Advanced deep learning methods for text generation in 2022-2024 literature: a systematic review

Artem V. Slobodianiuk[1], Serhiy O. Semerikov[1,2,3,4,5]

[1]Kryvyi Rih State Pedagogical University, 54 Universytetskyi Ave., Kryvyi Rih, 50086, Ukraine
[2]Institute for Digitalisation of Education of the NAES of Ukraine, 9 M. Berlynskoho Str., Kyiv, 04060, Ukraine
[3]Zhytomyr Polytechnic State University, 103 Chudnivsyka Str., Zhytomyr, 10005, Ukraine
[4]Center for Information-analytical and Technical Support of Nuclear Power Facilities Monitoring of the NAS of Ukraine, 34a Palladin Ave., Kyiv, 03142, Ukraine
[5]Academy of Cognitive and Natural Sciences, 54 Universytetskyi Ave., Kryvyi Rih, 50086, Ukraine

Abstract

This paper presents a systematic review of advanced deep learning methods used for text generation according to the literature published in 2022-2024. The review analyzes 43 research articles retrieved from the Scopus database focusing on the neural network architectures and approaches employed for text generation tasks. The results show that models based on the Transformer architecture, such as GPT-2, GPT-3, BERT, and their variants, are the most popular among the innovative approaches. Attention mechanisms and controllable text generation techniques are also gaining popularity. Overall, there is a trend of shifting from traditional methods to more innovative and effective models based on the Transformer architecture and attention mechanisms. Because the systematic search was executed in February 2024, the paper additionally provides a narrative, non-systematic discussion of developments between 2024 and 2026, which finds that Transformer dominance persisted while the locus of innovation moved to post-training, inference-time computation, retrieval and evaluation. The findings of this review can guide researchers and practitioners in selecting appropriate methods and architectures for developing text generation systems and identifying promising directions for future research in this field.

Keywords

text generation, natural language processing, deep learning, neural networks, systematic review, Transformer, GPT, BERT, attention mechanism, controllable text generation

1. Introduction

Text generation is a subfield of Natural Language Processing (NLP) that combines computational linguistics and artificial intelligence to generate natural language text [1]. The recent advancements in deep learning, particularly the emergence of large language models like GPT-3 [2], have fueled the interest in text generation and opened up new possibilities for automating the creation of various types of textual content.

Fatima et al. [1] conducted a systematic review of text generation using deep neural network models covering 90 sources from 2015 to 2021. However, the rapid development of large language models and their increasing accessibility in 2022-2023 [3] has led to a surge of interest in this field, necessitating an update to the previous review.

CS&SE@SW 2025: 8th Workshop for Young Scientists in Computer Science & Software Engineering, November 21, 2025, Kryvyi Rih, Ukraine
https://doi.org/10.55056/cssesw2025/paper42
minekosdid@kdpu.edu.ua (A. V. Slobodianiuk); semerikov@gmail.com (S. O. Semerikov)
https://acnsci.org/semerikov/ (S. O. Semerikov)
0009-0007-9425-1255 (A. V. Slobodianiuk); 0000-0003-0789-0272 (S. O. Semerikov)

2. Methodology

2.1. Search strategy and information sources

The Scopus database was selected as the sole information source for this review, considering its extensive coverage of relevant libraries and databases. The following search query was used to retrieve articles:
`TITLE-ABS-KEY(neural network) OR TITLE-ABS-KEY(machine learning) OR TITLE-ABS-KEY(deep learning)) AND TITLE("text generation")`

2.2. Eligibility criteria and study selection

The inclusion criteria for the articles were: (1) published between 2022 and 2024; (2) focus on text generation using artificial neural networks; (3) describe approaches, architectures, quality metrics, languages, datasets, or applications of text generation.

The exclusion criteria were: (1) published before 2022 or not containing data from 2022-2024; (2) not related to text generation or not using artificial neural networks; (3) not containing relevant information for the research questions.

The study selection process followed the PRISMA methodology [4]. Out of 248 documents retrieved from Scopus, 43 articles were selected for the final review after applying the eligibility criteria and removing duplicates.

The search was executed in February 2024, which fixes the evidence base of this review to the 2022–2024 window. Because the paper is published in 2026, section 4 adds a narrative, non-systematic discussion of developments after that cutoff; it does not extend the systematic search reported here.

3. Results

3.1. Overview of selected studies

The 43 articles selected for the review were published in 2022 (19 articles), 2023 (22 articles), and 2024 (2 articles as of February). There was a predominance of journal articles (22) over conference papers (21), which may indicate a more thorough coverage of the text generation issues in journals in recent years.

The analysis of 43 selected articles reveals a trend towards more in-depth coverage of text generation topics in scientific journals compared to conference proceedings. This trend is consistent with findings from previous reviews [1], but shows an acceleration in the shift towards journal publications. For instance, while Fatima et al. [1] reported a relatively even split between journal and conference papers in 2015-2021, our analysis shows a clear preference for journal publications in 2023-2024. This could indicate a maturation of the field, with researchers producing more comprehensive and thoroughly validated studies.

3.2. Advanced deep learning methods for text generation

Table 1 presents an overview of the neural network architectures used for text generation according to the 2022-2024 research data. The architectures are divided into traditional and innovative approaches.

Table 1 provides a comprehensive overview of both traditional and innovative neural network architectures used for text generation. The traditional approaches, such as RNN and LSTM, are still in use but are being increasingly supplanted by innovative approaches. Notably, Transformer-based models like BERT and GPT variants dominate the innovative approaches, reflecting the paradigm shift in natural language processing towards attention-based mechanisms. The diversity of architectures underscores the rapid evolution and experimentation in the field of text generation.

Among the innovative approaches, the most popular are models based on the Transformer architecture, such as GPT-2, GPT-3, BERT, and their variations. These models demonstrate high efficiency in generating coherent and semantically relevant text. Approaches using attention mechanisms and controllable text generation techniques are also gaining popularity.

Table 1
Neural network architectures for text generation.

Architecture	Description	Representatives	Articles
Traditional approaches			
RNN	Recurrent Neural Networks used for sequential data processing.	–	[5, 6, 7, 8, 9, 10, 11]
LSTM	A variant of RNN that better remembers long-term dependencies.	–	[5, 8, 12, 13, 14, 10, 11, 15, 9, 16, 17]
CNN	Convolutional Neural Networks often used for image processing.	YOLOv5	[5, 6, 18, 19]
Graph Neural Networks	Models that work with graph data structures.	GraphWriter, CGE-LW	[20, 6]
Innovative approaches			
Autoencoders	Networks used for learning effective encodings of unlabeled data.	AE, VAE, iVAE, clVAE+ MI, β0.4 VAE, SaVAE, LagVAE	[21, 14, 10]
Transformer	Architecture that uses an attention mechanism for sequential data processing.	T5, CodeT5, TrICY, DETR	[20, 6, 22, 23, 24, 7, 25, 26, 27, 28, 29, 16, 30, 31, 32, 33, 23, 34]
BERT	A Transformer-based model trained on large amounts of unlabeled text.	PubmedBERT, BioLinkBERT, RoBERTa, XLM-RoBERTa	[25, 35, 12, 22, 23, 26, 36, 37, 11, 38, 27, 39, 6, 30, 40, 41]
GPT-2, GPT-3	Transformer-based models used for text generation.	OPT, Llama, CodeBERT	[5, 25, 8, 9, 12, 7, 14, 22, 42, 28, 39, 43, 44, 36, 23, 15, 45, 27, 35, 32, 41, 46, 30, 31, 33]
Attention-based models	Models that use attention mechanisms to improve the quality of generated text.	–	[7, 25, 26, 36, 31, 33]
Seq2Seq	Architecture that uses an encoder and decoder for sequence generation.	S2ST, S2SL, S2SG, S2ST+, D+ Full, DSG	[30, 14, 17, 37, 24, 47, 31]
GAN	Generative Adversarial Networks consisting of a generator and discriminator.	EGAN, TILGAN, DoubAN-Full, WRGAN, CatGAN, SeqGAN, DGSAN	[5, 10, 45]

Table 2 presents a generalization of text generation approaches based on the data from table 1.

