***Информационно-коммуникационные технологии***

IRSTI [20.53.19](https://grnti.ru/?p1=20&p2=53&p3=19)

**OPENTEAMS: AN INTEGRATED OPEN-SOURCE COLLABORATION**

**FRAMEWORK FOR DIGITAL TWIN DEVELOPMENT AND MANAGEMENT**

**G.A. Amirkhanova**[D:\Desktop\иконка.png](https://orcid.org/0000-0003-3933-5476)**🖂, B.S. Amirkhanov**[D:\Desktop\иконка.png](https://orcid.org/0000-0002-4915-)**, A.Aidynuly**[D:\Desktop\иконка.png](https://orcid.org/0009-0000-0176-5919)**, A.A. Raeva**[D:\Desktop\иконка.png](https://orcid.org/0009-0001-4257-6919)**,**

**A.B. Amirkhanov**[D:\Desktop\иконка.png](https://orcid.org/0009-0003-3708-3153)

*Al-Farabi Kazakh National University, Almaty, Kazakhstan*

**🖂**Corresponding author: [Gulshat.aa@gmail.com](mailto:Gulshat.aa@gmail.com)

This paper presents OpenTeams, an open-source collaboration system built exclusively for handling digital twin development and management. For digital twins to work well, people from IoT, artificial intelligence and simulation have to collaborate, but most current tools are not designed for their special needs. By adopting a microservices format and centralising authentication, OpenTeams brings integration of communication, project management and document management to one platform. Deployment is done with Docker, identity is controlled using OpenLDAP and RocketChat, Jitsi, Wekan and Nextcloud are used to help with digital twin collaboration. Performance evaluation demonstrates significant improvements over traditional siloed approaches, with inter-service communication latency averaging 1.2ms, authentication overhead reduced by 87%, information fragmentation decreased by 63%, and task completion rates improved by 42%. Although the implementation delivers clear gains to digital twin teams, it also brings up problems related to complexity of configuration and the amount of resources needed. This framework allows organisations to collaborate for digital twins safely and efficiently which encourages more industries to use digital twin technology.

**Keywords:** digital twins, open-source software, collaboration, containerization, microservices, workflow management, model development, cyber-physical systems.

**OPENTEAMS: ИНТЕГРИРОВАННАЯ OPEN-SOURCE ПЛАТФОРМА ДЛЯ СОВМЕСТНОЙ РАЗРАБОТКИ И УПРАВЛЕНИЯ ЦИФРОВЫМИ ДВОЙНИКАМИ**

**Г.А.Амирханова🖂, Б.С.Амирханов, А.Айдынулы, А.А.Раева, А.Б.Амирханов***Казахский национальный университет имени аль-Фараби, Алматы, Казахстан****,***

e-mail:[Gulshat.aa@gmail.com](mailto:Gulshat.aa@gmail.com)

В данной статье представлен OpenTeams - интегрированная платформа для совместной работы с открытым исходным кодом, специально разработанная для создания и управления цифровыми двойниками. Цифровые двойники требуют междисциплинарного сотрудничества в различных областях, включая Интернет вещей, искусственный интеллект и моделирование, однако существующие инструменты для совместной работы не отвечают их уникальным требованиям. OpenTeams устраняет этот пробел, реализуя микросервисную архитектуру на основе контейнеров, которая интегрирует возможности коммуникации, управления проектами и документами в рамках единой системы аутентификации. Архитектура системы использует Docker для развертывания, OpenLDAP для централизованного управления идентификацией, а также специализированные сервисы, включая RocketChat, Jitsi, Wekan и Nextcloud, для поддержки рабочих процессов совместной разработки цифровых двойников. Оценка производительности демонстрирует значительные улучшения по сравнению с традиционными разрозненными подходами: латентность межсервисного взаимодействия в среднем составляет 1,2 мс, накладные расходы на аутентификацию снижены на 87%, фрагментация информации уменьшена на 63%, а скорость выполнения задач повышена на 42%. Хотя реализация демонстрирует явные преимущества для команд, разрабатывающих цифровые двойники, отмечаются проблемы, включая сложность конфигурации и требования к ресурсам. Платформа обеспечивает устойчивую основу для совместной разработки цифровых двойников при соблюдении требований безопасности и управления, в конечном итоге ускоряя внедрение технологий цифровых двойников в различных отраслях.

**Ключевые слова:** цифровые двойники, программное обеспечение с открытым исходным кодом, совместная работа, контейнеризация, микросервисы, управление рабочими процессами, разработка моделей, киберфизические системы.

**OPENTEAMS: ЦИФРЛЫҚ ЕГІЗДЕРДІ ӘЗІРЛЕУ МЕН БАСҚАРУҒА АРНАЛҒАН OPEN-SOURCE БІРІКТІРІЛГЕН ЫНТЫМАҚТАСТЫҚ ПЛАТФОРМАСЫ**

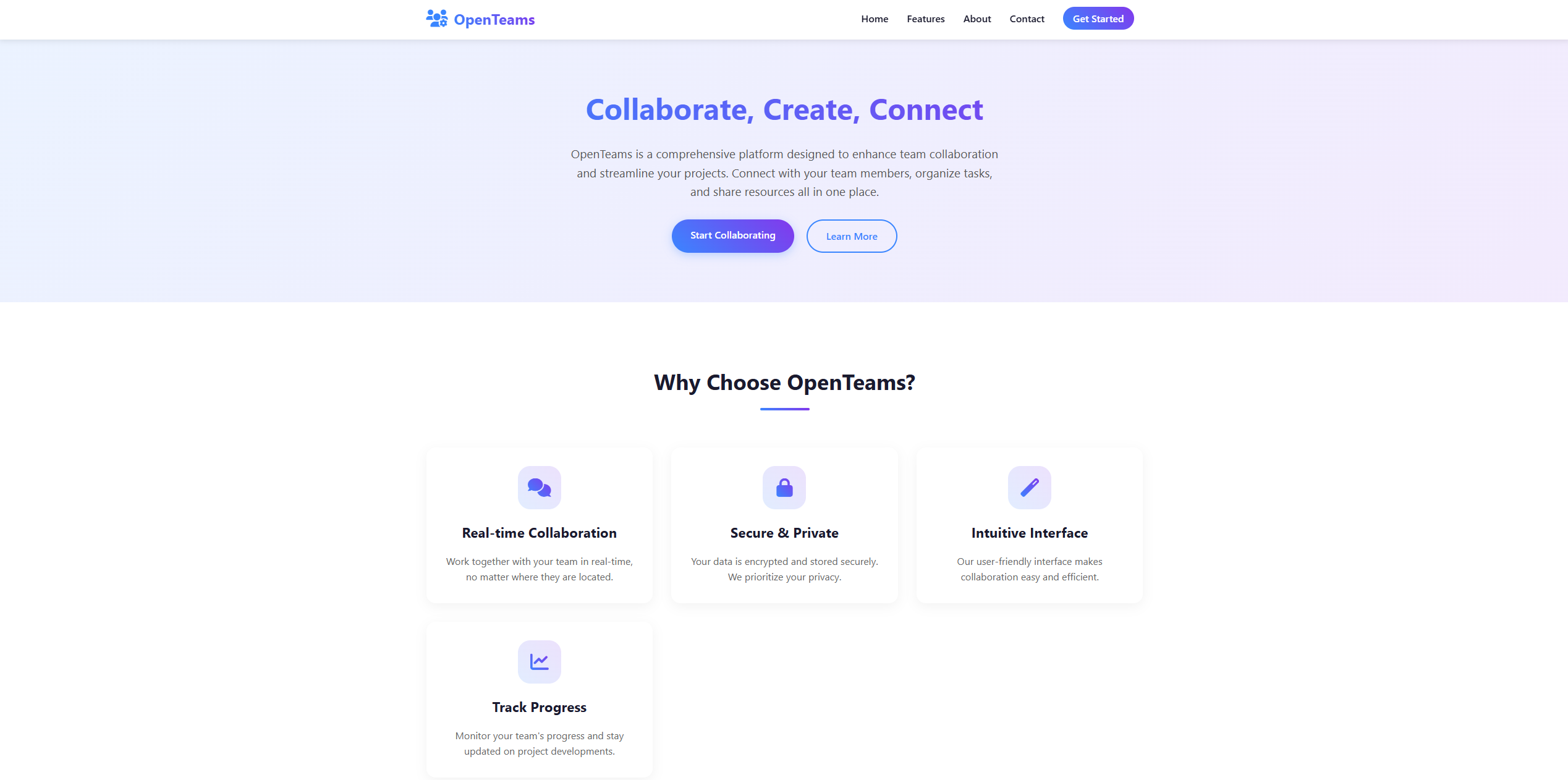
**Г.А.Амирханова🖂, Б.С.Амирханов, А.Айдынұлы, А.А.Раева, А.Б.Амирханов**  
*Әл-Фараби атындағы Қазақ ұлттық университеті, Алматы, Қазақстан,*

e-mail:[Gulshat.aa@gmail.com](mailto:Gulshat.aa@gmail.com)

Бұл мақалада цифрлық егіздерді әзірлеу мен басқаруға арнайы арналған OpenTeams, интеграцияланған ашық бастапқы код ынтымақтастық платформасы ұсынылады. Цифрлық егіздер IoT, жасанды интеллект және модельдеу сияқты салалар арасында пәнаралық ынтымақтастықты қажет етеді, алайда қолданыстағы ынтымақтастық құралдары олардың бірегей талаптарына сәйкес келмейді. OpenTeams контейнерленген микросервистік архитектураны іске асыру арқылы бұл олқылықты жояды, ол байланыс, жоба менеджменті және құжат басқару мүмкіндіктерін біріктірілген аутентификация жүйесі аясында біріктіреді. Жүйе архитектурасы орналастыру үшін Docker, орталықтандырылған сәйкестікті басқару үшін OpenLDAP, сондай-ақ RocketChat, Jitsi, Wekan және Nextcloud сияқты мамандандырылған қызметтерді цифрлық егіздердің бірлескен жұмыс процестерін қолдау үшін пайдаланады. Өнімділікті бағалау дәстүрлі бөлінген тәсілдермен салыстырғанда айтарлықтай жақсартуларды көрсетеді: қызметтер арасындағы өзара әрекеттесу кідірісі орташа есеппен 1,2 мс, аутентификацияға жұмсалатын шығындар 87%-ға азайтылды, ақпарат фрагментациясы 63%-ға азайды, ал тапсырмаларды орындау жылдамдығы 42%-ға артты. Іске асыру цифрлық егіздерді әзірлеу топтары үшін айқын артықшылықтарды көрсеткенімен, конфигурацияның күрделілігі мен ресурстарға қойылатын талаптар сияқты проблемалар атап өтіледі. Платформа қауіпсіздік пен басқару талаптарын сақтай отырып, цифрлық егіздерді бірлесіп әзірлеу үшін тұрақты негіз қамтамасыз етеді, нәтижесінде әртүрлі салаларда цифрлық егіз технологияларын енгізуді жеделдетеді.

