УДК 00.1082

Maksat Kalimoldayev*[1], Madina Mansurova[2], 2026.
[1]Institute of Information and Computer Technologies National Academy of Sciences of the Republic of Kazakhstan, Almaty, Kazakhstan
[2]Al-Farabi Kazakh National University, Almaty, Kazakhstan
**E-mail:mnk@ipic.kz*

ARTIFICIAL INTELLIGENCE DEVELOPMENT IN KAZAKHSTAN OVER THE PAST DECADE: A COMPREHENSIVE REVIEW

Kalimoldaev Maksat, Doctor of Physics and Mathematics, Professor, Honorary Academician of the National Academy of Sciences of the Republic of Kazakhstan under the President of the Republic of Kazakhstan, Advisor to the Director General, Head of the Laboratory, Institute of Information and Computer Technologies, Almaty, Kazakhstan.
E-mail: mnk@ipic.kz; https://orcid.org/0000-0003-0025-8880

Mansurova Madina, head of the Department of artificial intelligence and big data, professor of the Al-Farabi Kazakh National University.
E-mail: madina.mansurova@kaznu.edu.kz ; https://orcid.org/ 0000-0002-9680-2758

Abstract. Artificial intelligence's influence on society has never been more pronounced. Often described as a new wave of industrial transformation following the internet, AI is reshaping economies, public services, and everyday life worldwide. Positioned at the crossroads of Eurasia, Kazakhstan has set an ambitious goal to leverage AI as a core enabler of a fully digitalized society. In this context, this paper provides a comprehensive, evidence-oriented review of the current state of AI development in Kazakhstan, informed by global trends in expanding AI capabilities, declining deployment costs, rising adoption, and growing emphasis on governance and responsible use. The analysis is aligned with Kazakhstan's national AI development concept for 2024-2029, which prioritizes data management, infrastructure development, human capital, research and development including language technologies, and regulatory frameworks, and it uses these priorities to structure the evidence and highlight practical pathways for implementation and measurable progress.

Keywords: artificial intelligence, digital society, research, technology, infrastructure.
Conflict of interest: The authors declare that there is no conflict of interest.

М.Н. Калимолдаев[1]*, М.Е. Мансурова[2], 2026.
[1] Ақпараттық және есептеу технологиялар институты, Алматы, Қазақстан
[2]Әл-Фараби атындағы Қазақ ұлттық университеті, Алматы, Қазақстан
**E-mail:mnk@ipic.kz*

СОҢҒЫ ОН ЖЫЛДЫҚТА ҚАЗАҚСТАНДА ЖАСАНДЫ ИНТЕЛЛЕКТІҢ ДАМЫТУЫ: КЕШЕНДІ ШОЛУ

Мақсат Нұрәділұлы Калимолдаев, профессор, ф.-м. ғ. д., Ақпараттық және есептеу технологиялар институтының Бас директорының кеңесшісі, зертхана меңгерушісі, Қазақстан Республикасы Ұлттық ғылым академиясының құрметті академигі.
E-mail:mnk@ipic.kz; ttps://orcid.org/0000-0003-0025-8880

Мадина Есімханқызы Мансурова, Жасанды интеллект және үлкен деректер кафедрасының меңгерушісі, Әл-Фараби атындағы Қазақ ұлттық университетінің профессоры.
E-mail: madina.mansurova@kaznu.edu.kz; https://orcid.org/ 0000-0002-9680-2758

Аннотация. Жасанды интеллекттің қоғамға әсері ешқашан байқалмады. Жасанды интеллект көбінесе Интернеттен кейінгі өнеркәсіптік өзгерістердің жаңа толқыны деп аталады, бүкіл әлемдегі экономиканы, мемлекеттік қызметтерді және күнделікті өмірді өзгертеді. Еуразия қиылысында орналасқан Қазақстан өзінің алдына жасанды интеллектті (ЖИ) толық цифрлық қоғам құрудың негізгі құралы ретінде пайдалану мақсатын қойды. Осы тұрғыда бұл құжат Қазақстандағы ЖИ дамуының ағымдағы жай-күйіне жан-жақты, нақты деректерге бағдарланған шолу болып табылады, ол ЖИ мүмкіндіктерін кеңейтудегі жаһандық үрдістерге, енгізу шығындарын төмендетуге, енгізуді ұлғайтуға және басқаруға және жауапты пайдалануға өсіп келе жатқан назарға негізделген. Талдау Қазақстанның жасанды интеллектін дамытудың 2024-2029 жылдарға арналған ұлттық тұжырымдамасымен келісілген, онда деректерді басқару, инфрақұрылымды дамыту, адами капитал, тілдік технологияларды қоса алғанда, зерттеулер мен әзірлемелер және нормативтік-құқықтық база басым болып табылады және бұл басымдықтар нақты деректерді құрылымдау және енгізудің практикалық жолдары мен өлшенетін прогресті айқындау үшін пайдаланылады.

Түйін сөздер: жасанды интеллект, цифрлық қоғам, зерттеу, технология, инфрақұрылым.

Мүдделер қақтығысы: авторлар мүдделер қақтығысының жоқтығын мәлімдейді.

М.Н. Калимолдаев*[1], М.Мансурова[2], 2026.
[1]Институт информационных и компьютерных технологий, Алматы, Казахстан
[2]Казахский национальный университет имени аль-Фараби, Алматы, Казахстан

РАЗВИТИЕ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА В КАЗАХСТАНЕ ЗА ПОСЛЕДНЕЕ ДЕСЯТИЛЕТИЕ: ВСЕСТОРОННИЙ ОБЗОР

Калимолдаев Мақсат Нұрәділұлы, д. ф.-м.н., профессор, почетный академик НАН РК при Президенте |РК, советник Генерального директора, заведующий лабораторией, Институт информационных и компьютерных технологий, Алматы, Казахстан.
E-mail:mnk@ipic.kz; https://orcid.org/ 0000-0003-0025-8880

Мансурова Мадина Есімханқызы, к.ф.-м.н., профессор, заведующая кафедрой искусственного интеллекта и больших данных, Казахский национальный университет имени аль-Фараби, Алматы, Казахстан.
E-mail: madina.mansurova@kaznu.edu.kz ; https://orcid.org/ 0000-0002-9680-2758

Аннотация. Влияние искусственного интеллекта на общество еще никогда не было таким заметным. Искусственный интеллект, который часто называют новой волной промышленных преобразований, происходящих вслед за Интернетом, меняет экономику, общественные услуги и повседневную жизнь по всему миру. Расположенный на перекрестке Евразии, Казахстан поставил перед собой амбициозную цель – использовать искусственный интеллект (ИИ) в качестве основного средства создания полностью цифрового общества. В этом контексте данный документ представляет собой всеобъемлющий, ориентированный на фактические данные обзор текущего состояния развития ИИ в Казахстане, основанный на глобальных тенденциях в расширении возможностей ИИ, снижении затрат на внедрение, росте внедрения и растущем внимании к управлению и ответственному использованию.