Table 2
Text generation approaches.

Category	Articles
Traditional approaches	[13, 18, 19]
Innovative approaches	[22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47]
Combination of traditional and innovative approaches	[5, 6, 7, 8, 9, 10, 11, 12, 14, 15, 16, 17, 20]

Traditional approaches, although used less frequently, still find application in certain tasks, such as image-based text generation, machine translation, and others.

Overall, there is a trend of shifting from traditional approaches to more innovative and effective models based on the Transformer architecture and attention mechanisms. This allows improving the quality of the generated text and expanding the scope of these technologies.

Figure 1 shows that innovative approaches to text generation predominate in 2022 and 2023, while traditional approaches and a combination of approaches are less common.

Comparing the obtained results with the data from the previous systematic review [1], the following conclusions can be drawn:

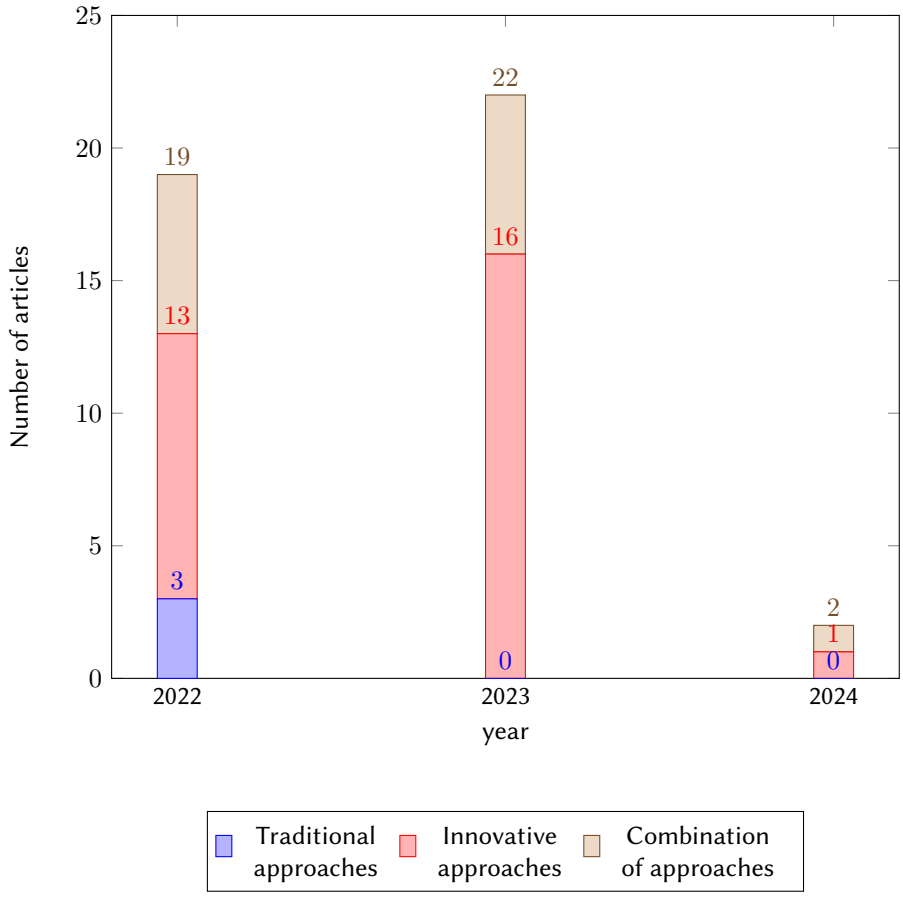

Figure 1: Distribution of articles by year and text generation approach categories.

- Traditional approaches, such as RNN, LSTM, and CNN, are still used for text generation but to a lesser extent compared to innovative approaches.
- The Transformer architecture and its variants (GPT-2, GPT-3, BERT) have gained significant popularity in 2022-2024, demonstrating high efficiency in generating coherent and semantically relevant text.
- New architectures and approaches have emerged, such as Diffusion Models and Memory Networks models, which were not presented in the previous review.
- Significant attention is paid to models using attention mechanisms and controllable text generation.
- There is a trend towards combining traditional and innovative approaches to achieve better results in text generation.
- Overall, in 2022-2024, there is a shift from traditional approaches to more innovative and effective models based on the Transformer architecture and attention mechanisms, which allows improving the quality of the generated text and expanding the scope of these technologies.

4. Discussion: neural text generation after the review window

4.1. Why this section is needed, and what it is not

The search reported in section 2 was executed in February 2024 and, by construction, describes work published between 2022 and the opening weeks of 2024. This paper appears in 2026. A review that stops at its own search date invites being read as a statement about the present rather than about the window it actually covers, and in this particular field the gap is not a formality: the two years that followed the cutoff reorganised how text generation systems are built, evaluated and deployed.

The status of what follows should therefore be stated plainly. The PRISMA process of section 2 is not re-run, no article is added to the 43 that constitute the evidence base, and tables 1 and 2 are left exactly as the systematic search produced them. This section is a *narrative*, non-systematic update. Its sources were gathered by putting twenty-six structured questions about post-cutoff developments, together with one synthesis request spanning the same ground, to Scopus AI over the Scopus corpus – the same database the review itself used – and keeping only returned references that carry a resolvable DOI. The procedure is reproducible and its provenance is documented, but it has no eligibility protocol, no duplicate screening and no claim to completeness. Section 4.12 states what follows from that.

4.2. The headline change: innovation moved off the backbone

The most useful single observation about 2024–2026 is that the backbone did not change much and almost everything around it did. Transformer-based models remained the prevailing paradigm for text generation throughout the period [48, 49, 50], so the central finding of this review survives intact. What moved was the locus of innovation: by 2026 the quality of a generation system was determined less by which architecture it used than by the stack assembled around it.

It is convenient to separate five layers, because the update is essentially a statement about where research attention travelled through them. The *backbone* is the pretrained model, still almost always a Transformer. *Post-training* adapts it through instruction tuning, parameter-efficient fine-tuning and preference optimisation. *Inference-time control* improves outputs without touching weights, through reasoning, search, retrieval and tools. The *evaluation stack* decides whether any of it worked. The *risk stack* decides whether it is safe to deploy. Between 2022 and 2024 the literature surveyed in section 3 was concentrated in the first layer, comparing architectures against one another. By 2026 the comparative question had largely been settled and the open problems had migrated upward.

This reframing matters for how the present review should be read. Its architecture-centred tables describe a period in which architecture was the variable of interest. That was an accurate description of 2022–2024 research, and it is no longer an accurate description of where the difficulty lies.

4.3. Architectures: continuity, with qualifications

Within the backbone layer the picture is one of narrowing rather than upheaval. Where 2022–2023 work treated encoder–decoder, recurrent and Transformer designs as competing options, 2025–2026 work is organised around decoder-only large language models and their retrieval-augmented variants, and increasingly asks which adaptation strategy to apply rather than which architecture to choose [48, 50]. The shift is from "Transformers versus older generators" to "which Transformer derivative, adapted how".