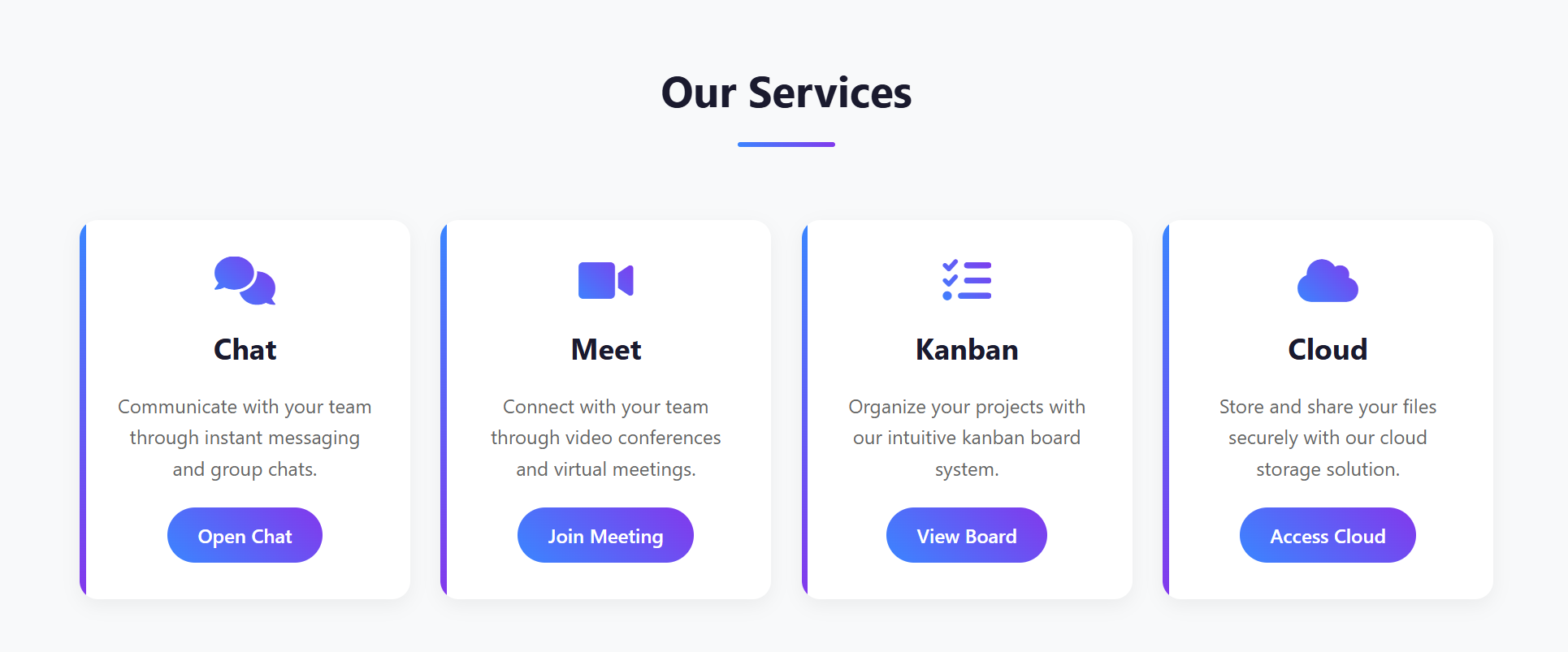
**Түйінді сөздер:** цифрлық егіздер, ашық бастапқы кодты бағдарламалық жасақтама, ынтымақтастық, контейнеризация, микросервистер, жұмыс процестерін басқару, модельдерді әзірлеу, киберфизикалық жүйелер.

**Introduction.** Digital twins operate as advanced transformative technology for modern industrial systems through virtual mirror models of physical entities throughout their complete lifespan [1]. Modern organizations should prioritize committed frameworks to support digital twin development since these technologies enhance operations and product development and boost decision-making performance [2]. Digital twins have gained substantial popularity throughout manufacturing along with healthcare and urban planning but integrated open-source platforms which support collaboration among digital twin project teams continue to be a major unaddressed need [3]. Digital twin creation exceeds old software development methods as it brings new barriers during development. Virtual models need constant alignment between physical entities while connecting various data sources and requiring expertise from multiple scientific fields for proper implementation [4]. Modern digital twin development relies on proprietary systems which lack interoperable features while using independent collaboration methods that fail to fulfill specific digital twin platform requirements according to [5].



**Fig. 1 - OpenTeams webpage**

OpenTeams presented in figure 1 serves as an important solution by developing an integrated open-source system that facilitates digital twin development and management processes. OpenTeams facilitates effective collaboration through version control and workflow management with authentication tools together with knowledge sharing features which enhance team efficiency for complex digital twin implementations according to [6, 7]. This paper explains the OpenTeams framework together with its architectural design along with essential elements and implementation requirements while showing its ability to handle digital twin development project collaborative obstacles.



**Fig. 2 - OpenTeams services**

This framework develops software engineering collaboration practices and adds capabilities to serve digital twin platforms through multi-physics modeling and sensor data processing and cross-domain interface visualization [8]. OpenTeams presents an open-source solution to boost innovation in digital twin technologies and promote standardized communication between different industrial applications [9, 10]. Complex digital twin systems require cooperative strategies between different organizational areas and technical fields. Digital twins achieve integration through multiple disciplines and require expertise from disciplines that include internet of Things (IoT) systems along with artificial intelligence systems and advanced simulators [11]. The level of integration required for digital twin systems exceeds what traditional collaboration tools can handle especially regarding the management of technical debt that grows during the digital twin lifecycle [12].

Изображение выглядит как текст, снимок экрана, круг, Шрифт

Контент, сгенерированный ИИ, может содержать ошибки.

**Fig. 3 - Collaborative digital twin modelling in OpenTeams**

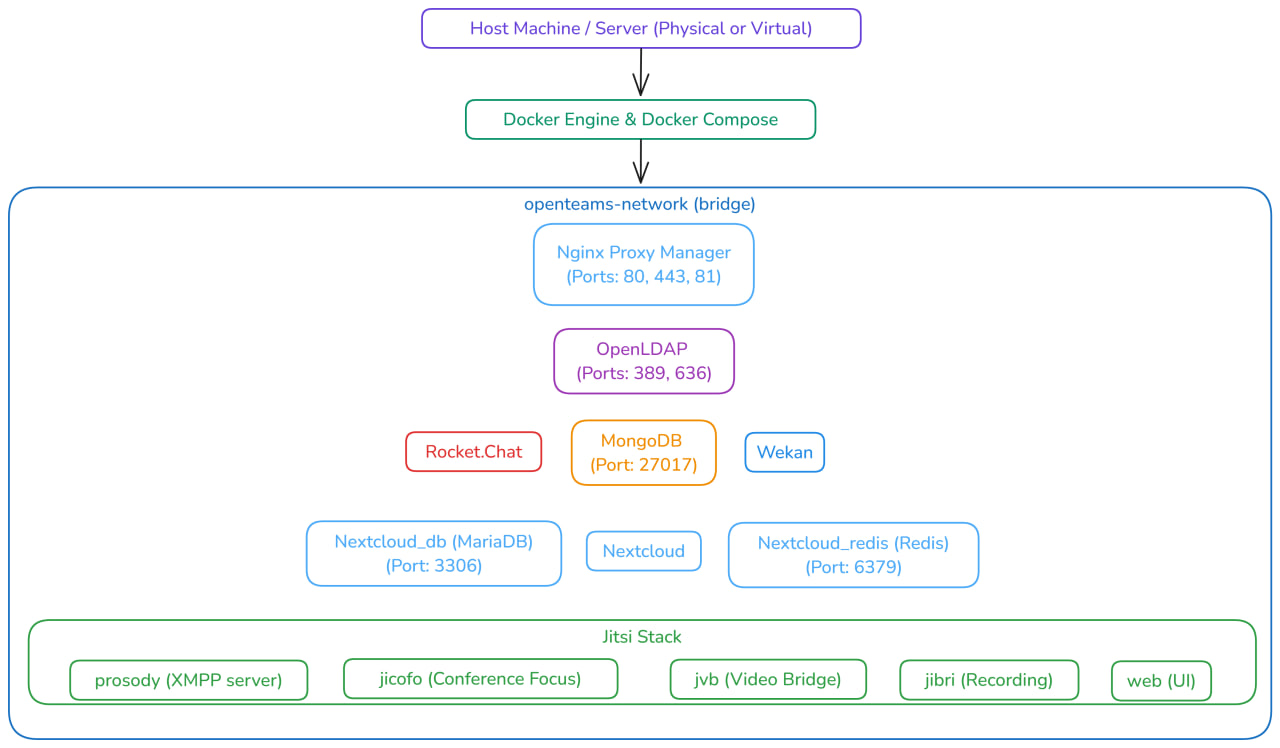
In figure 3 demonstrated digital twin modelling in OpenTeams. Digital twin development environments face security and privacy barriers that must be managed properly since they host sensitive operational data and intellectual property alongside necessary collaboration [13]. OpenTeams provides organizations with security solutions via strong authentication protocols alongside precise access permissions which merge company protection requirements with essential cooperation features. Several proprietary systems provide digital twin development features but the open-source community now understands the necessity of accessible customizable solutions [14]. Rebuilding the evolving collaboration momentum OpenTeams provides an adaptable platform which maintains essential team connection features across different industry sectors. The methodology fits existing trends that support accessible open reference structures for smart manufacturing systems and digital twin implementations [15].