Анализ согласован с Национальной концепцией развития искусственного интеллекта Казахстана на 2024-2029 годы, в которой приоритетными являются управление данными, развитие инфраструктуры, человеческий капитал, исследования и разработки, включая языковые технологии, и нормативно-правовая база, и эти

приоритеты используются для структурирования фактических данных и определения практических путей внедрения и измеримого прогресса.

Ключевые слова: искусственный интеллект, цифровое общество, исследование, технологии, инфраструктура.

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов.

Introduction

Artificial intelligence (AI) has progressed from early conceptual foundations, such as McCulloch and Pitts' neuron-inspired model [1] and Turing's proposal for evaluating machine intelligence [2], into today's data-driven era enabled by backpropagation [3], deep convolutional networks [4], and transformer architectures [5] that underpin modern foundation models. This transition has moved AI from research prototypes to large-scale public adoption with unprecedented speed. Widely reported adoption figures illustrate this shift: ChatGPT reached 100 million users in roughly two months, whereas YouTube required about 1.5 years to reach a similar milestone [6], highlighting how rapidly generative AI systems can diffuse once the enabling infrastructure and product interfaces mature.

Similarly, AI research in Kazakhstan is increasingly aligned with the global mainstream. The local ecosystem has moved from primarily using widely adopted transformer-based methods to developing nationally relevant large language models, including KazLLM [7] and the AlemLLM [8], alongside a growing body of applied research. This growth is also reflected in publication volume: Kazakhstan's AI-related research output has expanded sharply over the past two decades, rising from single-digit annual counts around 2000 to well over a thousand papers by 2023 (Fig. 1). In parallel, Kazakhstan has articulated an ambition to position AI as a key component of a fully digitalized society [9].

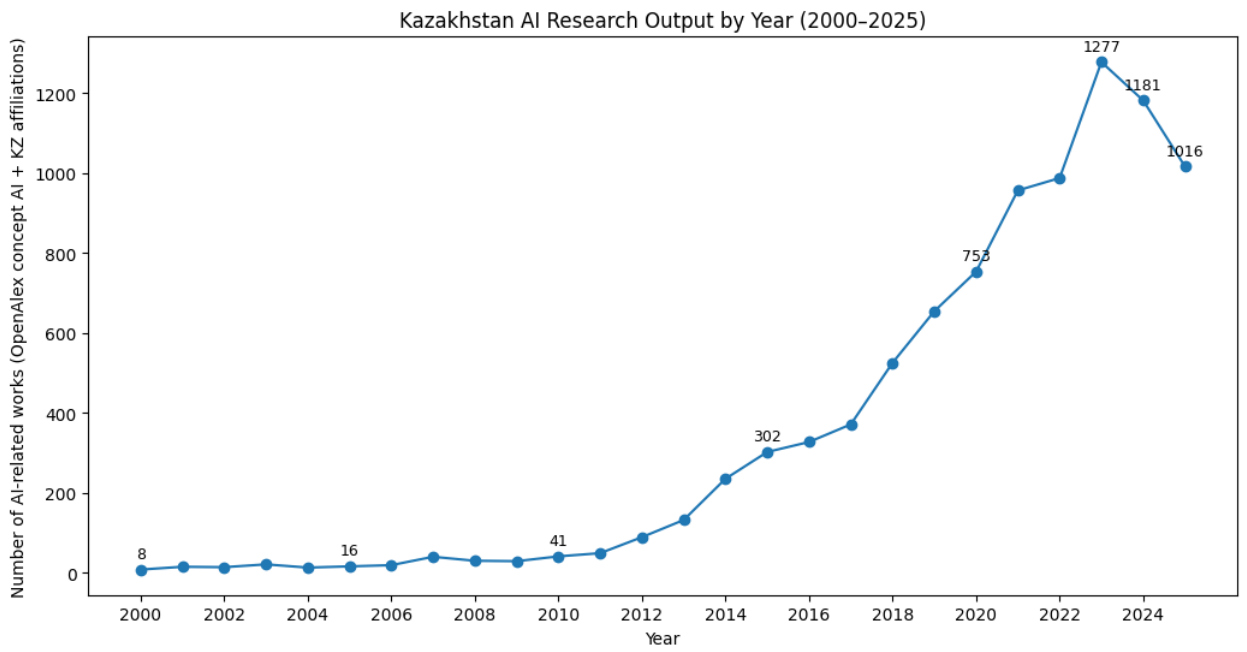

Fig. 1. Kazakhstan AI research output by year (2000–2025), measured as the annual count of AI-related works in OpenAlex with at least one author affiliated with an institution in Kazakhstan

Beyond research, AI is already being operationalized across major economic sectors, with a clear emphasis on measurable industrial impact. For example, Samruk-Kazyna reports implementing 62 AI-based projects, with a primary focus on production processes and an expected cumulative effect exceeding $1.3 billion over five years [10]. For 2026, portfolio companies have been assigned a KPI to increase EBITDA by 5% through the use of AI, and within two years it is planned to transition toward a model in which 70% of management decisions are

made with AI participation. Sectoral examples include KazMunayGas projects such as ABAI, where AI supports reservoir flooding management and recommends reservoir pressure actions to increase oil production, with a projected economic effect of about 326 billion tenge over 2025 to 2030 [11]. Additional initiatives include AI-based forecasting of deficits and surpluses of petroleum products to stabilize regional supply, with an expected economic effect of 22.5 billion tenge. In the energy sector, predictive defect detection systems are being introduced at GRES-1 and AlES to improve equipment reliability and reduce downtime, with a projected economic effect of about 36 billion tenge for 2026 to 2030. AI is also being used to forecast renewable energy generation based on internal meteorological services.