Two qualifications are needed, and both cut against a conclusion drawn in section 3. First, state space models emerged after the cutoff as a genuine alternative line of work rather than a curiosity: Mamba and its successors are explicitly positioned as alternatives to self-attention and reach parity on several sequence tasks, while remaining weaker on long-context generalisation and prone to local-pattern shortcuts, which is why the strongest systems in this line are extensions and hybrids rather than replacements [51, 52, 53, 54]. Second, diffusion models for text, which the 2022–2024 corpus registered only as an emerging possibility, matured into a practical contender with competitive quality and materially faster inference on selected tasks [55, 56, 57].

Generation also stopped being a text-only problem. Where the reviewed corpus treats image-conditioned captioning as one application among several, post-2024 work is organised around joint generative systems in which text is produced together with, and explicitly aligned to, images, audio or video, with evaluation designed per modality rather than borrowed from text [58, 59, 60, 61]. The operative question moved from whether a model can describe an image to whether its output stays coherent across modalities.

Sparse mixture-of-experts designs are the one architectural development with immediate economic consequences. The post-2024 evidence is that they deliver better quality per unit of training compute,

but that the saving is contingent on a serving stack built around routing, batching and memory locality; without it, latency and memory traffic absorb the gain [62, 63, 64, 65, 66].

4.4. Post-training became the main differentiator

The clearest training-side change is that pretraining stopped being the interesting part. Instruction tuning, supervised fine-tuning and parameter-efficient methods such as LoRA became the standard route to downstream quality, with the parameter-efficient variants preserving most of the gain at a fraction of the cost, though not without measurable effects on output diversity [67, 68].

Preference-based alignment moved further and faster. Direct preference optimisation spread widely because it dispenses with a separately trained reward model and shortens the pipeline, and the post-2024 literature continues to apply and extend it – to summarisation with small models, and alongside prefix- and process-level reinforcement schemes – rather than abandon it [69, 70, 71]. Reinforcement learning from verifiable rewards, in which correctness is checked automatically instead of judged by annotators, moved from a niche idea to a central post-training route for reasoning-heavy generation [71, 72, 73]. The most visible demonstration is a reasoning model trained largely through reinforcement learning with verifiable rewards [74]. It is worth recording that the field did not simply converge on the simpler method: the comparative evidence remains genuinely mixed, with policy-optimisation methods still outperforming direct preference optimisation on some testbeds.

Open weight releases are the enabling condition for much of this. Their research effect was to convert text generation from a largely API-mediated activity into a reproducible one, in which fine-tuning, benchmarking and domain adaptation can be carried out and replicated locally [75, 76, 77]. The same openness widened the attack surface, a point taken up in section 4.9.

Small models became a research programme in their own right rather than a compromise. The consistent post-2024 finding is a division of labour: distillation preserves generation quality when shrinking a model, quantisation makes the result deployable, and the strongest systems combine both, with the balance set by whether latency, memory, energy or fidelity dominates [78, 79, 80, 81].

4.5. Inference-time computation became a first-class variable

Nothing in the 2022–2024 corpus anticipates how much of the post-cutoff quality gain would come from spending computation at inference rather than at training. Chain-of-thought prompting, self-consistency, tree- and graph-structured search, and self-verification were treated by 2026 as core quality mechanisms rather than prompting tricks [82, 83, 84].

The evidence supports a conditional rather than a general claim. Additional inference compute helps when the task rewards deliberation, branching or verification, and the returns are best when compute is allocated adaptively per problem rather than spent uniformly [84]. On easy problems the same machinery produces redundancy and measurable over-reasoning, and self-revision can degrade an answer that was already correct [85]. Structured search combined with verification generally outperforms simply generating longer reasoning traces.

4.6. Retrieval, tools and agents

Retrieval-augmented generation developed from a single retrieve-then-generate step into modular pipelines with query rewriting, reranking, adaptive retrieval and uncertainty estimation, and then into graph-based and agentic designs that traverse structured knowledge [86, 87, 88]. The effect on factual accuracy is positive but conditional on retrieval quality, which is why uncertainty-aware and multi-evidence designs became necessary rather than optional [87, 88]; deployed systems in education and auditing show the same dependence on the quality of what is retrieved [89, 90].

Tool-using and code-executing agents became a substantial literature after the cutoff, with the field's centre of gravity moving from whether models can invoke tools to when they do so correctly, safely and efficiently [91, 92, 93]. Systematic taxonomies of agent failure modes now exist [94, 95], which is itself an indicator of maturity: the failures are reproducible enough to classify.

4.7. Control widened from decoding constraints to control stacks

Controllable generation, which this review identified as a rising theme, did rise – but its meaning widened enough that the 2022–2024 framing no longer contains it. Control in the earlier corpus meant decoding constraints and attribute conditioning. By 2026 it denoted a stack combining prompting, decoding-time constraints, fine-tuning, reinforcement learning, latent and activation steering, and post-processing, with measurable gains in fine-grained adherence to format and attribute constraints [96]. Dense multi-constraint tasks, safety robustness and faithfulness remain the standing failure cases.

A distinct strand constrains output to a formal grammar or schema, driven by the practical need to place generated text into software pipelines. Structured decoding reliably improves syntactic validity and schema compliance, but naive hard masking degrades reasoning quality even as it guarantees well-formed output; the better methods preserve the model's conditional distribution or align constraints with token boundaries [97, 98, 99, 100].

Long context deserves a note here because it was widely expected to solve control and grounding problems and largely did not. Context windows reached hundreds of thousands of tokens, but the capacity did not translate reliably into use: models still degrade with input length, lose factual precision, and miss relevant material that is demonstrably inside the window [101, 102, 103]. Long contexts also opened new safety failures peculiar to length [104, 105].

4.8. Evaluation was the quietest and most consequential change

The review's own corpus evaluated generated text largely with n-gram overlap against references. The post-2024 literature treats that as insufficient on its own – not obsolete, but incapable of answering the question that now matters, which is whether output is correct, faithful and useful under the task's real constraints [106, 107]. Evaluation moved toward combinations of semantic metrics, factuality checks, domain benchmarks and human judgement.

Using language models as judges became the standard way to make that affordable, and the post-2024 assessment of the practice is carefully bounded. Judge models are usable as scalable surrogates in narrow, well-structured settings and under human oversight, while remaining sensitive to prompt design, position and verbosity, and to domain mismatch [108, 109, 110, 111]. The defensible position is that they are a triage layer, not a replacement for human raters [112].

Two threats to benchmark validity received sustained attention for the first time. Contamination inflates scores when evaluation data has leaked into training data, and saturation means that even uncontaminated benchmarks stop discriminating between strong models; they are separate problems that compound [113, 114, 115, 116]. The response has been a move away from single static leaderboards toward harder, more diverse and less leak-prone designs, with newly collected reference texts used specifically to reduce contamination [106].

Hallucination matured from a defect to be eliminated into a residual risk to be managed. The literature now distinguishes intrinsic from extrinsic error and factuality from faithfulness, and treats mitigation as a pipeline property – grounding, then adaptive retrieval, uncertainty estimation, verification and human review – rather than a decoding fix [88, 117, 118]. The prevailing conclusion is that hallucination is not fully eliminable in general-purpose models, and that bounding it requires external verification.

4.9. Risks widened from output quality to system security

Risk research broadened more than any area except inference-time methods. Jailbreaks and prompt injection became first-class concerns, with attacks shown to generalise across models and modalities, and with retrieval and agent pipelines providing indirect injection surfaces that model-level alignment does not cover [119, 120, 121, 122]. Retrieval augmentation, adopted largely as a safety measure against fabrication, does not itself make a system safer [123]. The defensive consensus is layered: filtering, provenance tracking, sandboxing and continuous red-teaming, because model-only defences are insufficient [124, 125, 126].