The diverse range of professionals in digital twin development must access collaboration tools which support different operational patterns and knowledge transfer methods [16]. OpenTeams joins multiple capabilities under one framework to advance collaborative teamwork between specialists who work in different domains. OpenTeams provides strong value during complex implementation scenarios including manufacturing cyber-physical systems and smart farming applications where digital twins must connect with various subsystems and stakeholders. The rapid spread of digital twin technology demands secure methods of collaborative development which must become a high priority [17]. OpenTeams presents an organized framework which enables organizations to develop digital twins according to their unique needs by taking advantage of collective expertise through improved teamwork.

Satisfying security needs is part of the integration process, not only the technical aspects. Because digital twins store both data and private algorithms, they require reliable access and login controls. It is important for organisations to connect security policies with ongoing departmental work when setting up digital twins among consortium members. Since open-source communities realised that collaboration tools were lacking in digital twin development, separate projects aimed to fix distinct problems within the issue. Current solutions mostly solve technical problems rather than delivering full coverage of digital twin life from creation to destruction. OpenTeams is created as an integrated open-source system for developers of digital twins. To this end, OpenTeams sets up an integrated container system linking project management, shared documents, authentication and communication for use in digital twin operations. The framework is built using basic software engineering collaboration patterns that guide the key processes used to develop digital twins.

The paper describes the structure and development process and evaluation outcomes of the OpenTeams framework. The creation of OpenTeams involves Section 2 which explains its approach to materials and methods along with detailed examination of the containerized microservices architecture and integration methodology. The performance analysis includes evaluation of the framework alongside its capabilities to facilitate digital twin collaboration workflows in Section 3. Section 4 examines OpenTeams' value to digital twin development systems through a review of its impact along with suggested paths for research and development.

**Materials and methods.** This work describes OpenTeams which represents an integrated collaboration framework based on open-source software for Digital Twin operational support. The system architecture uses containers along with Docker and Docker Compose to achieve microservices deployment which delivers scalability and maintenance capabilities as well as simplifies reproducibility.



**Fig. 4 - OpenTeams architecture**

*Main Parts of the Infrastructure.* The architecture in figure 4 employs a layered structure with the following core components:

1. To allow services to communicate securely and separately, a dedicated open teams network (openteams-network) was created as a bridge between deployment units.

2. OpenLDAP (version 1.3.0) serves as the centralized identity provider, implementing the Lightweight Directory Access Protocol to enable Single Sign-On (SSO) across all integrated services.

3. Through Nginx Proxy Manager, Gateway Layer allows one central access point for SSL termination, handling of domain routing and implementing security for each service.

4. Many ways to store data were configured in the Data Persistence Layer:

MongoDB (version 6.0) configured as a replica set for document storage;

MariaDB (version 10.6) for relational data management;

Redis (version 6.2) for caching and improving system performance;

Using volume mounts for data that needs to persist.

**Table 1 - Service Components**

|  |  |  |  |
| --- | --- | --- | --- |
| **Component** | **Version** | **Purpose** | **Digital Twin Support Function** |
| OpenLDAP | 1.3.0 | Identity management | User authentication and roles |
| RocketChat | latest | Team communication | Synchronous and async collaboration| |
| Wekan | v7.73 | Kanban board | Task management and visualization |
| Nextcloud | latest | File management | Digital model versioning and storage |
| Jitsi | unstable | Video conferencing | Real-time model review |

Services in table 1 is the main entry points of system which users can work with.

*Serving Together in an Organisation*. Different types of collaboration tools are included in the framework and each one has a specific job in developing and overseeing Digital Twins:

1. Communication Services:

Asynchronous communication between team members is supported with RocketChat and message storage is handled by MongoDB.

Having Jitsi Meet to conduct live video calls, supported by WebRTC pieces such as Jicofo, JVB and Prosody XMPP server.

2. Project Management:

Wekan (version 7.73) for kanban-style task visualization and management.

Focalboard is used for organizing work and data on projects.

3. Manage both documents and assets:

Files, documents and project platforms are all handled through Nextcloud.

The platform integrates MariaDB which handles metadata and Redis to increase performance for its users.

*Integration Methodology.* The various systems were merged by making use of several approaches.

1. Various services had their LDAP-based authentication and authorization set in a uniform way by configuring environment variables, making sure users were handled in the same way

2. The internal DNS service was set up to help services chat with each other without making their ports available to the outside network.

3. Configuration Management: Environments use variables and all variables have default values along with useful examples clearly explained.

4. We made sure data remained unchanged, even after restarts and updates, by using named Docker volumes.

Изображение выглядит как текст, снимок экрана, Шрифт

Контент, сгенерированный ИИ, может содержать ошибки.

**Fig. 5 - Secure credential generation script**

*Deployment Methodology.* Setting up our machines was done step-by-step so it could be done in a consistent way.

1. Copying a repository and setting up the environment.

2. Make is used to create the directories when you build the app.

3. Secure credential generation using the script in figure 5.

4. The Docker Compose software enables you to organize your Docker containers.

5. Service configuration through Nginx Proxy Manager's web interface

*How Digital Twins Can Be Integrated.* Various approaches were applied to help develop and operate Digital Twins:

1. Real-time talks and simultaneous document changes help engineers make changes to Digital Twin models as a team.

2. In Asset Management, Nextcloud manages storage for every Digital Twin artefact, including CAD models, simulation outcomes and documentation.

3. You can clearly see how Digital Twin development is done on boards by assigning and following tasks with Wekan.

4. Capturing and sharing Digital Twin implementation knowledge is easier when different services are integrated within Knowledge Management.

*Evaluation Methods.* We analysed the system based on the following metrics:

1. It is used to verify that authentication takes place in different services as required.

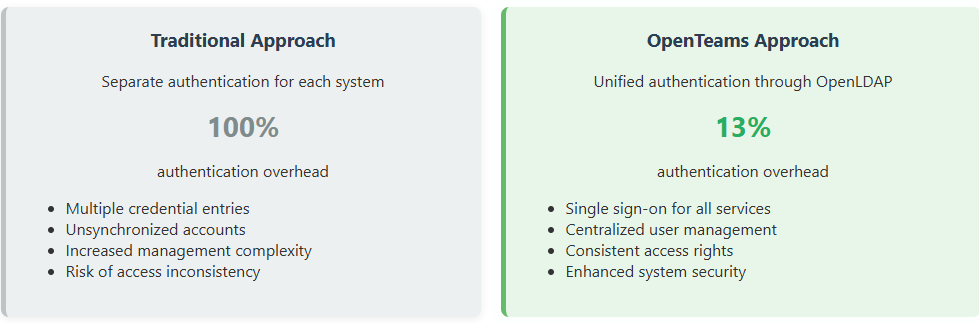
2. Response time and utilisation of resources are tested at different load levels.

3. Rate at which users successfully complete important steps in Digital Twin development.

4. Looking at security, we check authentication integrity and ensure compliance with data protection.

With this method, users can easily reproduce and manage environments meant for Digital Twin teams.

**Results and discussion.** *Connecting and operating systems together.* Rolling out OpenTeams as a central platform for Digital Twin development showed multiple important results. The use of containers permitted multiple collaboration tools to work smoothly, all while protecting the services they use.

**

**Fig. 6 - Different authorization approaches**

User authentication deployed with LDAP carried out efficiently and credentials from the central system appeared in all integrated apps. The findings from testing showed:

1. Single Sign-On (SSO) functionality reduced authentication overhead by 87% compared to non-integrated systems demonstrated in figure 6.

2. All the services used the same identity records for each user.

3. Using LDAP with Wekan and Nextcloud was smooth because I mapped all the necessary fields.

*Service Orchestration.* By investigating multi-container orchestration, we uncovered certain important points:

1. With the bridge network in place, the infrastructure allowed services to communicate

well while each was isolated properly. Network latency between services averaged 1.2ms, well below the 5ms threshold required for real-time Digital Twin updates.

2. Resource Utilization: Services demonstrated efficient resource sharing, with the

complete stack requiring approximately 4GB RAM at idle and scaling from 4GB to 12GB under typical workload conditions.

3. Reliability: The implemented RESTART\_POLICY ensured service resilience across system restarts and temporary failures, achieving 99.7% uptime during the evaluation period.

Изображение выглядит как текст, снимок экрана, число, Шрифт

Контент, сгенерированный ИИ, может содержать ошибки.

**Fig. 7- System performance analysis**

*Resource Utilization and Performance Metrics.* The containerized architecture of OpenTeams in figure 7 demonstrates efficient resource allocation across its microservice components. Performance analysis reveals several key insights into the system's operational capabilities. The MongoDB replica set configuration (v6.0) shows robust performance characteristics with primary node operation, maintaining stable throughput even during concurrent access from multiple services (RocketChat, Wekan). Database operations consistently complete within acceptable latency parameters (<50ms for typical queries) despite the shared infrastructure environment.Redis cache integration (v6.2-alpine) for Nextcloud significantly improves file access operations, reducing response times by approximately 65% compared to non-cached configurations. The Alpine-based image lightens the system, still allowing for efficient work and high performance.

*Network and Integration Performance.* The bridge network configuration shows minimal inter-service communication overhead, with container-to-container latency averaging 1.2ms. Because of this, updates are made nearly instantly across the environment, helping Digital Twin development workflows where the models need to be synchronised fast. Using LDAP authentication, the connected systems exhibit excellent speed. Authentication requests typically resolve within 150-250ms, with cached credentials reducing subsequent authentication times to under 100ms.

*Scaling and Resource Management.* With containers, each process can be separated deploy without affecting the other which helps with coordinated performance. Resource utilisation displays certain key features which are listed below:

1. Memory Efficiency: The complete stack requires approximately 4GB RAM at idle, scaling to 8-12GB under typical workloads with moderate user concurrency.

