени ерекшеліктерді ескеретін жасанды интеллекттің цифрлық егемендікті қолдау, ұлттық мұраны сақтау және жергілікті ақпараттық қызметтерге қолжетімділікті жақсарту үшін маңыздылығын айқындайды.

Түйін сөздер: ірі тілдік модельдер, қайта оқыту, синтетикалық деректер, мамандандырылған чат-боттар, мәдени мұраны сақтау.

АДАПТАЦИЯ БОЛЬШИХ ДАННЫХ ДЛЯ ЛОКАЛЬНЫХ ИНФОРМАЦИОННЫХ ЯЗЫКОВЫХ МОДЕЛЕЙ: РАЗРАБОТКА ЧАТ-БОТ СИСТЕМЫ BIGTOR

Адгозалов Фарид
Азербайджанский государственный университет нефти и промышленности,
Баку, Азербайджан
e-mail: feridadgozelov.0@gmail.com

Абстракт. В данной статье представлена разработка BigTor — специализированной чат-бот системы, предназначенной для устранения культурных, административных и информационных пробелов в Азербайджане посредством адаптации больших языковых моделей к локализованным условиям. Существующие глобальные системы искусственного интеллекта часто демонстрируют ограниченную эффективность при работе с языками с низким уровнем цифровых ресурсов, такими как азербайджанский язык, особенно в задачах, требующих культурного понимания, юридической терминологии и контекстно-зависимого рассуждения. Для преодоления этих ограничений в качестве базовой архитектуры была выбрана модель DeepSeek-R1-Distill-Llama-8B благодаря балансу между вычислительной эффективностью и расширенными возможностями логического вывода. Модель была дообучена с использованием методов Parameter-Efficient Fine-Tuning (PEFT) и Low-Rank Adaptation (LoRA), а также 4-битной квантизации и точности bfloat16 для оптимизации использования аппаратных ресурсов и снижения потребления памяти.

Крупный синтетический датасет в формате Alpaca instruction-tuning был создан с помощью платформы AICTIA AI Studio и охватывал более 50 тематических категорий, связанных с азербайджанской культурой, государственным управлением, историей, политикой и государственными услугами. Экспериментальная оценка показала, что BigTorV1 значительно превосходит базовые модели в локализованных задачах, достигая 92% точности в запросах, связанных с национальной музыкой, и 88% точности в задачах исторических знаний при сохранении низкой задержки ответа. Полученные результаты подтверждают, что генерация синтетических данных и эффективные методы дообучения способны существенно повысить производительность локализованных AI-систем для языков с ограниченными ресурсами. Кроме того, проект подчеркивает важность культурно-ориентированного искусственного интеллекта для обеспечения цифрового суверенитета, сохранения национального наследия и улучшения доступа к локализованным информационным сервисам.

Ключевые слова: большие языковые модели, дообучение, синтетические данные, специализированные чат-боты, сохранение культурного наследия.