These deployments are supported by expanding compute and data infrastructure and by a large-scale human capital agenda. Samruk-Kazyna indicates that its AI projects are implemented within a closed circuit on the Al FARABIUM supercomputer operated by Kazakhtelecom, with a portion of capacity used by the group and the remainder leased to commercial customers, including abroad [12]. Parallel to this, market projections suggest growth in Kazakhstan's data centre sector, with revenue expected to rise toward roughly $417 million by 2028 [13]. On the workforce side, Kazakhstan is scaling AI literacy through the AI Movement initiative, reporting more than 400,000 people trained through programs such as AI-Sana, AI-Qyzmet, and AI-People, and launching AI-Corporate for major state holdings. The national goal is to train 1 million citizens in five years. In collaboration with the Ministry of Education, Day of AI content has also been introduced for primary school grades 1 to 4, with plans to expand across all grade levels [14]. AI-related applications are also emerging in construction through unified digital platforms for planning and tracking, and in agriculture and water management through satellite monitoring and digital optimization tools.

Motivated by this combination of rapid adoption, expanding research capacity, and cross-sector deployment, this paper reviews current AI development trends in Kazakhstan and discusses the country's emerging large language model projects, focusing on ecosystem capabilities, constraints, and near-term directions for research and implementation.

Overview of current LLM trends

As illustrated in Table 1, modern large language models (LLMs) have demonstrated rapid performance gains alongside an unprecedented rise in computational demands [15]. This surge is driven by increasing model scales and the heightened complexity of training and deployment. We examine the evolution of these requirements across three key dimensions: model development, dataset expansion, and application-level requirements.

Table 1: Recent, vendor-reported benchmark results for popular LLMs across widely used evaluations (MMLU, GPQA, HumanEval, GSM8K, MATH, SWE-bench Verified, MMMU)

Model	MMLU	GPQA (Diamond)	HumanEval	GSM8K	MATH	SWE-bench Verified	MMMU
Claude 3.5 Sonnet	88.7	59.4	92.0	96.4	71.1	-	68.3
GPT-4.1	90.2	66.3	-	-	-	54.6	-
GPT-4o	88.7	53.6	90.2	-	76.6	33.2	69.1
Llama 3.1 405B Instruct	87.3	50.7	89.0	96.8	73.8	-	-
Gemini 2.5 Pro	-	-	-	-	-	63.8	-

Training costs have increased dramatically shown in Table 2, rising from less than one thousand dollars for early models such as the 2017 Transformer to tens or even hundreds of millions for state-of-the-art systems [16]. Despite continuous efficiency improvements, overall

resource needs remain high because frontier training increasingly relies on longer training runs, higher-quality data, larger context windows, and more complex training objectives, for example, the multi-stage post-training, tool-use, and multimodal alignment. Contemporary training typically unfolds in two phases: a pretraining stage using vast corpora that can exceed one trillion tokens, followed by instruction tuning, in many cases, preference optimization to align the model with downstream tasks and user intent [15]. This tuning process depends on large volumes of curated prompts and expert-annotated data, adding additional compute and data-engineering cost.

Table 2: The demand for computing resources is growing rapidly

Model	Parameters (B)	Training Tokens (T)	Compute Cost ($M)	GPU Count (est.)	Year	Notable Features
BERT-Large	0.34	0.003	<1	<100	2018	Baseline transformer
GPT-3	175	0.3	10–20	10,000	2020	Few-shot learning
PaLM	540	0.78	Unknown	6,144 TPUs	2022	Massive scale, Pathways system
Chinchilla	70	1.4	Unknown	Unknown	2022	Data-efficiency focus
PaLM 2	340	3.6	Unknown	Unknown	2023	Compute-optimal design
GPT-4 (est.)	1700	Unknown				

At the same time, the field is shifting from “bigger models only” toward a more nuanced scaling strategy: compute-optimal training, data-quality and system-level scaling [18]. Multilingual and domain-specific modeling increases dataset diversity requirements, while multimodal training compounds cost by adding vision encoders, richer inputs, and heavier preprocessing. As a result, data pipelines have become a core competitive advantage: large-scale training increasingly depends on efficient tokenization, sharding, streaming, dynamic batching, and careful dataset mixing to maintain stable training and avoid wasted compute.

Once the challenges associated with large model architectures are addressed, the next critical focus shifts to ensuring data efficiency. Table 3 provides an overview of the number of tokens used during the pretraining phase of various large language models. From an infrastructure perspective, managing training with datasets at the scale of trillions of tokens requires highly optimized data pipelines [19]. This includes the use of memory-mapped files, efficient tokenization, dynamic shuffling, and parallel data loaders. Even the process of reading and streaming such massive datasets consumes a substantial amount of GPU time. As training objectives grow more diverse and context-aware, the demands on preprocessing and I/O systems continue to increase accordingly.

Table 3 The number of tokens used in pre-training stage of popular LLMs

MODEL	TOKENS (B)
BERT-LARGE	3.0 billion
GPT-3	300.0 billion
PALM	780.0 billion
CHINCHILLA	1400.0 billion
PALM 2	3600.0 billion
GPT-4 (EST.)	Unknown
GEMINI ULTRA (EST.)	Unknown

Inference has emerged as a dominant driver of computational and economic constraints in modern LLM deployments. In many production settings, the marginal cost of serving large models at scale can rival or exceed the original training expenditure, particularly for applications with large user populations and stringent latency requirements [20]. This has intensified the focus on inference efficiency through model- and system-level techniques, including smaller and task-specialized models, low-precision quantization (for example 8-bit and 4-bit), knowledge distillation, kernel and runtime optimizations, batching strategies, speculative decoding, and memory-efficient attention mechanisms. In parallel, contemporary deployments increasingly favor compound architectures over single-pass model invocation. Retrieval-Augmented Generation pipelines typically integrate retrieval, optional re-ranking, and generation. Tool-augmented agentic systems introduce function calling, iterative planning loops, and external API interactions. Safety layers add moderation, policy enforcement, and compliance checks. Each added component increases end-to-end compute per query, making holistic pipeline optimization and robust GPU scheduling central to reliable and cost-effective deployment.