Provenance became a research area in its own right after the cutoff, along two lines: post-hoc detection of machine-generated text, and watermarking at generation time [127, 128, 129, 130]. Both are brittle in the ways that matter – detectors under distribution shift and adversarial rewriting, watermarks under paraphrase and translation [131].

Training on machine-generated text raised a question with no analogue in the 2022–2024 corpus, since the web it presupposes was still mostly human-written. Unfiltered recursive training on synthetic output degrades models, but the post-2024 finding is narrower than the popular reading: synthetic data is useful when its distribution is controlled, verified or mixed with sufficient real data, and the field is moving toward data-quality engineering rather than away from synthetic corpora [132, 133, 134].

Two further constraints appear that the earlier corpus does not register as research problems at all. The environmental cost of generation is now measured over the full lifecycle, and is dominated by inference volume, hardware and datacentre operation rather than by training a model once [135, 136, 137, 138]. And regulation became a design input: the EU AI Act, journal policies and publication-ethics guidance converge on no machine authorship, explicit disclosure and human accountability, while uptake has outrun compliance and left a measurable transparency gap [139, 140, 141, 142].

4.10. Languages, domains and the writing process

Low-resource generation, which the 2022–2024 corpus treated as a data problem, became a method problem: parameter-efficient fine-tuning, synthetic data pipelines and language-specific benchmarks now carry the work [143, 144, 145, 146]. Ukrainian is a legible instance of the pattern. The period produced an instruction-tuned Ukrainian text-editing model, a shared task devoted to fine-tuning open-weight models for the language, and targeted adaptation of open models to Ukrainian representation [147, 148, 149]; subsequent benchmark work found that base models can outperform instruction-tuned variants in few-shot settings, and that tokenizer inefficiency remains a measurable cost for Ukrainian generation [150, 151]. Both findings are cautionary for the assumption, natural in a corpus dominated by English, that gains transfer across languages unchanged.

Application activity concentrated unevenly. Medicine and biomedicine dominate post-2024 publication volume by a clear margin, with education and code generation forming substantial secondary clusters [50, 107, 152, 153]. Finally, a literature on human–model co-writing emerged that has direct bearing on how this paper and others like it are produced: collaboration raises speed and often quality, but as the model takes over more of the drafting the output becomes less diverse, less original and less owned by the writer [154, 155, 156, 157, 158].

4.11. What this means for the conclusions of this review

Table 3 sets the conclusions drawn in section 3 against the post-cutoff evidence. The summary judgement is that they hold in direction and are incomplete in scope: each remains a defensible statement about 2022–2024, and each requires a qualification that could not have been derived from that window.

Two findings have no counterpart in the original conclusions and would have to be added were the review repeated today. Inference-time computation is now a primary quality variable, on a par with model choice. And evaluation has become a research problem rather than a reporting step, because the metrics used throughout the reviewed corpus cannot answer the questions the field now asks.

4.12. Limitations of this update

This section is narrative and must be read as such. It applies no eligibility criteria, screens no duplicates and makes no completeness claim, so it cannot support quantitative statements of the kind reported in section 3 – there is no denominator here, and no counts are given. Its sources were surfaced through a retrieval system whose ranking is opaque, which favours recent and highly visible work and will under-represent negative results and small-venue publications; the recency skew is visible in the fact that most cited items are from 2025–2026. Only DOI-bearing references were retained, which excludes preprints that have been influential in this field, several of which introduced methods discussed above.

Table 3
Conclusions of the 2022–2024 review and their standing in the 2024–2026 literature.

Conclusion (2022–2024)	Standing	Qualification required by 2024–2026 work
Transformer models and their variants dominate	Holds	Dominance became system-level rather than architectural: the differentiator is the post-training and inference stack, not the backbone
Attention mechanisms are increasingly central	Holds	Foundational rather than emergent; the frontier moved to what surrounds an attention-based model
Controllable generation is gaining popularity	Holds, widened	Control now denotes a stack spanning prompting, decoding constraints, tuning, steering and post-processing, not attribute conditioning
RNN, LSTM, CNN, Seq2Seq and GAN approaches are declining	Holds with exceptions	State space models emerged after the cutoff as a distinct non-attention line reaching parity on some sequence tasks; the trend is a falling share rather than disappearance
Diffusion and memory models are newly emerging	Understated	Diffusion language models became practical contenders; memory and retrieval became core infrastructure
Traditional and innovative approaches are increasingly combined	Holds, redefined	Combination now means hybrid *systems* – retrieval, tools, verifiers around a Transformer – rather than hybrid architectures

Finally, the interval covered is short enough that much of the cited work has not yet accumulated citation history or independent replication, and some of it will not survive. The systematic evidence of this review remains the 43 articles of 2022–2024; this section is context for reading them, not an extension of them.

5. Conclusion

This systematic review analyzed the application of artificial neural networks for text generation in 2022-2024, focusing on the advanced deep learning methods employed. The results show that models based on the Transformer architecture, such as GPT-2, GPT-3, BERT, and their variants, are the most popular among the innovative approaches. Attention mechanisms and controllable text generation techniques are also gaining popularity. There is an overall trend of shifting from traditional methods to more innovative and effective models.

The findings of this review can guide researchers and practitioners in selecting appropriate methods and architectures for developing text generation systems and identifying promising directions for future research in this rapidly evolving field. As section 4 sets out, these conclusions hold in direction for the period they cover, but a reader in 2026 should apply them with the qualifications summarised in table 3: the decisive variables have moved from the choice of architecture to the post-training and inference-time stack built around it, and to the evaluation practices used to judge the result. Further studies may focus on developing new neural network architectures and text generation methods, creating more effective metrics for evaluating the quality of generated text, expanding the range of languages and domains of application, and addressing the ethical and social issues associated with the use of text generation systems.

Author contributions

Conceptualization, Artem V. Slobodianiuk and Serhiy O. Semerikov; methodology, Serhiy O. Semerikov; software, Artem V. Slobodianiuk; validation, Artem V. Slobodianiuk and Serhiy O. Semerikov; formal analysis, Artem V. Slobodianiuk; investigation, Artem V. Slobodianiuk; resources, Serhiy O. Semerikov; data curation, Artem V. Slobodianiuk; writing – original draft preparation, Artem V. Slobodianiuk; writing – review and editing, Artem V. Slobodianiuk and Serhiy O. Semerikov; visualization, Artem

V. Slobodianiuk; supervision, Serhiy O. Semerikov. All authors have read and agreed to the published version of the manuscript.

Funding

This research received no external funding.

Conflicts of interest

The authors declare no conflict of interest.

Declaration on Generative AI

The authors have not employed any generative AI tools.

References

[1] N. Fatima, A. S. Imran, Z. Kastrati, S. M. Daudpota, A. Soomro, A Systematic Literature Review on Text Generation Using Deep Neural Network Models, IEEE Access 10 (2022) 53490–53503. doi:10.1109/ACCESS.2022.3174108.

[2] R. Liashenko, S. Semerikov, Bibliometric analysis of chatbot training research: Key concepts and trends, Information Technologies and Learning Tools 101 (2024) 181–199. doi:10.33407/itlt.v101i3.5622.

[3] large language models - Google Trends, 2023. URL: https://trends.google.com/trends/explore?date=2022-01-01%202023-12-21&q=large%20language%20models&hl=en.