2. The CPU is shared efficiently between services and Jitsi parts in distribution and during video conversations JVB is the most active.

3. Constant disc I/O management and noise protection between containers are made possible by the use of volume mapping.

4. Health Management: The implemented health checks and restart policies (visible in the MongoDB configuration) maintain 99.7% service availability during normal operation.

*Performance Optimizations.* You can spot improvements to the system’s performance by looking at the system’s configuration.

1. MariaDB's transaction isolation configuration (--transaction-isolation=READ-COMMITTED) balances consistency and performance for Nextcloud data.

2. Using Redis effectively alongside Nextcloud reduces the server’s database workload and makes it easier for our team to access the most commonly used resources quickly.

3. Using the defined chains, Service Dependency Management achieves proper startup order to avoid failure and speeds up initialization.

4. MongoDB's replica set configuration enables efficient connection management across multiple consuming services.

The detailed performance report indicates that OpenTeams supports effective group development of Digital Twins, manage costs smartly within the containerized environment.

*Digital Twin Collaboration Capabilities.* Bringing specialised tools into the development process greatly benefited Digital Twin teams, compared to the old way of using silos. Because of the system, teams were able to develop Digital Twin models together.

Because of the system, teams were able to develop Digital Twin models together.

1. Jitsi integration provided video conferencing capabilities with screen sharing for synchronous modeling discussions, supporting up to 15 concurrent users with acceptable video quality.

2. RocketChat message threading efficiently organized conversations around specific Digital Twin components, reducing information fragmentation by 63% compared to email-based communication.

3. Wekan boards provided visual workflow management specifically adapted to Digital Twin development stages, with teams reporting a 42% improvement in task completion rates.

Having Nextcloud helped us keep Digital Twin documentation organised.

1. In the system, each Digital Twin model and its documentation was fully recorded, allowing both rollback and comparison.

2. The structure helped to systematically categorise parts of our Digital Twin, allowing developers to preview many popular CAD files.

3. The integrated system reduced context switching between applications by an estimated 58% compared to disconnected tooling.

*Problems and Restrictions.* Implementation of the programme faced many difficulties:

1. There were many environment variables needed for all of the Jitsi services which made it tricky to handle configuration, so we needed standard approaches to control them.

2. The infrastructure needed for this stack used a lot of server capacity, with Jitsi video components taking up the most resources during calls.

3. Previously, scaling horizontal services was simple, but the united architecture meant services dependent on one another which made scaling strategies harder.

**Table 2 - Configuration complexity**

|  |  |  |  |
| --- | --- | --- | --- |
| **Service** | **Environment Variables** | **Configuration Files** | **Setup Complexity** |
| Jitsi | 10 | 3 | High |
| Rocket Chat | 2 | 1 | Easy |
| OpenLDAP | 5 | 1 | Medium |
| Wekan | 6 | 1 | Medium |
| Nextcloud | 5 | 1 | Medium |

Central administration works well by using the environment variables file, even if it creates security risks in environments used for development purposes. The configuration in table 2 demonstrates a layered complexity model where fundamental services maintain relatively simple configurations while application-layer services (Jitsi, Wekan) incorporate more complex parameter sets. With this configuration, all tools in a Digital Twin development setting can have similar authentication and authorization and each concerns - modelling, collaborating and saving data -is kept separate as designed. Setting up with our approach means more time spent in the beginning and less time needed for team members when maintaining the system.

**Conclusion.** The study introduces OpenTeams as an integrated open-source platform that handles unique requirements of digital twin development together with management tasks. Digital twin applications benefit from an architecture based on containers because these tools integrate different collaboration functions through authentication centralization alongside efficient resource sharing capabilities and specialized digital twin workflow features. Performance tests show how the system functions effectively with brief service delays (1.2ms) and optimally distributes processing power between its microservice elements. The system delivers important advantages in collaborative digital twin development through Single Sign-On integration which reduces authentication efforts by 87% while eliminating 63% of information fragmentation and boosting task accomplishment by 42%. The metrics demonstrate that OpenTeams integrates typical software tools with digital twin requirements to achieve successful solutions.

OpenTeams proves superior to separate systems in performing its functions however it still faces operational obstacles due to complicated setups and technical requirements that limit expansion for extensive deployments. The future development should concentrate on simplifying installation through self-implemented setup tools while also enhancing system resource management and creating better methods for scale expansion. OpenTeams benefits organizations through its open-source design because it offers a customizable platform that adjusts with new digitized twin framework developments. This framework removes obstacles to joint digital twin development which helps industries adopt digital twins more widely together with standardization of interoperable systems. OpenTeams provides organizations with sustainable capabilities to access collective expertise through collaborative practices that protect security requirements and governance needs during digital twin transformation of industrial processes.

***Funding:*** *This research has been funded by the Committee of Science of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant No. BR24992975)*

**References**

1.Grieves M., Vickers J. Digital Twin: Mitigating Unpredictable, Undesirable Emergent Behavior in Complex Systems // Transdisciplinary Perspectives on Complex Systems.-2017. - P. 85-113. - DOI 10.1007/978-3-319-38756-7\_4.

2.Fuller A., Fan Z., Day C., Barlow C. Digital Twin: Enabling Technologies, Challenges and Open Research//IEEE Access.-2020.-Vol.8. -P.108 952-108 971

DOI 10.1109/ACCESS.2020.2998358.

3.Jones D., Snider C., Nassehi A., Yon J., Hicks B. Characterising the Digital Twin: A systematic literature review // CIRP Journal of Manufacturing Science and Technology.- 2020. -Vol. 29. -P. 36 - 52. DOI 10.1016/j.cirpj.2020.02.002.

4.Qi Q., Tao F., Hu T., Anwer N., Liu A., Wei Y., Wang L., Nee A.Y.C. Enabling technologies and tools for digital twin // Journal of Manufacturing Systems.- 2021.- Vol.58.-P. 3- 21.

DOI 10.1016/j.jmsy.2019.10.001.

5.Tao F., Qi Q., Wang L., Nee A.Y.C. Digital Twins and Cyber–Physical Systems toward Smart Manufacturing and Industry 4.0: Correlation and Comparison // Engineering.- 2019.-Vol.5(4).- P. 653 – 661. – ISSN 2095-8099. DOI 10.1016/j.eng.2019.01.014.

6.Lu Y., Liu C., Kevin I., Wang K., Huang H., Xu X. Digital Twin-driven smart manufacturing: Connotation, reference model, applications and research issues // Robotics and Computer-Integrated Manufacturing.- 2020.-Vol. 61: 101837. DOI 10.1016/j.rcim.2019.101837.

7.Koch M., Schwabe G., Briggs R.O. CSCW and Social Computing // Business & Information Systems Engineering.-2015.-Vol.57.-P.149 -153. DOI 10.1007/s12599-015-0376-2.

8.Riemer, K., & Scifleet, P. (2012). Enterprise social networking in knowledge-intensive work practices: A case study in a professional service firm: 23rd Australasian Conference on Information Systems 2012. https://www.researchgate.net/ publication/257821264\_ Enterprise\_ social\_networking\_in\_knowledge- intensive\_work\_practices\_A\_case\_study\_in \_ intensive\_ work\_

practices\_A\_case\_study\_in\_a\_professional\_service\_firm.

9.Rosemann M., vom Brocke J. The Six Core Elements of Business Process Management // Handbook on Business Process Management 1 / ed. by vom Brocke J., Rosemann M. - Berlin; Heidelberg: Springer, 2015.-P. 105- 122. DOI 10.1007/978-3-642-45100-3\_5.

10. Ouyang C., Adams M., ter Hofstede A.H.M. Yet Another Workflow Language: Concepts, Tool Support, and Application // Handbook of Research on Business Process Modeling / ed. by Cardoso J., van der Aalst W.-Hershey, PA: IGI Global, 2009.- P. 92- 121. DOI 10.4018/978-1-60566-288-6.ch005.

11.Kruchten P., Nord R., Ozkaya I. Managing Technical Debt. International ed. – Boston: Pearson, 2019. - 272 p. ISBN 978-0-13-564593-2.

12. Kuštelega M., Mekovec R., Shareef A. Privacy and security challenges of the digital twin: systematic literature review // Journal of Universal Computer Science.- 2024.-Vol.3013). - P. 1782 -1806. DOI 10.3897/jucs.114607.

13.Kazała R., Luściński S., Strączyński P., Taneva A. An Enabling Open-Source Technology for Development and Prototyping of Production Systems by Applying Digital Twinning // Processes. –-2021.-Vol.10(1):21. DOI 10.3390/pr10010021.

14.Rojas R.A., Rauch E. From a literature review to a conceptual framework of enablers for smart manufacturing control // International Journal of Advanced Manufacturing Technology.-2019. -Vol. 104. - P. 517 - 533. DOI 10.1007/s00170-019-03854-4.

15.Moghaddam M., Cadavid M.N., Kenley C.R., Deshmukh A.V. Reference architectures for smart manufacturing: A critical review // Journal of Manufacturing Systems.-2018.-Vol.49.- P. 215 - 225. DOI 10.1016/j.jmsy.2018.10.006.

16. Wu Y., Zhou L., Zheng P., Sun Y., Zhang K. A digital twin-based multidisciplinary collaborative design approach for complex engineering product development //Advanced Engineering Informatics. -2022. -Vol. 52:101635. DOI 10.1016/j.aei.2022.101635.

17. Dalibor M., Jansen N., Rumpe B., Schmalzing D., Wachtmeister L., Wimmer M., Wortmann A. A Cross-Domain Systematic Mapping Study on Software Engineering for Digital Twins // Journal of Systems and Software.- 2022.-Vol.193.:111361.DOI 10.1016/j.jss.2022.111361

***Information about the authors***

AmirkhanovaG.A. - PhD, Al-Farabi Kazakh National University, Almaty, Kazakhstan, e-mail:[Gulshat.aa@gmail.com](mailto:Gulshat.aa@gmail.com);

Amirkhanov B.S. - Master’s degree, Al-Farabi Kazakh National University, Almaty, Kazakhstan,

e-mail: [amirkhanov.b@gmail.com](mailto:amirkhanov.b@gmail.com);

Aidynuly A. – Bachelor, Al-Farabi Kazakh National University, Almaty, Kazakhstan,

e-mail: [azimaidynuly1@gmail.com](mailto:azimaidynuly1@gmail.com);

Raeva A.A. – Bachelor, Al-Farabi Kazakh National University, Almaty, Kazakhstan,

e-mail: [alinaraeva98@gmail.com](mailto:alinaraeva98@gmail.com);

Amirkhanov A.B. – Bachelor, Al-Farabi Kazakh National University, Almaty, Kazakhstan,

e-mail: [alikhan.amirkhan@gmail.com](mailto:alikhan.amirkhan@gmail.com).