Post-training alignment has become a core stage of the LLM lifecycle. Methods such as reinforcement learning from human feedback [21] and preference optimization [22], including direct preference optimization variants, improve helpfulness, controllability, and safety. However, they introduce additional iterative training cycles, auxiliary models such as reward or preference models, and substantial evaluation infrastructure. Governance and responsible deployment have also shifted from broad principles to operational constraints. As a result, organizations invest in standardized evaluation suites, red-teaming protocols, dataset documentation, privacy-preserving controls, and traceability mechanisms. The field's trajectory is no longer defined solely by scaling parameter counts. It increasingly reflects the scaling of full systems, including data pipelines, context length, multimodal inputs, tool use, and rigorous alignment and evaluation, under strong pressure to reduce both training and inference costs.

Beyond training-centric scaling, several broader trends have shaped the LLM landscape and help explain the rapid acceleration of adoption. Deep learning moved from research to mainstream production as advances in representation learning and GPU availability enabled strong performance in vision, speech, and recommendation systems before LLMs became dominant. This period consolidated the foundation-model paradigm, in which large pretrained models are adapted to many tasks through prompting, fine-tuning, or parameter-efficient adapters, reducing the need to train narrowly specialized models from scratch. Open-source ecosystems also expanded rapidly, lowering barriers by releasing checkpoints, training recipes, and optimized inference runtimes. This enabled universities and mid-sized organizations to experiment with modern architectures under limited resources. Cloud computing and managed AI platforms further simplified access to accelerators and deployment via APIs, while simultaneously reinforcing hybrid strategies in which sensitive data, regulation, and governance constraints motivate on-premise or sovereign deployments.

A further shift has been the move from language modeling as an isolated capability to system-level intelligence as a deployment reality. In practice, LLMs increasingly serve as orchestrators embedded within larger workflows [23]. They call tools, query databases, retrieve and synthesize documents, generate structured outputs, and interact with enterprise software. This agentic direction has been strengthened by advances in function calling, structured prompting, and planning-oriented methods, as well as by the widespread adoption of Retrieval-Augmented Generation for grounding outputs in verifiable knowledge. In parallel, multimodality has become increasingly central, as modern systems combine text with images, audio, and video to support richer interfaces such as voice assistants, document understanding, and multimodal search. Context windows have expanded to support long-document processing and multi-step tasks, but this increases memory pressure and latency, motivating continued research in efficient attention, long-context training, and inference optimization.

In deployment practice, cost, reliability, and privacy requirements have driven a bifurcation of the model ecosystem. On one side are smaller enterprise and edge models optimized for latency, operational cost, and data locality. On the other are frontier-scale models

optimized for maximal capability. This bifurcation is sustained by rapid progress in compression and efficiency methods, including quantization, distillation, sparsity, kernel fusion, and hardware-aware compilation, as well as architectural approaches such as mixture-of-experts that can increase capability without proportionally increasing inference cost for every token. At the same time, evaluation has become a first-class engineering requirement. Organizations increasingly rely on standardized benchmarks, domain-specific test suites, and continuous monitoring to detect regressions, hallucinations, bias, and safety failures. Responsible AI has also matured into operational practice, with stronger emphasis on data provenance, privacy protection, documentation, auditability, and risk management, particularly in regulated domains including healthcare, finance, and public services.

Against this global backdrop, Kazakhstan's LLM ecosystem can be characterized by increasing visibility, openness, and end-to-end completeness. A notable shift has occurred from research outputs remaining confined to internal reports or closed pilots toward publishing models, datasets, and evaluation artifacts in open ecosystems. This transition is consequential because it transforms isolated efforts into a cumulative ecosystem. Shared baselines, replicable results, and accessible artifacts enable faster iteration and more credible scientific comparison across institutions and projects.

One representative entry point is the machine learning community, the repository of kz-transformers [24] on Hugging Face is one example, which functions as an open repository of checkpoints, datasets, and demonstration spaces that lower the friction of experimentation and replication. The output is not restricted to a single direction. It includes foundational components, practical utilities, and benchmarking-oriented resources that collectively support incremental community progress. Such public, iterative release practices establish shared reference points that future work can systematically improve upon.

Building on this community-level momentum, institutional-scale efforts provide evidence of capacity to manage the full lifecycle of foundation-model development. In this context, ISSAI's KazLLM [7] is notable not merely as a larger model release, but as an indicator of growing capability to curate and clean large corpora, coordinate compute-intensive training, and package artifacts for public use. The significance lies in what this infrastructure enables next. Once a national-scale model exists, research can accelerate toward systematic evaluation, alignment, domain specialization, and application development without repeatedly reconstructing the entire pipeline.

As artifacts proliferate, the credibility of progress increasingly depends on evaluation practices that are transparent and reproducible [25]. The growing attention to benchmarks, leaderboards, and structured evaluation spaces is therefore a critical signal of ecosystem maturation. By emphasizing comparative measurement and repeatable testing, these efforts align Kazakhstan's trajectory with global norms in which models are accompanied by evaluation infrastructure that makes claims scientifically interpretable and practically actionable.

Finally, Kazakhstan's trajectory should be situated within the broader international research network that shapes contemporary AI. This framing reinforces the interpretation that Kazakhstan's ecosystem is increasingly integrated into global flows of methods, tooling, and collaboration. Overall, the emerging picture is not only of increased model production, but of a more complete modern AI pipeline spanning data, training, evaluation, multimodality, and deployment practices.

Project context and implementation case

After outlining global AI developments and Kazakhstan's national priorities, we introduce the program-targeted project "Creating a Large Language Model (LLM) to Support the Kazakh Language and Technological Progress" as a concrete case study. Designed as a three-year effort, the project aims to develop a modern LLM that strengthens Kazakh as the state language and a language of intercultural communication, while also supporting technological innovation, data security, education, and scientific research in Kazakhstan.

In general, progress in model capability is driven by three interacting factors: data, model scale and architecture, and algorithms. For Kazakh, data is the dominant constraint. As a low-

resource language, Kazakh occupies only a tiny fraction of online text, and until recently it was barely represented in the pretraining corpora of many widely used foundation models such as LLAMA 2 [26] shown in Figure 2.