[4] M. J. Page, J. E. McKenzie, P. M. Bossuyt, I. Boutron, T. C. Hoffmann, C. D. Mulrow, L. Shamseer, J. M. Tetzlaff, E. A. Akl, S. E. Brennan, R. Chou, J. Glanville, J. M. Grimshaw, A. Hróbjartsson, M. M. Lalu, T. Li, E. W. Loder, E. Mayo-Wilson, S. McDonald, L. A. McGuinness, L. A. Stewart, J. Thomas, A. C. Tricco, V. A. Welch, P. Whiting, D. Moher, The PRISMA 2020 statement: an updated guideline for reporting systematic reviews, BMJ 372 (2021) n71. doi:10.1136/bmj.n71.

[5] A. Bas, M. O. Topal, c. Duman, I. Van Heerden, A Brief History of Deep Learning-Based Text Generation, in: J. M. Alja'Am, S. AlMaadeed, S. A. Elseoud, O. Karam (Eds.), Proceedings of the International Conference on Computer and Applications, ICCA 2022 - Proceedings, Institute of Electrical and Electronics Engineers Inc., 2022. doi:10.1109/ICCA56443.2022.10039545.

[6] W. Yu, C. Zhu, Z. Li, Z. Hu, Q. Wang, H. Ji, M. Jiang, A Survey of Knowledge-enhanced Text Generation, ACM Computing Surveys 54 (2022). doi:10.1145/3512467.

[7] L. Mou, Search and learning for unsupervised text generation, AI Magazine 43 (2022) 344–352. doi:10.1002/aaai.12068.

[8] J. Wu, Y. Guo, C. Gao, J. Sun, An automatic text generation algorithm of technical disclosure for catenary construction based on knowledge element model, Advanced Engineering Informatics 56 (2023) 101913. doi:10.1016/j.aei.2023.101913.

[9] H. Du, W. Xing, B. Pei, Automatic text generation using deep learning: providing large-scale support for online learning communities, Interactive Learning Environments 31 (2023) 5021–5036. doi:10.1080/10494820.2021.1993932.

[10] S. Shahriar, GAN computers generate arts? A survey on visual arts, music, and literary text generation using generative adversarial network, Displays 73 (2022) 102237. doi:10.1016/j.displa.2022.102237.

[11] H. Strobelt, J. Kinley, R. Krueger, J. Beyer, H. Pfister, A. M. Rush, GenNI: Human-AI Collaboration for Data-Backed Text Generation, IEEE Transactions on Visualization and Computer Graphics 28 (2022) 1106–1116. doi:10.1109/TVCG.2021.3114845.

[12] I. Alonso, E. Agirre, Automatic Logical Forms improve fidelity in Table-to-Text generation, Expert Systems with Applications 238 (2024) 121869. doi:10.1016/j.eswa.2023.121869.

[13] E. Kreiss, F. Fang, N. D. Goodman, C. Potts, Concadia: Towards Image-Based Text Generation with a Purpose, in: Y. Goldberg, Z. Kozareva, Y. Zhang (Eds.), Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Association for Computational Linguistics (ACL), 2022, pp. 4667–4684.

[14] K. Y. Rao, K. S. Rao, S. V. S. Narayana, Conditional-Aware Sequential Text Generation In Knowledge-Enhanced Conversational Recommendation System, Journal of Theoretical and Applied Information Technology 101 (2023) 2820–2836.

[15] N. Fatima, S. M. Daudpota, Z. Kastrati, A. S. Imran, S. Hassan, N. S. Elmitwally, Improving news headline text generation quality through frequent POS-Tag patterns analysis, Engineering Applications of Artificial Intelligence 125 (2023) 106718. doi:10.1016/j.engappai.2023.106718.

[16] Z. Lin, Y. Gong, Y. Shen, T. Wu, Z. Fan, C. Lin, N. Duan, W. Chen, Text Generation with Diffusion Language Models: A Pre-training Approach with Continuous Paragraph Denoise, in: A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, J. Scarlett (Eds.), Proceedings of Machine Learning Research, volume 202, ML Research Press, 2023, pp. 21051–21064.

[17] M. S. Amin, A. Mazzei, L. Anselma, Towards Data Augmentation for DRS-to-Text Generation, CEUR Workshop Proceedings 3287 (2022) 141–152.

[18] A. Hanafi, M. Bouhorma, L. Elaachak, Machine Learning-Based Augmented Reality For Improved Text Generation Through Recurrent Neural Networks, Journal of Theoretical and Applied Information Technology 100 (2022) 518–530.

[19] T. Tazalli, Z. A. Aunshu, S. S. Liya, M. Hossain, Z. Mehjabeen, M. S. Ahmed, M. I. Hossain, Computer Vision-Based Bengali Sign Language To Text Generation, in: 5th IEEE International Image Processing, Applications and Systems Conference, IPAS 2022, Institute of Electrical and Electronics Engineers Inc., 2022. doi:10.1109/IPAS55744.2022.10052928.

[20] J. Zhu, X. Ma, Z. Lin, P. De Meo, A quantum-like approach for text generation from knowledge graphs, CAAI Transactions on Intelligence Technology 8 (2023) 1455–1463. doi:10.1049/cit2.12178.

[21] Z. Teng, C. Chen, Y. Zhang, Y. Zhang, Contrastive Latent Variable Models for Neural Text Generation, in: J. Cussens, K. Zhang (Eds.), Proceedings of Machine Learning Research, volume 180, ML Research Press, 2022, pp. 1928–1938.

[22] C. An, J. Feng, K. Lv, L. Kong, X. Qiu, X. Huang, CONT: Contrastive Neural Text Generation, in: S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, A. Oh (Eds.), Advances in Neural Information Processing Systems, volume 35, Neural information processing systems foundation, 2022.

[23] H. Seo, S. Jung, J. Jung, T. Hwang, H. Namgoong, Y.-H. Roh, Controllable Text Generation Using Semantic Control Grammar, IEEE Access 11 (2023) 26329–26343. doi:10.1109/ACCESS.2023.3252017.

[24] X. Yin, X. Wan, How Do Seq2Seq Models Perform on End-to-End Data-to-Text Generation?, in: S. Muresan, P. Nakov, A. Villavicencio (Eds.), Proceedings of the Annual Meeting of the Association for Computational Linguistics, volume 1, Association for Computational Linguistics (ACL), 2022, pp. 7701–7710.

[25] H. Zhang, H. Song, S. Li, M. Zhou, D. Song, A Survey of Controllable Text Generation Using Transformer-based Pre-trained Language Models, ACM Computing Surveys 56 (2023). doi:10.1145/3617680.

[26] W. Zhou, Y. E. Jiang, E. Wilcox, R. Cotterell, M. Sachan, Controlled Text Generation with Natural Language Instructions, in: A. Krause, E. Brunskill, C. K., B. Engelhardt, S. Sabato, J. Scarlett (Eds.), Proceedings of Machine Learning Research, volume 202, ML Research Press, 2023, pp. 42602–42613.

[27] S. Montella, A. Nasr, J. Heinecke, F. Bechet, L. M. Rojas-Barahona, Investigating the Effect of Relative Positional Embeddings on AMR-to-Text Generation with Structural Adapters, in: EACL 2023 - 17th Conference of the European Chapter of the Association for Computational Linguistics,

Proceedings of the Conference, Association for Computational Linguistics (ACL), 2023, p. 727 – 736.

[28] S. Hong, S. Moon, J. Kim, S. Lee, M. Kim, D. Lee, J.-Y. Kim, DFX: A Low-latency Multi-FPGA Appliance for Accelerating Transformer-based Text Generation, in: Proceedings of the Annual International Symposium on Microarchitecture, MICRO, volume 2022-October, IEEE Computer Society, 2022, pp. 616–630. doi:10.1109/MICRO56248.2022.00051.