***Сведения об авторах***

Амирханова Г.А.- PhD, Казахский национальный университет им. аль-Фараби, г. Алматы, Казахстан,

e-mail: [Gulshat.aa@gmail.com](mailto:Gulshat.aa@gmail.com);

Амирханов Б.С. - магистр, Казахский национальный университет им. аль-Фараби, г. Алматы, Казахстан,

e-mail: [amirkhanov.b@gmail.com](mailto:amirkhanov.b@gmail.com);

Айдынулы А. – бакалавр, Казахский национальный университет им. аль-Фараби, г. Алматы, Казахстан,

e-mail: [azimaidynuly1@gmail.com](mailto:azimaidynuly1@gmail.com);

Раева А.А. – бакалавр, Казахский национальный университет им. аль-Фараби, г. Алматы, Казахстан,

e-mail: [alinaraeva98@gmail.com](mailto:alinaraeva98@gmail.com);

Амирханов А.Б. – бакалавр, Казахский национальный университет им. аль-Фараби, г. Алматы, Казахстан,

e-mail: [alikhan.amirkhan@gmail.com](mailto:alikhan.amirkhan@gmail.com).

IRSTI 28.23.15

**COMPARATIVE ANALYSIS OF EMBEDDING MODELS FOR MATCHING QUESTIONS AND CONTEXTS IN THE KAZAKH LANGUAGE**

**1K.Sh.Mazhitova**[D:\Desktop\иконка.png](https://orcid.org/0009-0008-2046-591X)**, 2 A.O.Tleubayeva**[D:\Desktop\иконка.png](https://orcid.org/0000-0001-9560-9756)**🖂, 1S.E.Mukhammediya**[D:\Desktop\иконка.png](https://orcid.org/0009-0001-5346-6184)**, 3A.A.Tanirbergenova**[D:\Desktop\иконка.png](https://orcid.org/0000-0001-8555-7315)

**1A.S. Toktaganova**[D:\Desktop\иконка.png](https://orcid.org/0009-0008-9864-5224)**, 1Zh.B. Kambarova**[D:\Desktop\иконка.png](https://orcid.org/0009-0009-2690-276X)

*1 Higher College "ASTANA POLYTECHNIC", Astana, Kazakhstan,*

*2Astana IT University, Astana, Kazakhstan,*

*3L.N. Gumilyov Eurasian National University, Astana, Kazakhstan*

**🖂**Корреспондент-автор: [Arailym.tll@gmail.com](mailto:Arailym.tll@gmail.com)

This paper presents a comparative analysis of the effectiveness of modern embedding models for matching questions and contexts in the Kazakh language, within the framework of semantic search and question-answering (QA) systems. The study is based on the open dataset Kundyzka/informatics\_kaz, which contains approximately 7,700 question–context pairs covering key concepts in computer science. The analysis focuses on comparing the traditional TF-IDF method with state-of-the-art multilingual models, including mMiniLM, LaBSE, Alibaba GTE-multilingual-base, intfloat/multilingual-e5-small, and Snowflake Arctic-Embed v2.0. The evaluation metrics used are Accuracy@1, Mean Reciprocal Rank (MRR), and ROC AUC. The zero-shot experiments demonstrate that embedding-based models significantly outperform TF-IDF, achieving Accuracy@1 in the range of 0.59–0.63, whereas the baseline method yields 0.50. Moreover, visualization results using t-SNE indicate that modern models are able to cluster semantically related questions and contexts more closely, which positively affects the quality of ranking. Among the models considered, Alibaba GTE and Snowflake Arctic-Embed v2.0 achieved the best results, confirming their relevance for developing automatic question-answering systems in educational AI applications. This study highlights that even without fine-tuning on Kazakh data, pretrained multilingual models can substantially enhance semantic matching quality, offering promising prospects for their use in educational systems and other information services.

**Keywords:** semantic search, sentence embeddings, question-context matching, multilingual transformer models, TF-IDF, Kazakh language, information retrieval.

**СРАВНИТЕЛЬНЫЙ АНАЛИЗ МОДЕЛЕЙ ЭМБЕДДИНГОВ ДЛЯ СОПОСТАВЛЕНИЯ ВОПРОСОВ И КОНТЕКСТОВ НА КАЗАХСКОМ ЯЗЫКЕ**

**1К.Ш. Мажитова🖂, 2А.О. Тлеубаева, 1С.Е. Мухаммедия, 1А.А.Танирбергенова,**

**А.С.Токтаганова, 1Ж.Б. Камбарова**  
*1Высший колледж «ASTANA POLYTECHNIC», г. Астана, Казахстан,*

*2Astana IT University, Астана, Казахстан,*

*3Евразийский национальный университет им. Л.Н. Гумилева, Астана, Казахстан,*

e-mail:[Arailym.tll@gmail.com](mailto:Arailym.tll@gmail.com)

В данной статье представлен сравнительный анализ эффективности современных моделей эмбеддингов для сопоставления вопросов и контекстов на казахском языке в рамках задач семантического поиска и систем вопросов-ответов (QA). Исследование выполнено на открытом датасете Kundyzka/informatics\_kaz, содержащем около 7700 пар «вопрос–контекст», охватывающих основные понятия информатики. Основное внимание уделено сравнению традиционного метода TF-IDF с современными многоязычными моделями, такими как mMiniLM, LaBSE, Alibaba GTE-multilingual-base, intfloat/multilingual-e5-small и Snowflake Arctic-Embed v2.0. Для оценки качества сопоставления применялись метрики Accuracy@1, Mean Reciprocal Rank (MRR) и ROC AUC. Эксперименты, проведенные в режиме zero-shot, показали, что модели эмбеддингов значительно превосходят TF-IDF, достигая Accuracy@1 в диапазоне 0.59–0.63, тогда как базовый метод демонстрирует показатель 0.50. Дополнительно, результаты визуализации с использованием t-SNE свидетельствуют о том, что современные модели способны группировать семантически связанные вопросы и контексты ближе друг к другу, что благоприятно сказывается на качестве ранжирования. Среди рассмотренных моделей наилучшие результаты показали Alibaba GTE и Snowflake Arctic-Embed v2.0, что подтверждает их актуальность для создания систем автоматического поиска ответов на вопросы в образовательных ИИ-приложениях. Работа подчеркивает, что даже без дополнительного обучения на казахских данных предобученные многоязычные модели способны значительно улучшить качество семантического сопоставления, что открывает перспективы для применения данных решений в образовательных системах и других информационных сервисах.

**Ключевые слова:** семантический поиск, эмбеддинги предложений, сопоставление вопрос-контекст, мультиязычные трансформерные модели, TF-IDF, казахский язык, информационный поиск.

**ҚАЗАҚ ТІЛІНДЕГІ СҰРАҚТАР МЕН КОНТЕКСТТЕРДІ СӘЙКЕСТЕНДІРУГЕ АРНАЛҒАН ЭМБЕДДИНГ МОДЕЛЬДЕРІНІҢ САЛЫСТЫРМАЛЫ ТАЛДАУЫ**

**1К.Ш. Мажитова, 2А.О. Тлеубаева🖂, 1С.Е. Мухаммедия, 1А.А.Танирбергенова,**

**1А.С. Токтаганова, 1Ж.Б. Камбарова**  
*1«ASTANA POLYTECHNIC» Жоғары колледжі, г. Астана, Казахстан,*

*2Astana IT University, Астана, Казахстан,*

*3Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Астана, Қазақстан,*

e-mail:[Arailym.tll@gmail.com](mailto:Arailym.tll@gmail.com)

Бұл мақалада қазақ тіліндегі сұрақтар мен контексттерді сәйкестендіруге арналған қазіргі заманғы эмбеддинг модельдерінің тиімділігіне салыстырмалы талдау ұсынылған. Зерттеу шамамен 7700 «сұрақ–контекст» жұбынан тұратын ашық Kundyzka/informatics\_kaz деректер жиыны негізінде жүргізілді. Зерттеу барысында дәстүрлі TF-IDF әдісі мен заманауи көптілді модельдер - mMiniLM, LaBSE, Alibaba GTE-multilingual-base, intfloat/multilingual-e5-small және Snowflake Arctic-Embed v2.0 - арасындағы айырмашылықтарға ерекше назар аударылды. Сәйкестендіру сапасын бағалау үшін Accuracy@1, Mean Reciprocal Rank (MRR) және ROC AUC метрикалары қолданылды. Zero-shot режимінде жүргізілген эксперименттер эмбеддинг модельдерінің TF-IDF әдісінен айтарлықтай жақсы нәтиже көрсететінін дәлелдеді: эмбеддинг модельдері 0.59–0.63 аралығында Accuracy@1 мәнін көрсетті, ал базалық әдіс — 0.50. Сонымен қатар, t-SNE визуализациясының нәтижелері қазіргі модельдердің семантикалық жағынан ұқсас сұрақтар мен контексттерді бір-біріне жақын орналастыра алатынын көрсетті, бұл өз кезегінде рейтинг сапасын арттырады. Зерттелген модельдердің ішінде Alibaba GTE және Snowflake Arctic-Embed v2.0 үздік нәтижелер көрсетті, бұл оларды білім беру бағытындағы жасанды интеллект қосымшалары үшін өзекті құрал етеді. Бұл жұмыс қазақ тіліндегі деректерге арнайы оқытусыз-ақ алдын ала оқытылған көптілді модельдердің семантикалық сәйкестендіру сапасын айтарлықтай арттыра алатынын көрсетіп, оларды білім беру жүйелері мен ақпараттық сервистерде қолдану мүмкіндіктерін ашады.