Language	Percent	Language	Percent
en	89.70%	uk	0.07%
unknown	8.38%	ko	0.06%
de	0.17%	ca	0.04%
fr	0.16%	sr	0.04%
sv	0.15%	id	0.03%
zh	0.13%	cs	0.03%
es	0.13%	fi	0.03%
ru	0.13%	hu	0.03%
nl	0.12%	no	0.03%
it	0.11%	ro	0.03%
ja	0.10%	bg	0.02%
pl	0.09%	da	0.02%
pt	0.09%	sl	0.01%
vi	0.08%	hr	0.01%

Figure 2: Language distribution in pretraining data with percentage >= 0.005% of meta LLAMA 2 in 2024.

As a result shown in table 4, many general-purpose LLMs show limited Kazakh understanding out of the box and require additional adaptation, such as targeted fine-tuning and instruction tuning, to perform reliably in Kazakh. This motivates the project's emphasis on building high-quality training data and ensuring that Kazakh is not treated as a marginal component in the model's learning process.

Table 4. Comparative performance and adaptability of popular LLMs for Kazakh language tasks.

Issue	**GPT-4 (OpenAI)**	**LLaMA 3 (Meta AI)**	**DeepSeek**	**TinyLlama**
Data scarcity (number of Kazakh tokens in training)	1.5B tokens (Kazakh < 0.1% of data)	600M tokens (limited support)	400M tokens (primary language: Chinese)	100M tokens
Translation quality (BLEU score for Kazakh)	32.1%	25.4%	20.2%	18.5%
Morphology and agglutination (suffix recognition accuracy, %)	78.2%	68.4%	65.1%	59.3%
Answer reliability (hallucination rate on factual data, %)	9.3%	15.7%	18.9%	22.5%
Customization availability (fine-tuning capability, % of full dataset)	0% (closed model)	100%	100%	100%

The project therefore focuses on expanding both the quantity and the quality of Kazakh-language resources. Existing assets such as the Kazakh National Corpus [28] provide an important foundation, but many of these resources were created for linguistic research and were not

designed for modern LLM training workflows. Today's LLM training typically involves two stages. The first is pretraining on large-scale, mostly unlabeled text to learn general language representations. The second is instruction tuning, where the model learns to follow user intent through instruction–response pairs, similar to teaching a student through examples. This second stage is especially important for real-world usefulness, because it trains the model not only to generate fluent text, but also to respond in a helpful, task-oriented way that matches everyday language use.

To produce instruction data that reflects real Kazakh usage rather than artificial patterns generated by models, the project relies on expert curation. In practice, this means that linguists and language specialists design instruction–response pairs that mirror natural Kazakh phrasing, cultural context, and expected answer styles. This work is supported by consortium partners, including the Akhmet Baitursynov Institute of Language Education and the Shaisultan Shayakhmetov National Scientific and Practical Center "Til-Qazyna." In parallel, the project scales unlabeled data through large document digitization. Because Al-Farabi Kazakh National University maintains one of the largest university libraries in Central Asia, the project has digitized thousands of books using OCR [28] and is also processing large volumes of news and multimodal content. Both expert curation and large-scale digitization are time-consuming, but they create durable national assets that can support research and applications over the next decade.

If data is the fuel of AI, then the model is its engine. While today's frontier systems show improving support for low-resource languages, the situation was notably weaker only a few years ago, and relying solely on external models is not sufficient for strategic goals such as data sovereignty, education, and public-sector deployment. For this reason, the project trained a compact LLaMA 3 variant with 1.9 billion parameters using two 100 GB GPUs. The model was trained on a curated dataset compiled from KazNU's dissertation archives and enriched with instruction data authored by Kazakh language experts. This experience shows that even "compact" LLM development requires substantial computational resources when the goal is high-quality, domain-relevant performance.

Beyond data and base model training, algorithm has become a third critical pillar such as alignment [29]. A capable model with strong data coverage still needs reliable control so that its responses follow human intent, remain safe, and reflect social and cultural expectations . In modern LLM pipelines, this control layer is often built through reinforcement learning from human feedback [21] or related preference-optimization methods [22]. In the context of Kazakhstan, alignment is also about ensuring that the model's behavior matches local norms, educational requirements, and public-sector values. The human experts process reinforcement learning from human feedback is shown in Figure 3.

Fig 3. Reinforcement learning from human feedback.

After addressing alignment, the next major challenge is hallucination. In practice, hallucination occurs when a model produces fluent but unsupported statements, including

information that is inaccurate or does not exist. This behavior is strongly linked to the way standard LLMs generate responses: if the model answers directly from its internal parameters, it may rely on statistical associations rather than verifiable evidence. A common mitigation is to require the model to first retrieve relevant source material and then generate an answer grounded in that context. This is the core idea of Retrieval-Augmented Generation [30], whose workflow is illustrated in the corresponding figure 4.

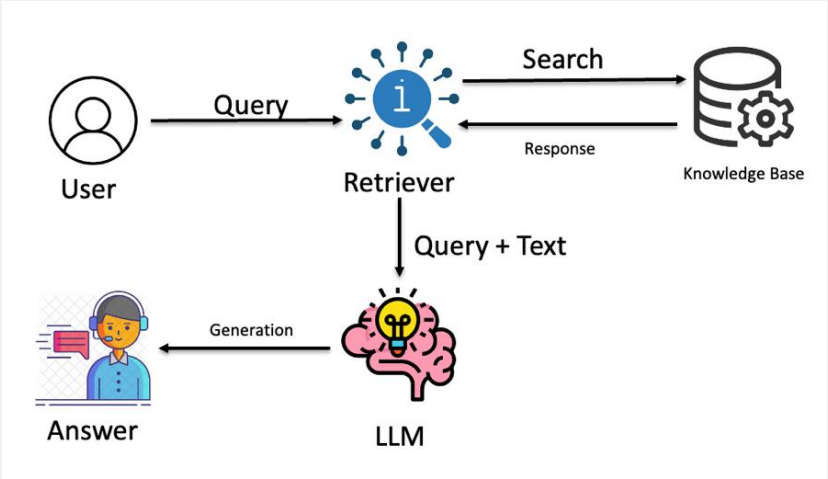

Figure 4: Work logic of RAG

Building on this foundation, real-world deployments have also been pursued across multiple domains. Two representative cases illustrate practical impact and implementation maturity: an admission-office assistant and a newsroom agent. The newsroom scenario aligns well with core LLM capabilities, since editorial workflows depend heavily on summarization, question answering, structured information extraction, and content drafting.