[29] Y. Li, L. Cui, J. Yan, Y. Yin, W. Bi, S. Shi, Y. Zhang, Explicit Syntactic Guidance for Neural Text Generation, in: Proceedings of the Annual Meeting of the Association for Computational Linguistics, volume 1, Association for Computational Linguistics (ACL), 2023, pp. 14095–14112.

[30] H. Le, D.-T. Le, V. Weber, C. Church, K. Rottmann, M. Bradford, P. Chin, Semi-supervised Adversarial Text Generation based on Seq2Seq models, in: EMNLP 2022 - Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track, Association for Computational Linguistics (ACL), 2022, pp. 264–272.

[31] M. Chen, X. Lu, T. Xu, Y. Li, J. Zhou, D. Dou, H. Xiong, Towards Table-to-Text Generation with Pretrained Language Model: A Table Structure Understanding and Text Deliberating Approach, in: Y. Goldberg, Z. Kozareva, Y. Zhang (Eds.), Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Association for Computational Linguistics (ACL), 2022, pp. 8199–8210.

[32] E. Seifossadat, H. Sameti, Improving semantic coverage of data-to-text generation model using dynamic memory networks, Natural Language Engineering 30 (2023) 454–479. doi:10.1017/S1351324923000207.

[33] V. Agarwal, S. Ghosh, H. BSS, H. Arora, B. R. K. Raja, TrICy: Trigger-Guided Data-to-Text Generation With Intent Aware Attention-Copy, IEEE/ACM Transactions on Audio, Speech, and Language Processing 32 (2024) 1173–1184. doi:10.1109/TASLP.2024.3353574.

[34] D. Taunk, S. Sagare, A. Patil, S. Subramanian, M. Gupta, V. Varma, XWikiGen: Cross-lingual Summarization for Encyclopedic Text Generation in Low Resource Languages, in: ACM Web Conference 2023 - Proceedings of the World Wide Web Conference, WWW 2023, Association for Computing Machinery, Inc, 2023, pp. 1703–1713. doi:10.1145/3543507.3583405.

[35] W. Xu, Y. Tuan, Y. Lu, M. Saxon, L. Li, W. Y. Wang, Not All Errors Are Equal: Learning Text Generation Metrics using Stratified Error Synthesis, in: Findings of the Association for Computational Linguistics: EMNLP 2022, Association for Computational Linguistics (ACL), 2022, p. 6588 – 6603.

[36] T. Wu, H. Wang, Z. Zeng, W. Wang, H.-T. Zheng, J. Zhang, Enhancing Text Generation with Cooperative Training, Frontiers in Artificial Intelligence and Applications 372 (2023) 2704–2711. doi:10.3233/FAIA230579.

[37] X. Chu, Feature Extraction and Intelligent Text Generation of Digital Music, Computational Intelligence and Neuroscience 2022 (2022). doi:10.1155/2022/7952259.

[38] C. Meyer, D. Adkins, K. Pal, R. Galici, A. Garcia-Agundez, C. Eickhoff, Neural text generation in regulatory medical writing, Frontiers in Pharmacology 14 (2023). doi:10.3389/fphar.2023.1086913.

[39] Q. Chen, H. Sun, H. Liu, Y. Jiang, T. Ran, X. Jin, X. Xiao, Z. Lin, H. Chen, Z. Niu, An extensive benchmark study on biomedical text generation and mining with ChatGPT, Bioinformatics 39 (2023). doi:10.1093/bioinformatics/btad557.

[40] X. Yue, H. A. Inan, X. Li, G. Kumar, J. McAnallen, H. Shajari, H. Sun, D. Levitan, R. Sim, Synthetic Text Generation with Differential Privacy: A Simple and Practical Recipe, in: Proceedings of the Annual Meeting of the Association for Computational Linguistics, volume 1, Association for Computational Linguistics (ACL), 2023, pp. 1321–1342.

[41] W. M. Si, M. Backes, Y. Zhang, A. Salem, Two-in-One: A Model Hijacking Attack Against Text Generation Models, in: 32nd USENIX Security Symposium, USENIX Security 2023, volume 3, USENIX Association, 2023, pp. 2223–2240.

[42] M. Bayer, M.-A. Kaufhold, B. Buchhold, M. Keller, J. Dallmeyer, C. Reuter, Data augmentation in natural language processing: a novel text generation approach for long and short text classifiers,

International Journal of Machine Learning and Cybernetics 14 (2023) 135–150. doi:10.1007/s13042-022-01553-3.

[43] M. Ghazvininejad, V. Karpukhin, V. Gor, A. Celikyilmaz, Discourse-Aware Soft Prompting for Text Generation, in: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Association for Computational Linguistics (ACL), 2022, pp. 4570–4589.

[44] J. J. Koplin, Dual-use implications of ai text generation, Ethics and Information Technology 25 (2023). doi:10.1007/s10676-023-09703-z.

[45] A. Pautrat-Lertora, R. Perez-Lozano, W. Ugarte, EGAN: Generatives Adversarial Networks for Text Generation with Sentiments, in: International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, IC3K - Proceedings, volume 1, Science and Technology Publications, Lda, 2022, pp. 249–256.

[46] X. Lu, S. Welleck, P. West, L. Jiang, J. Kasai, D. Khashabi, R. Le Bras, L. Qin, Y. Yu, R. Zellers, N. A. Smith, Y. Choi, NEUROLOGIC AFesque Decoding: Constrained Text Generation with Lookahead Heuristics, in: NAACL 2022 - 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, Association for Computational Linguistics (ACL), 2022, pp. 780–799.

[47] H. Gong, X. Feng, B. Qin, Quality Control for Distantly-Supervised Data-to-Text Generation via Meta Learning, Applied Sciences 13 (2023) 5573. doi:10.3390/app13095573.

[48] A. D. S, G. Charanmayee, I. Varsha, T. N. Sri, K. N. Sri, D. P P, An Empirical Evaluation of Transformer-Based LLMs for Domain-Aware Text Generation, in: 2026 IEEE Madhya Pradesh Section Conference (MPCON), IEEE, 2026, pp. 292–297. doi:10.1109/MPCON69668.2026.11508572.

[49] H. Xiao, F. Zhou, X. Liu, T. Liu, Z. Li, X. Liu, X. Huang, A comprehensive survey of large language models and multimodal large language models in medicine, Information Fusion 117 (2025) 102888. doi:10.1016/j.inffus.2024.102888.

[50] Z. Cao, V. K. Keloth, Q. Xie, L. Qian, Y. Liu, Y. Wang, R. Shi, W. Zhou, G. Yang, J. Zhang, X. Peng, E. Zhen, et al., The Development Landscape of Large Language Models for Biomedical Applications, Annual Review of Biomedical Data Science 8 (2025) 251–274. doi:10.1146/annurev-biodatasci-102224-074736.

[51] X. Zhang, Q. Zhang, H. Liu, T. Xiao, X. Qian, B. Ahmed, E. Ambikairajah, H. Li, J. Epps, Mamba in Speech: Towards an Alternative to Self-Attention, IEEE Transactions on Audio, Speech and Language Processing 33 (2025) 1933–1948. doi:10.1109/TASLPRO.2025.3566210.

[52] K. Miyazaki, Y. Masuyama, M. Murata, Exploring the Capability of Mamba in Speech Applications, in: Interspeech 2024, interspeech_2024, ISCA, 2024, pp. 237–241. doi:10.21437/Interspeech.2024-994.