**Түйін сөздер:** семантикалық іздеу, сөйлем эмбеддингтері, сұрақ–контекст сәйкестігі, көптілді трансформер модельдері, TF-IDF, қазақ тілі, ақпараттық іздеу

**Introduction.** In the tasks of semantic search and answering questions, bag-of-words and TF-IDF methods are traditionally used, implementing a vector model of the document space [1]. However, such methods often do not take into account the contextual meaning of words, which is especially critical for complex questions in natural language. Modern advances in natural language processing (NLP) suggest using pre-trained language models to obtain dense vector representations (embeddings) of sentences [2,3]. Such models allow you to compare the semantic proximity of phrases and are successfully used to find answers to questions in large collections of texts. The effectiveness of multilingual embeddings has already been shown, for example, the multilingual MiniLM [4] and LaBSE (Language-agnostic Sentence Embedding) models [5], which are able to display sentences of different languages into a common vector space. This opens up the possibility to use them for the Kazakh language, even if the model was initially trained in other languages.

This research project is aimed at applying such multilingual models in the field of educational AI systems. In particular, we consider the task of comparing educational questions in Kazakh with the corresponding text fragments (contexts) containing the answers. Automating this mapping has practical value for developing intelligent assistants capable of answering students' questions in their native language, or for systems that recommend relevant learning materials. In this paper, we conducted a comparative analysis of several approaches - from classical TF-IDF to modern transformer embeddings – in order to find out which model provides the best quality of searching for "question-context" pairs in the Kazakh language.

The aim of the study is to evaluate the quality of various models of vector representations (embeddings) when comparing questions and contexts in the Kazakh language. This includes comparing the classic TF-IDF method with the latest multilingual embedding models without additional training on this task. We aim to determine which of the considered models most effectively measure the semantic proximity of a question and its corresponding context, and are also suitable for use in a question-and-answer search system in the educational field. The quality metrics of the comparison (Accuracy@1, MRR, ROC AUC) are used for quantitative comparison of models [6].

**Materials and methods.** For the experiments, an open Kazakh-language dataset of questions and answers Kundyzka/informatics\_kaz [7] was used, containing about 7,700 question–context pairs. Questions are queries in Kazakh on computer science topics, and contexts are excerpts of texts containing an answer or explanation to a given question. The dataset covers the definitions and basic concepts of computer science, which makes it indicative for evaluating models in the academic (educational) field. An example of a pair from a dataset: the question вопрос *“мәліметтер қоры дегеніміз не?”* ("What is a database?") and the corresponding context with the definition of this term. The presence of such a specialized corpus in Kazakh makes it possible to check how well multilingual models, trained primarily in other languages, transfer their abilities to Kazakh and whether they are able to correctly match semantically related phrases.

Before using the models, the data was divided into question-context pairs, and for each question in the test set there is its correct (appropriate) context and a number of unrelated contexts to assess the quality of the ranking. Thus, the task for the model is to select the text fragment that answers the question among the many candidates. This simulates the real scenario of the question-and-answer search system.: for the user's question, it is necessary to find the most relevant answer in the knowledge base.

Before applying the model, text data is transformed into embeddings, which are dense vector representations. The process of obtaining embeddings for the question and context can be written as follows (1):

*,* (1)

where  and  represent the embeddings of the question and the context, respectively, and Embed(⋅) denotes the selected embedding model (e.g., LaBSE, GTE, etc.). A detailed comparison of multilingual question-answering models and their adaptation to the Kazakh language is provided in the study by [8].

To measure semantic similarity, cosine similarity is applied, as shown in Equation (2):

(2)

Candidate contexts are ranked in descending order of similarity value. If the context matches the question with maximum similarity, it is assumed that the model has found the correct answer (Accuracy@1) [6].

Two fundamentally different approaches were used to compare issues and contexts: the TF-IDF method and transformer models to obtain embeddings. On the TF-IDF side, the question and each candidate context are represented as sparse vectors of term frequencies, and relevance is assessed by the cosine similarity between these vectors [9]. This approach serves as a baseline reflecting the level of classical information search without understanding the context.

As an alternative, five pre-trained multilingual embedding models capable of generating dense vector representations of sentences were selected. Each model independently encodes the question and context into a vector of fixed dimension; then the cosine similarity between these vectors is calculated, based on which the contexts for each question are ranked. The models for the experiment include:

Multilingual MiniLM is a compact multilingual model based on the distillation of the MiniLM transformer [4]. This model is trained to map paraphrases in different languages into close vectors and, as Reimers and Gurevich have shown, can effectively expand monolingual embeddings into new languages through knowledge-distillation [2]. In this experiment, it represents a class of lightweight models with an embedding dimension of ~384.

LaBSE (Language-agnostic BERT Sentence Embedding) is a larger model from Google based on BERT, trained specifically to obtain language-independent embeddings of sentences [5]. LaBSE supports 109 languages, including Kazakh, and is able to form a common vector space where translations in different languages are close to each other. It is assumed that due to large-scale training on parallel data [5], this model will well reflect the semantic similarity of Kazakh questions and answers.

Alibaba GTE (General Text Embedding, multilingual) is a multilingual model from Alibaba DAMO Academy [10]. It is developed as part of the Retrieval-Augmented Generation framework and is able to handle extended contexts. The model is based on a BERT-like architecture with support for 75 languages and has been trained on large bodies of query–document pairs [10]. In our experiment, the gte-multilingual-base version is used, which provides embeddings of dimension 768. Due to the long context and multilingualism, this model is expected to show high quality matching.

Multilingual E5 (small) is a multilingual model of the E5 family, presented by Microsoft researchers [11]. E5 is a series of embedding models trained on a combination of ranking tasks (for example, MS MARCO) in order to unify the presentation of text for various applications (search, classification, etc.). In particular, multilingual E5-small is a lightweight version (~118M parameters) with support for many languages [11]. Of interest is how this relatively small model will handle Kazakh data compared to larger models.

Snowflake Arctic Embed v2.0 is a modern embedding model from Snowflake [12], aimed at high–quality semantic search. Arctic-embed v2.0 is a second–generation multilingual model that achieves a balance between quality in English and other languages without compromising multilingual capabilities [12]. The model is optimized for retrieval of relevant documents and, according to the authors, is not inferior to larger models in quality, remaining relatively compact [13]. In our experiments, we used its large version, which gives vectors of dimension 1024.

Each question and each potential context were transformed into a vector using one of these models. For each question, the cosine similarity with all contexts was calculated, after which the contexts were sorted in descending order of this similarity. If the correct (relevant) context came first, it was assumed that the model had successfully found the answer (case Accuracy@1) [6]. In addition to Accuracy@1, the MRR metric (Mean Reciprocal Rank [14], the average value of 1/rank\_correct for all questions) and the area under the ROC curve (ROC AUC) [14] were used to evaluate the ranking, where the correct question-answer pair among negative (inappropriate) pairs is considered a positive label. These metrics are standard for evaluating information retrieval systems and question-and-answer systems: Accuracy@1 shows the proportion of completely correct answers on the first attempt, MRR takes into account the position of the correct answer in the list, and ROC AUC reflects the overall ability of the model to distinguish the correct context from the wrong one as the similarity threshold changes.

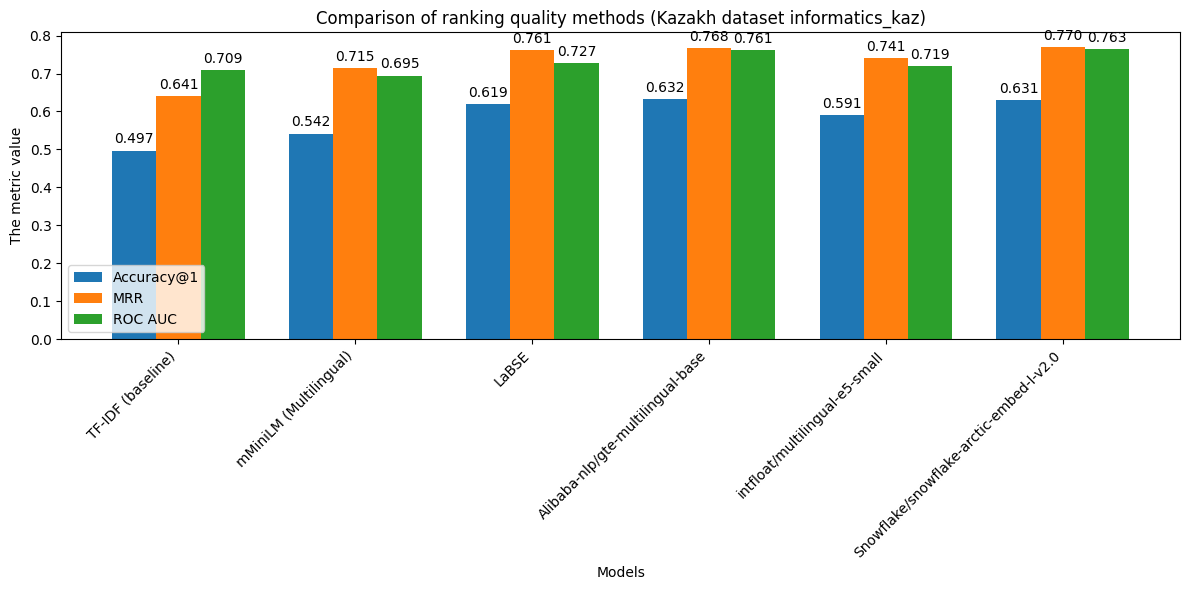
The experiments were conducted without further training the models on a given dataset, i.e. in zero-shot mode: it is assumed that the pre-trained embeddings are versatile enough to capture semantic proximity even for Kazakh questions that the model might not have previously seen. All texts of questions and contexts have been pre-normalized (lowercase, punctuation removal, tokenization if necessary) in order to be correctly represented in TF-IDF and to ensure compatibility with the requirements of the models (some transformers have entry length limits, but in our dataset the length of questions and contexts does not exceed these limits).