To improve usability under limited compute, parameter- and inference-efficient techniques were incorporated in a collaborative project titled “AI QazMedia,” conducted with the Faculty of Journalism in the AI Media Lab supported by LG Electronics Kazakhstan and focused on applications in media and communication science. In that study, the quantized model “issai/llama3.1-70b-GGUF4” was deployed as part of an AI assistant built on the open-source RAG platform RagFlow [31] as shown in Figure. Using the compressed GGUF4 format reduced inference-time GPU memory consumption by more than 70%, enabling deployment in resource-constrained settings. Whereas full-precision inference for a 70B-parameter model typically requires well over 100 GB of GPU memory, the quantized variant can operate effectively on a single 48 GB GPU, improving feasibility for both academic and production use. The underlying idea and system structure are summarized in the accompanying figure 5.

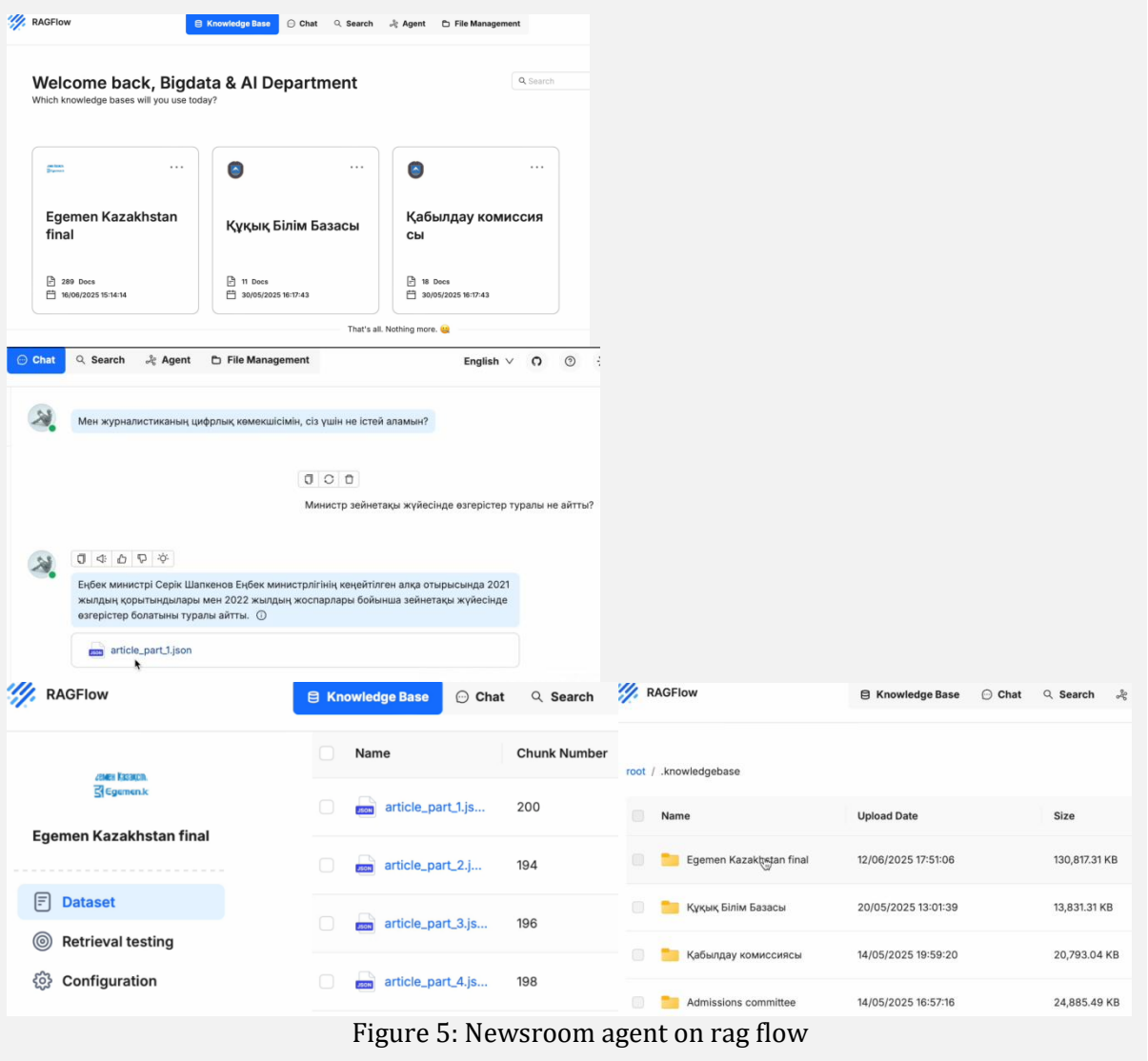

Figure 5: Newsroom agent on rag flow

This efficiency gain is primarily driven by quantization [32], which reduces model precision (commonly to 8-bit or 4-bit) while preserving much of the model's practical capability. In addition to quantization, distillation and optimized compute kernels are widely used to reduce serving cost and latency. Together, these methods have contributed to a major shift in the field: inference efficiency is now a first-order constraint, and successful deployments increasingly depend on carefully engineered serving stacks rather than model quality alone. Once architecture-level constraints are addressed, attention naturally shifts toward data efficiency, because scalable deployment and continuous improvement require high-quality data, robust retrieval indexes, and reliable update workflows.

The admission-office assistant addresses a different but equally high-impact need. Al-Farabi Kazakh National University receives tens of thousands of applications and an even larger volume of questions about programs, requirements, deadlines, and career pathways from prospective students across Kazakhstan and abroad. A recurring difficulty is that applicants often struggle to distinguish between closely related majors such as Computer Science, Software Engineering, Data Science, and Information Systems. In addition, educational-program information is frequently distributed across multiple webpages and documents, making it hard for students to connect fragmented details into a coherent understanding and a realistic study plan. The proposed assistant therefore uses a RAG-based architecture (illustrated in the figure 6) to consolidate trusted materials, retrieve relevant passages, and generate grounded answers that help applicants navigate programs and make informed decisions.