[53] D. Yuan, J. Liu, B. Li, H. Zhang, J. Wang, X. Cai, D. Zhao, ReMamba: Equip Mamba with Effective Long-Sequence Modeling, in: Findings of the Association for Computational Linguistics: EMNLP 2025, Association for Computational Linguistics, 2025, pp. 6830–6840. doi:10.18653/v1/2025.findings-emnlp.361.

[54] W. You, Z. Tang, J. Li, L. Yao, M. Zhang, Revealing and Mitigating the Local Pattern Shortcuts of Mamba, in: Findings of the Association for Computational Linguistics: ACL 2025, Association for Computational Linguistics, 2025, pp. 12156–12178. doi:10.18653/v1/2025.findings-acl.629.

[55] J. Liu, P. Cheng, J. Dai, J. Liu, DiffuPercom: A novel simplex-based diffusion framework for personalized comment generation, Applied Soft Computing 201 (2026) 115493. doi:10.1016/j.asoc.2026.115493.

[56] N. Gao, Y. Lu, P. Chen, G. Sun, R. Liang, Y. Zhang, TWDT: Training-free word-level controllable diffusion model for text generation, Knowledge-Based Systems 330 (2025) 114437. doi:10.1016/j.knosys.2025.114437.

[57] J. Zhao, J. Zheng, C. Yu, H. Fu, R. Zhang, Improvement of Diffusion Models Based on Radial Basis Function Neural Networks and Consistency Regularization, Machine Learning 115 (2026). doi:10.1007/s10994-026-07028-8.

[58] A. S, H. S, A. B, S. V, K. S, Y. K. A, From text to multimodality: A comprehensive survey of LLMs and MLLMs across diverse domains, ICT Express (2026). doi:10.1016/j.icte.2026.07.023.
[59] S. Li, H. Tang, Multimodal Alignment and Fusion: A Survey, International Journal of Computer Vision 134 (2026). doi:10.1007/s11263-025-02667-1.
[60] K. Taha, Generative AI for multimodal content: a survey with empirical and experimental evaluations, Artificial Intelligence Review 59 (2026). doi:10.1007/s10462-026-11525-6.
[61] Y. Hu, Z. Yang, X. Yang, J. Li, TCHFNet: Topic-guided cross-modal alignment and hidden layer-optimized fusion network for multimodal sentiment classification, Neurocomputing 692 (2026) 133844. doi:10.1016/j.neucom.2026.133844.
[62] T. M. Ghazal, Sparse Mixture-of-Experts Transformers for Efficient Scaling of Large Language Models, Procedia Computer Science 275 (2026) 923–930. doi:10.1016/j.procs.2026.01.105.
[63] J. Liu, P. Tang, W. Wang, Y. Ren, X. Hou, P. A. Heng, M. Guo, C. Li, A Survey on Inference Optimization Techniques for Mixture of Experts Models, ACM Computing Surveys 58 (2026) 1–37. doi:10.1145/3794845.
[64] S. Abuzakuk, O. Balmau, J. Chen, A.-M. Kermarrec, R. Pires, R. Prabhu, M. de Vos, The Cost of Expertise: Understanding MoE Decode Performance, in: Proceedings of the Sixth European Workshop on Machine Learning and Systems, EuroMLSys '26, ACM, 2026, pp. 466–472. doi:10.1145/3805621.3807659.
[65] A. Hegde, A. Kumble, R. Gupta, Parallelism Strategies and Concurrency Effects for Mixture-of-Experts Inference on GPU Systems, in: 2026 Systems and Information Engineering Design Symposium (SIEDS), IEEE, 2026, pp. 1–6. doi:10.1109/SIEDS69358.2026.11540265.
[66] Q. Zhang, X. Li, J. Kong, X. Liu, B. Ji, S. Li, J. Ma, J. Yu, SPAR: Step-wise Path Dispatching and Asymmetric Re-routing for Efficient MoE Inference, Springer Nature Singapore, 2026, pp. 386–399. doi:10.1007/978-981-92-3450-9_32.
[67] L. Jiang, S. Shi, C. Yang, Instruction Fine-Tuning Guidance: How PEFT Methods Impact the Generation of Language Model via Different Attributions, Springer Nature Singapore, 2025, pp. 277–292. doi:10.1007/978-981-96-5123-8_19.
[68] M. Szep, D. Rueckert, R. von Eisenhart-Rothe, F. Hinterwimmer, Fine-tuning Large Language Models with Limited Data: A Survey and Practical Guide, Transactions of the Association for Computational Linguistics 14 (2026) 341–377. doi:10.1162/TACL.a.627.
[69] Y. Tamer, A. Badawi, M. Bahgat, Efficient Summarization with Small Language Models via Direct Preference Optimization, in: 2026 ICEENG International Conference for Innovations in Intelligent Computing and Cybersecurity (IICC), IEEE, 2026, pp. 1–7. doi:10.1109/IICC69623.2026.11581993.
[70] Y. Sun, Z. Zhao, Y. Wei, Y. Zhang, C. Gong, Well Begun, Half Done: Reinforcement Learning with Prefix Optimization for LLM Reasoning, Proceedings of the AAAI Conference on Artificial Intelligence 40 (2026) 33144–33152. doi:10.1609/aaai.v40i39.40598.
[71] S. Chen, Y. Liu, Hierarchical process-level generative reward modeling with adaptive long-horizon reasoning for robust LLM alignment, Information Processing & Management 63 (2026) 104957. doi:10.1016/j.ipm.2026.104957.
[72] W. Jia, J. Lu, H. Yu, S. Wang, G. Tang, A.-L. Wang, W. Yin, D. Yang, Y. Nie, B. Shan, H. Feng, I. Li, et al., MEML-GRPO: Heterogeneous Multi-Expert Mutual Learning for RLVR Advancement, Proceedings of the AAAI Conference on Artificial Intelligence 40 (2026) 31283–31291. doi:10.1609/aaai.v40i37.40391.
[73] J. Xu, S. Fang, Y. Wang, Z. Wang, Y. Qu, H. Xie, C-GRPO: Calibrating Multimodal Reasoning via Entropy-Aware Data Curation and Policy Optimization, IEEE Transactions on Big Data (2026) 1–12. doi:10.1109/TBDATA.2026.3688996.
[74] D. Guo, D. Yang, H. Zhang, J. Song, P. Wang, Q. Zhu, R. Xu, R. Zhang, S. Ma, X. Bi, X. Zhang, X. Yu, et al., DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning, Nature 645 (2025) 633–638. doi:10.1038/s41586-025-09422-z.
[75] J. Mahapatra, U. Garain, A Comprehensive Performance Evaluation of LLMs for Data-to-Text Generation and Divergence-Weighted Training, ACM Transactions on Intelligent Systems and

Technology 17 (2026) 1–25. doi:10.1145/3795137.

[76] M. Nadaş, L. Dioşan, A. Tomescu, A. Pişcoran, Building large-scale English–Romanian literary translation resources with open models, Frontiers in Artificial Intelligence 9 (2026). doi:10.3389/frai.2026.1807431.

[77] K.-T. Tran, D.-H. Nguyen, B. O'Sullivan, H. D. Nguyen, IRLBench: A Multi-modal, Culturally Grounded, Parallel Irish-English Benchmark for Open-Ended LLM Reasoning Evaluation, in: Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.1, ACM, 2026, pp. 2794–2805. doi:10.1145/3770854.3785693.

[78] S. Dikici, T. T. Bilgin, Small Language Models: A Systematic Review of Computational Trade-Offs, Privacy Advantages and Deployment in Intelligent Systems, Expert Systems 43 (2026). doi:10.1111/exsy.70331.