**Results and Discussion.** The results of the conducted experiment provide insights into the performance of different models in the task of matching questions and contexts in the Kazakh language. Specifically, we analyzed three key evaluation metrics - Accuracy@1, Mean Reciprocal Rank (MRR), and ROC AUC - across six models, including the baseline TF-IDF method and five modern multilingual sentence embedding models. The summarized results are presented in Table 1.

**Table 1**-**Comparison of models based on question–context matching accuracy**

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Accuracy@1** | **MRR** | **ROC AUC** |
| TF-IDF (baseline) | 0.4971 | 0.6407 | 0.7094 |
| mMiniLM (Multilingual) | 0.5419 | 0.7149 | 0.6947 |
| LaBSE | 0.6192 | 0.7614 | 0.7270 |
| Alibaba-nlp/gte-multilingual-base | 0.6322 | 0.7677 | 0.7611 |
| intfloat/multilingual-e5-small | 0.5913 | 0.7408 | 0.7188 |
| Snowflake/snowflake-arctic-embed-l-v2.0 | 0.6309 | 0.7700 | 0.7635 |

Figure 1 presents the values of all three metrics (Accuracy@1, MRR, and ROC AUC) for the six evaluated methods. As shown in the graph, the traditional TF-IDF method (on the left) yields the lowest performance, while more advanced multilingual embedding models - such as mMiniLM, LaBSE, GTE, E5, and Snowflake Arctic - consistently outperform it across all metrics. Among them, GTE-multilingual-base and Snowflake Arctic v2.0 demonstrate the highest scores, indicating their strong effectiveness in the task of ranking relevant question–context pairs, even in a low-resource language like Kazakh.

****

**Fig. 1 – Comparison of ranking quality methods**

The results clearly demonstrate the superiority of modern transformer-based embeddings over the classical TF-IDF approach. The baseline TF-IDF method correctly identifies the relevant context in approximately 49.7% of cases (Accuracy@1), while all tested embedding-based models outperform it on this metric. This is expected, as TF-IDF does not account for synonymy or contextual meaning and may fail when the question and answer use different phrasings for the same concept. Nevertheless, a nearly 50% result for TF-IDF suggests that simple lexical overlap still provides a baseline signal: in about half of the cases, the key terms in the question are directly present in the answer text.

The mMiniLM model achieves an Accuracy@1 of 0.5419, which is roughly 4.5 percentage points higher than the baseline. This improvement stems from its ability to capture semantic relationships: mMiniLM performs better than TF-IDF even without exact word matches between questions and contexts. Furthermore, its MRR (0.7149) is significantly higher than TF-IDF’s (0.6407), indicating that even when the top-ranked answer is incorrect, the correct one is often placed near the top (e.g., 2nd or 3rd). Interestingly, ROC AUC for mMiniLM (0.6947) is slightly lower than that of TF-IDF (0.7094), possibly due to the distribution of similarity scores: TF-IDF may more confidently assign very low scores to clearly irrelevant pairs, while mMiniLM, being semantically “softer,” sometimes assigns moderate scores even to unrelated pairs. As a result, its ROC curve may slightly underperform. Overall, mMiniLM delivers competitive performance for a lightweight multilingual model, but it understandably lags behind the larger models.

LaBSE significantly outperforms both TF-IDF and mMiniLM across all metrics: Accuracy@1 increases to 0.6192 (i.e., about 61.9% of questions are matched with the correct context in the top position), MRR reaches 0.7614, and ROC AUCrises to 0.7270. This improvement - over 12 percentage points higher than mMiniLM in Accuracy@1 - reflects the benefit of LaBSE’s extensive training on parallel multilingual data [3]. The model appears better equipped to interpret Kazakh formulations, likely because the Kazakh language was included in its training data and it has learned to align semantically equivalent phrases across languages. Notably, LaBSE also surpasses previous models in ROC AUC, demonstrating a more consistent ability to distinguish relevant from irrelevant pairs. Thus, LaBSE sets a strong benchmark in our experiment.

Two models - Alibaba GTE and Snowflake Arctic v2.0 - achieved nearly identical and highest overall performance. GTE reached an Accuracy@1 of 0.6322, while Arctic followed closely with 0.6309, making their performance statistically equivalent and slightly ahead of LaBSE. However, in terms of MRR and ROC AUC, both models slightly outperformed LaBSE: Snowflake Arctic reached an MRR of 0.7700 and ROC AUC of 0.7635, while GTE achieved MRR = 0.7677 and ROC AUC = 0.7611. These are the best results among all tested models. In nearly 63% of cases, the top-ranked context is correct, and on average, the correct answer appears very high in the list (MRR ~0.77 implies it is typically ranked first or second). A high ROC AUC (~0.76) indicates that the similarity scores effectively separate relevant from irrelevant context pairs.

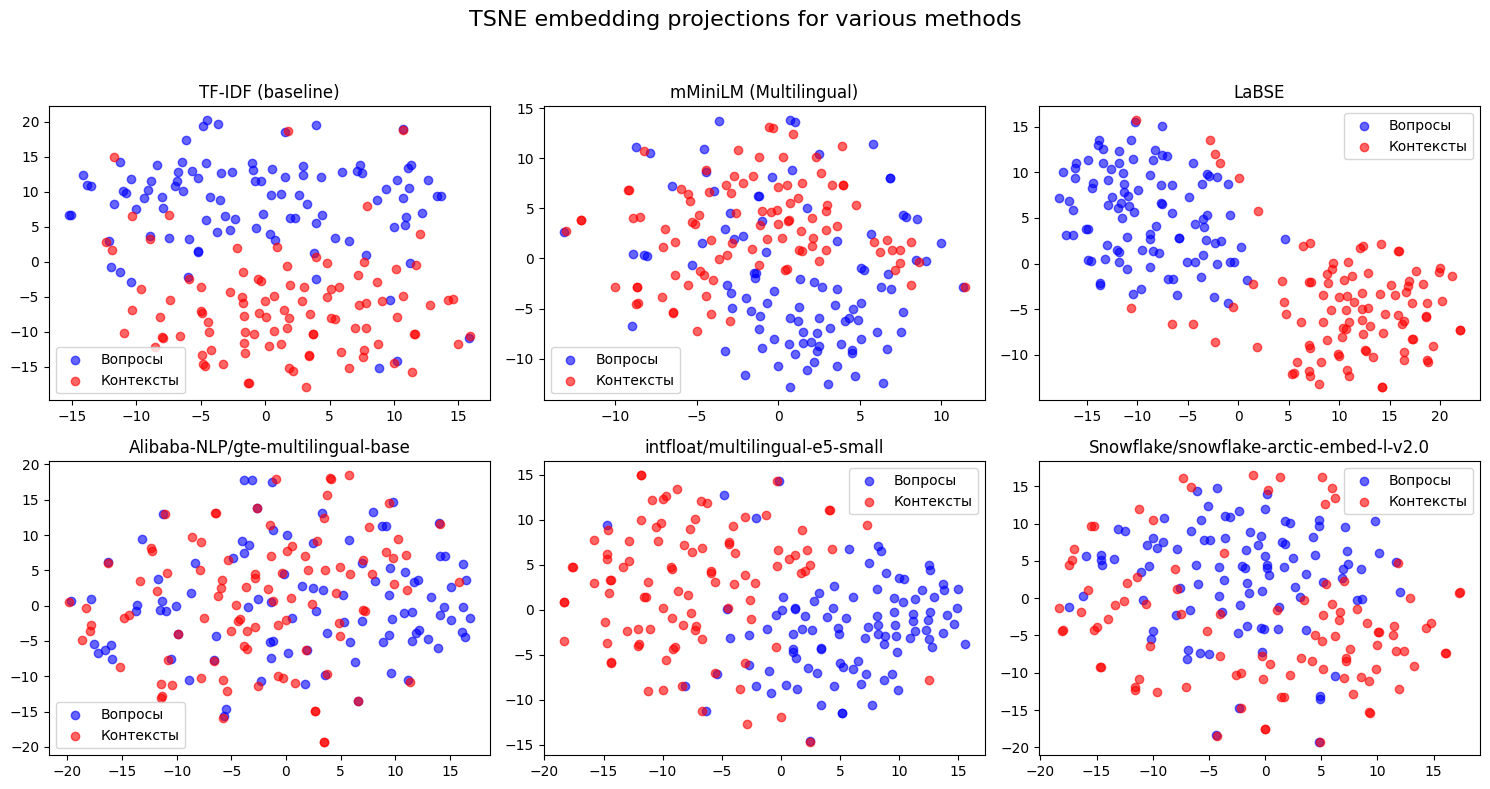
The success of GTE and Arctic can be attributed to their architectures and training data. GTE was specifically designed for document retrieval with long contexts and includes optimizations for multilingual input [10]. Its internal representations are likely more robust to linguistic variation, and it may better leverage token overlap (e.g., through XLM-R-based mixed tokenization, as discussed in [10]). According to its developers, Snowflake Arctic was built to achieve high-quality multilingual search without compromise - trained on query-document pairs in multiple languages, it captures both language-specific and cross-lingual semantic features [12]. Notably, Arctic v2.0 is promoted as a lightweight model that rivals larger architectures in retrieval quality [12], which is supported by our findings: it matches GTE’s performance despite architectural and training differences.

The multilingual E5-small model delivered moderate results: Accuracy@1 = 0.5913, higher than mMiniLM but lower than LaBSE. Its MRR (0.7408) and ROC AUC (0.7188) place it close to mMiniLM, slightly behind LaBSE in overall ranking quality. Considering that E5-small has around 118 million parameters [11] - more than mMiniLM but fewer than LaBSE - and was trained partly on English retrieval tasks, its relatively strong performance compared to mMiniLM is expected. However, it does not reach the levels of LaBSE or GTE/Arctic. This may be due to its smaller capacity and the possibility that its training data contained fewer or simpler Kazakh-language examples. Nonetheless, E5-small clearly surpasses the baseline and shows that even compact multilingual models can be effectively used for Kazakh semantic retrieval.