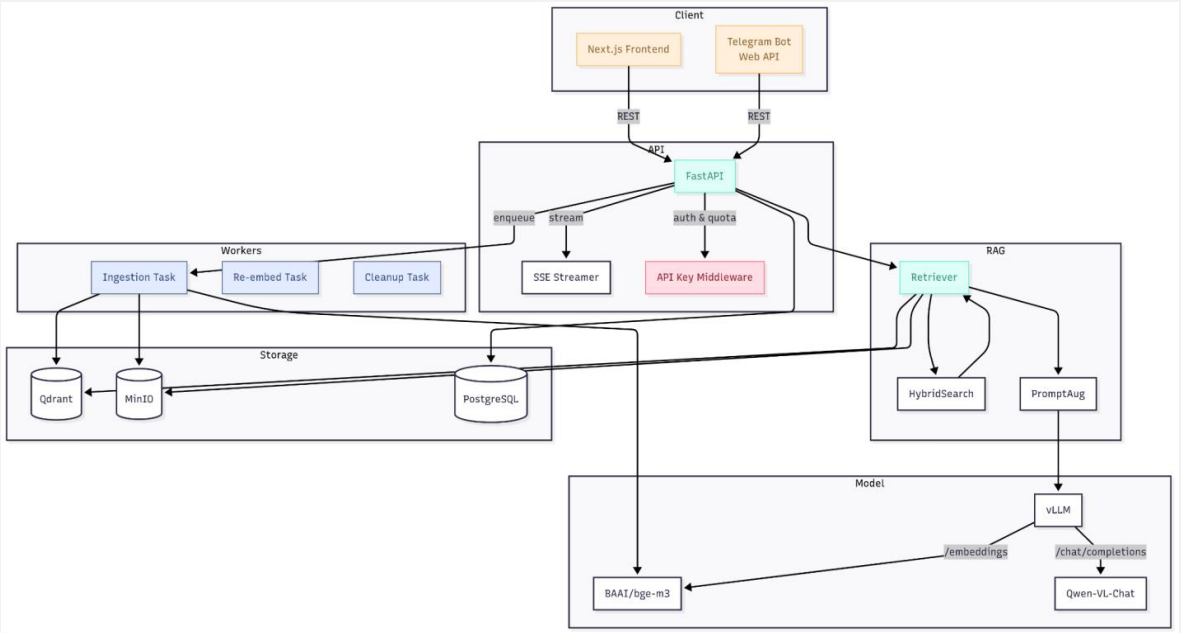

Figure 6: architecture of admission assistant

From a deployment perspective, the admission assistant is designed around four operational goals. First, full data control ensures that sensitive content and logs remain within the university infrastructure. Second, transparency and adaptability allow flexible choices of models, indexing strategies, and output formats as requirements evolve. Third, trilingual support (Kazakh, Russian, and English) enables broad accessibility without maintaining separate codebases. Fourth, independence from external APIs keeps computation local and reduces vendor lock-in. In the initial pilot, the system encountered two practical bottlenecks: response latency and residual hallucination in some queries. To address these issues, the next iteration plans to use stronger on-premise GPU infrastructure, including an NVIDIA DGX-class server, and to incorporate the ReAct framework to improve reasoning traceability and tool-use behavior. Together, these upgrades aim to improve both response speed and factual reliability, enabling wider adoption in university-wide admission operations..

Discussion

The evidence reviewed in this paper suggests that Kazakhstan's AI development over the past decade has moved from early-stage adoption toward ecosystem formation. Research output has increased sharply, and AI is no longer confined to academic prototypes. Large state and industrial actors are deploying AI in production settings, while parallel efforts are expanding compute capacity, data infrastructure, and workforce training. This combination indicates that Kazakhstan is building not only "AI projects," but also the institutional pathways needed for sustained adoption, including procurement, deployment practices, and skills pipelines.

At the same time, the main bottleneck is not model architecture alone but end-to-end capability. For Kazakhstan, the critical constraints are data availability and quality, compute access for both training and inference, and rigorous evaluation that can support deployment decisions. This is especially visible in language technologies. Because Kazakh is a low-resource language, it remains underrepresented in many global pretraining corpora, which leads to weaker baseline performance and makes additional adaptation unavoidable. The case study shows that improving performance requires investment in data assets that are suitable for LLM training, including curated instruction datasets and large-scale digitization with careful preprocessing.

The project case also demonstrates that "local LLM development" is primarily a systems problem. Even when training compact models, a viable pipeline must cover corpus acquisition, cleaning and deduplication, tokenization, training stability, and post-training alignment. In

deployment, reliability depends on controlling hallucination and ensuring traceability. Retrieval-Augmented Generation is a practical mechanism for grounding answers in verifiable sources, but it introduces new requirements: high-quality document repositories, indexing strategies, re-ranking, monitoring, and periodic refresh of knowledge bases. These components determine real-world usefulness as much as the base model itself.

Finally, operational constraints shape what can be deployed at scale. Inference cost, latency, and memory are limiting factors for universities, public services, and many enterprises. The newsroom case illustrates how quantization and other efficiency methods can enable advanced models in resource-constrained environments, while still delivering useful capabilities such as summarization, question answering, and content drafting. The admission assistant illustrates a second reality: user-facing systems must support multilingual interaction, transparent behavior, and local data control. Early pilots revealed common failure modes, including slow responses and residual hallucination, which points to the need for stronger on-premise GPU infrastructure and more robust agentic workflows, such as tool-use and reasoning frameworks, to improve reliability.

Conclusion

Kazakhstan's AI progress over the past decade is best understood as a transition toward a full-stack ecosystem that connects research, infrastructure, skills, and deployment. The country's near-term success will depend on strengthening the components that convert models into dependable services: trusted and continuously updated data assets, sufficient compute for training and serving, benchmark-driven evaluation, and responsible deployment practices including alignment, monitoring, and traceability. If these elements advance together, Kazakhstan can move from isolated demonstrations to scalable AI systems that deliver measurable value in public services, industry, education, and scientific research.

Reference

[1] Zamora-Cárdenas, W., Zumbado, M., & Trejos-Zelaya, I. (2020). McCulloch-Pitts Artificial Neuron and Rosenblatt's Perceptron: An abstract specification in Z. *Technology Inside by CPIC*, *5*, 16-29.

[2] Qu, P., Yan, J., Zhang, Y. H., & Gao, G. R. (2017). Parallel turing machine, a proposal. *Journal of Computer Science and Technology*, *32*(2), 269-285.

[3] LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural computation*, *1*(4), 541-551.

[4] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 25.

[5] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

[6] Chatterji, A., Cunningham, T., Deming, D. J., Hitzig, Z., Ong, C., Shan, C. Y., & Wadman, K. (2025). How people use chatgpt (No. w34255). National Bureau of Economic Research.

[7] Institute of Smart Systems and Artificial Intelligence. (n.d.). KazLLM. Nazarbayev University. Retrieved January 20, 2026, from https://issai.nu.edu.kz/kazllm/

[8] Astana Hub. (n.d.). AlemLLM (astanahub/alemllm) [Large language model]. Hugging Face. Retrieved January 20, 2026, from https://huggingface.co/astanahub/alemllm

[9] President Kassym-Jomart Tokayev's State of the Nation Address to the people of Kazakhstan: Kazakhstan in the era of artificial intelligence: Current challenges and solutions through digital transformation. Official website of the President of the Republic of Kazakhstan. https://www.akorda.kz/en/president-kassym-jomart-tokayevs-state-of-the-nation-address-to-the-people-of-kazakhstan-kazakhstan-in-the-era-of-artificial-intelligence-current-challenges-and-solutions-through-digital-transformation-1083029

[10] Interfax-Kazakhstan. (2026, January 6). Kazakhstan's Samruk-Kazyna projects $1.3 bln gain from AI over five years. https://www.interfax.kz/?int_id=21&lang=eng&news_id=77600

[11] Prime Minister of the Republic of Kazakhstan. (2026, January 6). Kazakhstan actively introduces AI solutions into production processes. https://primeminister.kz/en/news/kazakhstan-actively-introduces-ai-solutions-into-production-processes-30932

[12] Kazinform News Agency. (2025, November 18). Kazakhstan joins the world's TOP-500 most powerful supercomputers. https://qazinform.com/news/kazakhstan-joins-the-worlds-top-500-most-powerful-supercomputers-b42a40

[13] ENERGY Insights & Analytics. (2025, August 20). Kazakhstan Energy Outlook 2025 (Report). EXia. https://s3-prod.exia.kz/articles/Kazakhstan_Energy_Outlook_2025_EN.pdf

[14] Ministry of Artificial Intelligence and Digital Development of the Republic of Kazakhstan. (n.d.). Почти 1 млн казахстанцев прошли обучение по искусственному интеллекту: как выполняется поручение Главы государства по подготовке кадров для цифровой страны. GOV.KZ. Retrieved January 20, 2026, from https://www.gov.kz/memleket/entities/maidd/press/news/details/1142334?lang=ru

[15] Mussa, A., Tuimebayev, Z., & Mansurova, M. (2025). Make Large Language Models Efficient: A Review. *IEEE Access*.

[16] AI Index Steering Committee. (2025). AI Index Report 2025. Retrieved January 20, 2026, from https://hai.stanford.edu/ai-index/2025-ai-index-report

[17] Xia, M., Malladi, S., Gururangan, S., Arora, S., & Chen, D. (2024). Less: Selecting influential data for targeted instruction tuning. *arXiv preprint arXiv:2402.04333*.

[18] Xu M., Yin W., Cai D., Yi, R., Xu, D., Wang, Q., ... & Liu, X. (2024). A survey of resource-efficient llm and multimodal foundation models. *arXiv preprint arXiv:2401.08092*.

[19] Lin X., Wang W., Li Y., Yang S., Feng F., Wei Y., & Chua T. S. (2024, July). Data-efficient Fine-tuning for LLM-based Recommendation. In *Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval* (pp. 365-374).

[20] Zhou Z., Ning X., Hong K., Fu T., Xu J., Li S., ... & Wang Y. (2024). A survey on efficient inference for large language models. *arXiv preprint arXiv:2404.14294*.

[21] Griffith S., Subramanian K., Scholz J., Isbell C. L., & Thomaz, A. L. (2013). Policy shaping: Integrating human feedback with reinforcement learning. Advances in neural information processing systems, 26.

[22] Xu H., Sharaf A., Chen Y., Tan W., Shen L., Van Durme B., ... & Kim, Y. J. (2024). Contrastive preference optimization: Pushing the boundaries of llm performance in machine translation. *arXiv preprint arXiv:2401.08417*.

[23] Acharya D. B., Kuppan K., & Divya B. (2025). Agentic ai: Autonomous intelligence for complex goals–a comprehensive survey. IEEe Access.

[24] Kaz-Transformers. (n.d.). *kz-transformers* (Hugging Face organization page). Hugging Face. Retrieved January 20, 2026, from https://huggingface.co/kz-transformers

[25] Cao Y., Hong, S. Li, X. Ying J., Ma Y., Liang H., ... & Jiang Y. G. (2025). Toward generalizable evaluation in the llm era: A survey beyond benchmarks. arXiv preprint arXiv:2504.18838.

[26] Touvron H., Martin L., Stone K., Albert P., Almahairi A., Babaei Y., ... & Scialom, T. (2023). Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

[27] Ahmet Baitursynuly Institute of Linguistics. (n.d.). *National Corpus of the Kazakh Language* (QazCorpus). Retrieved January 20, 2026, from https://qazcorpus.kz

[28] Najam R., & Faizullah S. (2023). Analysis of recent deep learning techniques for Arabic handwritten-text OCR and post-OCR correction. Applied Sciences, 13(13), 7568.

[29] Liu Y., Yao Y., Ton J. F., Zhang X., Guo R., Cheng H., ... & Li H. (2023). Trustworthy llms: a survey and guideline for evaluating large language models' alignment. *arXiv preprint arXiv:2308.05374*.

[30] Fan W., Ding Y., Ning L., Wang S., Li H., Yin D., ... & Li Q. (2024, August). A survey on rag meeting llms: Towards retrieval-augmented large language models. In *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining* (pp. 6491-6501).

[31] RAGFlow. (n.d.). *RAGFlow* (official website). Retrieved January 20, 2026, from https://ragflow.io

[32] Li M., Huang Z., Chen L., Ren J., Jiang M., Li F., ... & Gao C. (2024, June). Contemporary advances in neural network quantization: A survey. In *2024 International Joint Conference on Neural Networks (IJCNN)* (pp. 1-10). IEEE.