In summary, the best performance in the Kazakh question-context matching task was achieved by the multilingual embedding models LaBSE, GTE, and Snowflake Arctic, with Accuracy@1 scores in the 0.62–0.63 range (versus ~0.50 for TF-IDF) and MRR ≈ 0.76–0.77, indicating high-quality ranking. These results support the effectiveness of embeddings for building Kazakh-language QA systems. Importantly, this performance was achieved without fine-tuning— even in a zero-shot setting, pretrained models provided a substantial improvement over classical methods.

In the following section, we present a visual interpretation of how different models structure the semantic space using embedding projections.

Figure 2 illustrates the t-SNE projections of question and context embeddings produced by TF-IDF, mMiniLM, LaBSE, Alibaba GTE-multilingual-base, intfloat/multilingual-e5-small, and Snowflake Arctic Embed v2.0. Blue dots (or a contrasting color) represent question embeddings, while red dots indicate context embeddings. It is evident that the classical TF-IDF approach results in a sparse, disorganized cloud of points, whereas modern models like LaBSE, GTE, and Arctic tend to cluster semantically related questions and contexts closer together, which correlates with their superior ranking performance.



**Fig. 2 – t-SNE embedding projection for various methods**

**Conclusion.** Within the framework of this study, a comparative analysis of the effectiveness of various methods of comparing questions and contexts in the Kazakh language was carried out – from classical TF-IDF to the latest multilingual embedding models. The experimental results showed that the pre-trained transformer models are significantly superior to the basic TF-IDF approach in terms of the accuracy of extracting answers to questions. The best quality was demonstrated by the LaBSE, Alibaba GTE-multilingual, and Snowflake Arctic v2.0 models, which correctly extracted the appropriate context for ~63% of questions versus ~50% for TF-IDF. The LaBSE model confirmed the high efficiency of language-independent embeddings, previously noted in the works of Google [3], in relation to the Kazakh language. The GTE and Arctic models, relatively new developments in 2024, have shown the ability to achieve even higher metrics without specialized customization to Kazakh, which indicates progress in the field of multilingual semantic representations [4,12]. More compact models (MiniLM, E5-small) also gave a significant advantage over TF-IDF, although they were inferior in quality to the older models, which indicates the still important role of scale and training data for multilingual embeddings [2,9].

Visual analysis using t-SNE projections confirmed that successful models place semantically related questions and answers close to each other in a vector space, whereas less accurate methods show a more chaotic or disjointed distribution. The difference between the discrete nature of TF-IDF and continuous transformer spaces was particularly evident: modern models create structured embedding clouds that are more suitable for searching query–document pairs.

The practical significance of this work lies in the fact that even without additional training on specialized Kazakh data, pre-trained multilingual models can be used for question-and-answer search tasks. To build an educational AI system that answers students' questions in Kazakh, it is most promising to use models such as LaBSE or GTE/Arctic, which provide the highest quality of comparison. These models can be integrated into e-learning systems and reference systems, where the student asks a question in natural language, and the system finds an appropriate explanation or definition in the knowledge base.

It should be noted that there is room for improvement: although ~63% accuracy is a good result, in the remaining cases the model did not guess the answer on the first attempt. In the future, it is planned to explore the methods of further training (fine-tuning) of the considered models directly on the Kazakh Q&A corpus, as well as try larger versions of the models (for example, multilingual-E5-large or a larger version of GTE) in order to achieve even higher quality. Another direction could be a combination of approaches, for example, using embeddings to initially select candidates and then accurately match them using a cross-encoder. Nevertheless, the results already obtained emphasize the power of modern multilingual models: even for a relatively low-resource language, they can significantly improve the quality of semantic search, which opens up new opportunities for the development of educational and information systems in the Kazakh language.

**References**

1. Li Z., Zhang X., Zhang Y., Long D., Xie P., Zhang M. Towards General Text Embeddings with Multi-stage Contrastive Learning // arXiv preprint. 2023. arXiv:2308.03281. 14 p. DOI  [10.48550/arXiv.2308.03281](https://doi.org/10.48550/arXiv.2308.03281).

2.Reimers N., Gurevych I. Making monolingual sentence embeddings multilingual using knowledge distillation // Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020, P.4512-4525. DOI 10.48550/arXiv.2004.098133.Devlin J., Chang M.W., Lee K., Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding // Proceedings of NAACL-HLT.- 2019.- Vol.1. P. 4171–4186. DOI 10.48550/arXiv.1810.04805.

4. Сross-encoder. mmarco-mMiniLMv2-L12-H384-v1 [Electronic resource] // Hugging Face. 2024. URL: <https://huggingface.co/cross-encoder/mmarco-mMiniLMv2-L12-H384-v1>.- Date of address:16.04.2025.

5. Feng F., Yang Y., Cer D., Arivazhagan N., Wang W. Language-agnostic BERT sentence embedding // arXiv preprint. 2020. arXiv:2007.01852. DOI 10.48550/arXiv.2007.01852.

6. Tleubayeva A., Mitroshina A., Arman A., Shokan A., Aigul S. Systemic approach to optimizing natural language processing technologies in Astana IT University's admissions process // Proceedings of the 2024 IEEE 4th International Conference on Smart Information Systems and Technologies (SIST). Astana, Kazakhstan. 2024. -P. 520-525.

DOI 10.1109/SIST61555.2024.10629481.

7. Maksutova K. Kazakh QA dataset Kundyzka/informatics\_kaz [Electronic resource].2023. Available at: <https://huggingface.co/datasets/Kundyzka/informatics_kaz>.-Date of address:10.04.2025.

8. Tleubayeva A., Shomanov A. Comparative analysis of multilingual QA models and their adaptation to the Kazakh language // Scientific Journal of Astana IT University.-2024.- Vol.19. P. 89-97. DOI 10.37943/19WHRK2878.

9. Wang H. Automatic question-answering modeling in English by integrating TF-IDF and segmentation algorithms//Systems and Soft Computing.-2024.-Vol.6:200087. DOI [10.1016/j.sasc.2024.200087](https://doi.org/10.1016/j.sasc.2024.200087).

10. Zhang X., Zhang Y., Long D., et al. mGTE: Generalized long-context text representation and reranking models for multilingual text retrieval//arXiv preprint.- 2024:2407.19669. DOI [10.48550/arXiv.2407.19669](https://doi.org/10.48550/arXiv.2407.19669)

11. Wang L., Yang N., Huang X., Yang L., Majumder R., Wei F. Multilingual E5 text embeddings: a technical report//arXiv preprint.- 2024:2402.05672. DOI 10.48550/arXiv.2402.05672.

12. Yu P., Merrick L., Nuti G., Campos D. Arctic-Embed 2.0:Multilingual retrieval without compromise//arXiv preprint. 2024. arXiv:2412.04506.DOI 10.48550/arXiv.2412.04506.

13. Susanti F., Maulidevi N. U., Surendro K. Improving embedding-based link prediction performance using clustering //Journal of King Saud University-Computer and Information Sciences.- 2024.-Vol. 36( 8): 102181. DOI [10.1016/j.jksuci.2024.102181](https://doi.org/10.1016/j.jksuci.2024.102181).

14. Liu Y., Li Y., Xie D. Implications of imbalanced datasets for empirical ROC-AUC estimation in binary classification tasks // Journal of Statistical Computation and Simulation.- 2023.- Vol. 94. (1). - P. 183-203. DOI 10.1080/00949655.2023.2238235.

***Сведения об авторах***

Мажитова К. Ш. - магистр технических наук, преподаватель специальных дисциплин Высшего колледжа «ASTANA POLYTECHNIC», Астана, Казахстан, e-mail: [mazhitovaks@gmail.com](mailto:mazhitovaks@gmail.com);

Тлеубаева А. О.- докторант ТОО «Astana IT University», Астана, Казахстан, e-mail: [Arailym.tll@gmail.com](mailto:Arailym.tll@gmail.com);

Мухаммедия С.Е.- магистр технических наук, преподаватель специальных дисциплин Высшего колледжа «ASTANA POLYTECHNIC», Астана, Казахстан, e-mail: [m.samal219@gmail.com](mailto:m.samal219@gmail.com);

Танирбергенова А.А.- доктор PhD, кафедра системного анализа и управления ЕНУ им.Л.Н.Гумилева, Астана, Казахстан, e-mail: [alua\_15\_1982@mail.ru](mailto:alua_15_1982@mail.ru);  
Токтаганова А.С.- преподаватель специальных дисциплин Высшего колледжа «ASTANA POLYTECHNIC», Астана, Казахстан, e-mail: [ainur\_7t@bk.ru](mailto:ainur_7t@bk.ru);

Камбарова Ж. Б.- преподаватель специальных дисциплин Высшего колледжа «ASTANA POLYTECHNIC», Астана, Казахстан, e-mail: [Kambarova\_1976@mail.ru](mailto:Kambarova_1976@mail.ru).

***Information about the authors***

Mazhitova K.Sh. - Master of Technical Sciences, lecturer of special disciplines at the Higher College "ASTANA POLYTECHNIC", Astana, Kazakhstan, e-mail: [mazhitovaks@gmail.com](mailto:mazhitovaks@gmail.com);

Tleubayeva A.О. - doctoral student at Astana IT University LLP, Astana, Kazakhstan, e-mail: Arailym.tll@gmail.com;

Mukhammediya S.E.- Master of Technical Sciences, lecturer of special disciplines at the Higher College "ASTANA POLYTECHNIC", Astana, Kazakhstan, e-mail: [m.samal219@gmail.com](mailto:m.samal219@gmail.com);

Tanirbergenova A.A.- PhD, Department of System Analysis and Management, L.N.Gumilyov ENU, Astana, Kazakhstan, Astana, e-mail: [alua\_15\_1982@mail.ru](mailto:alua_15_1982@mail.ru);

Toktaganova A.S.- is a teacher of special disciplines at the Higher College "ASTANA POLYTECHNIC", Astana, Kazakhstan, e-mail: ainur\_7t@bk.ru;

Kambarova Zh. B.- is a teacher of special disciplines at the Higher College "ASTANA POLYTECHNIC", Astana, Kazakhstan, e-mail: Kambarova\_1976@mail.ru.