[79] S.-V. Oprea, A. Bâra, Quantized Transformers in Practice: Benchmarking Full- and Low-Precision LLMs across Two Processors, Computers, Materials & Continua 87 (2026) 1–10. doi:10.32604/cmc.2026.078985.

[80] G. Yim, N. Ko, M. Bharadwaj, Beyond One-Step Distillation: Bridging the Capacity Gap in Small Language Models via Multi-Step Knowledge Transfer, in: Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 4: Student Research Workshop), Association for Computational Linguistics, 2026, pp. 182–187. doi:10.18653/v1/2026.eacl-srw.13.

[81] S. Das, D. Rudrapal, Investigating Factors Affecting Student Model Performance in Knowledge Distillation, Procedia Computer Science 283 (2026) 718–727. doi:10.1016/j.procs.2026.06.154.

[82] M. A. Ferrag, N. Tihanyi, M. Debbah, Reasoning beyond limits: Advances and open problems for LLMs, ICT Express 11 (2025) 1054–1096. doi:10.1016/j.icte.2025.09.003.

[83] P. Bento, A. Buzelin, A. Chagas, Y. Aquino, W. Meira, G. L. Pappa, Evo-Reasoner: Evolutionary Optimization of Structured Reasoning in LLMs, Springer Nature Switzerland, 2026, pp. 150–166. doi:10.1007/978-3-032-23607-4_10.

[84] G. Oh, S. Kim, S. Park, B.-H. Kim, Model and Task-Aware Test-Time Scaling Strategies for Large Language and Vision-Language Models in Medicine: Evaluation Study, Journal of Medical Internet Research 28 (2026) e90693–e90693. doi:10.2196/90693.

[85] O. Bamgbose, M. Hashemi, S. T. Madhusudhan, J. S. Nair, A. Tiwari, V. Yadav, DNR Bench: Benchmarking Over-Reasoning in Reasoning LLMs, Proceedings of the AAAI Conference on Artificial Intelligence 40 (2026) 37240–37248. doi:10.1609/aaai.v40i44.41055.

[86] A. A. Kamalipour, S. Asadi, M. M. Amiri Chimeh, From vectors to knowledge graphs: A comprehensive analysis of modern retrieval-augmented generation architectures, Computer Science Review 61 (2026) 100925. doi:10.1016/j.cosrev.2026.100925.

[87] S. Xu, Z. Yan, C. Dai, F. Wu, MEGA-RAG: a retrieval-augmented generation framework with multi-evidence guided answer refinement for mitigating hallucinations of LLMs in public health, Frontiers in Public Health 13 (2025). doi:10.3389/fpubh.2025.1635381.

[88] M. Niu, J. Su, C. Tian, UA-RAG: Uncertainty-aware dynamic retrieval-augmented generation, Neurocomputing 699 (2026) 134357. doi:10.1016/j.neucom.2026.134357.

[89] X. Shen, L. Feng, S. Hua, D. Liu, Z. Xie, B. Liu, Towards sustainable AI knowledge-base assistants in computer science education: on-premise deployment and optimization with open educational resources, Frontiers in Psychology 17 (2026). doi:10.3389/fpsyg.2026.1843444.

[90] T. L. Föhr, M. Schreyer, K. C. Moffitt, K.-U. Marten, Deep learning meets risk-based auditing: A holistic framework for leveraging foundation and task-specific models in audit procedures, International Journal of Accounting Information Systems 57 (2026) 100758. doi:10.1016/j.accinf.2025.100758.

[91] B. Elder, A. Murthi, J. Kang, A. Naik, K. Basu, K. Kate, D. Contractor, Live API-Bench: 2500+ Live APIs for Testing Multi-Step Tool Calling, in: Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, 2026, pp. 3092–3124. doi:10.18653/v1/2026.eacl-long.143.

[92] X. Guo, Y. Zhang, B. Zhang, R. Kawahara, M. Takeuchi, Y. Zhu, UniToolBench: A Benchmark for Tool-Augmented LLMs in Cross-Domain, Universal Task Automation, in: Findings of the Association for Computational Linguistics: EACL 2026, Association for Computational Linguistics, 2026, pp. 4726–4736. doi:10.18653/v1/2026.findings-eacl.248.

[93] Y. Liu, Z. Abulawi, A. Garimidi, D. Lim, Embedding Bayesian optimization in a multi-agent large language model framework for critical heat flux modeling with uncertainty quantification, Nuclear Engineering and Technology 58 (2026) 104573. doi:10.1016/j.net.2026.104573.

[94] V. Vinay, A System-Level Taxonomy of Failure Modes in Large Language Model Applications, in: 2026 IEEE Conference on Artificial Intelligence (CAI), IEEE, 2026, pp. 715–722. doi:10.1109/CAI68641.2026.11536235.

[95] S. Yu, F. Carroll, B. L. Bentley, Operational Hallucination and Safety Drift in AI Agents, in: 2026 IEEE International Conference on AI and Data Analytics (ICAD), IEEE, 2026, pp. 1–8. doi:10.1109/ICAD69378.2026.11608655.

[96] B. Qian, Y. Wu, S. Zeng, Z. Wang, Q. Wang, MPRL: Multi-Perspective Reinforcement Learning for Enhancing Format Adherence Capability of Large Language Models, Springer Nature Singapore, 2026, pp. 570–581. doi:10.1007/978-981-92-1300-9_45.

[97] Y.-L. Yang, C.-J. Chen, T.-K. Lin, S.-C. Lin, M. Koo, Schema Enforcement and Structured-Output Stability in Locally Deployed LLMs for Clinical Admission-Note Editing: A Proxy-Based Pre-Deployment Evaluation, Healthcare 14 (2026) 2150. doi:10.3390/healthcare14142150.

[98] A. Shrimal, A. Jain, S. Chowdhury, P. Yenigalla, PARSE: LLM Driven Schema Optimization for Reliable Entity Extraction, in: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track, Association for Computational Linguistics, 2025, pp. 2749–2763. doi:10.18653/v1/2025.emnlp-industry.184.

[99] S. Mishra, Reliability-Aware Structured Synthetic Data Generation via Schema Enforcement and Layered Repair, in: Proceedings of the Workshop on Synthetic Data Generation and Management for Building AI Systems, SynthAI '26, ACM, 2026, pp. 7–12. doi:10.1145/3814574.3816747.

[100] H. Shakil, W. Armstrong, Z. H. Pugh, D. Tokita, J. Kalita, Structured grounding is critical for reliable intelligence report generation with LLMs, Expert Systems with Applications 331 (2026) 133146. doi:10.1016/j.eswa.2026.133146.

[101] Y. Du, M. Tian, S. Ronanki, S. Rongali, S. B. Bodapati, A. Galstyan, A. Wells, R. Schwartz, E. A. Huerta, H. Peng, Context Length Alone Hurts LLM Performance Despite Perfect Retrieval, in: Findings of the Association for Computational Linguistics: EMNLP 2025, Association for Computational Linguistics, 2025, pp. 23281–23298. doi:10.18653/v1/2025.findings-emnlp.1264.

[102] J. X. Zhao, J. Z. Liu, B. Hooi, S.-K. Ng, How Does Response Length Affect Long-Form Factuality, in: Findings of the Association for Computational Linguistics: ACL 2025, Association for Computational Linguistics, 2025, pp. 3102–3125. doi:10.18653/v1/2025.findings-acl.161.

[103] H. Que, W. Rong, PIC: Unlocking Long-Form Text Generation Capabilities of Large Language Models via Position ID Compression, in: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, 2025, pp. 6982–6995. doi:10.18653/v1/2025.acl-long.347.

[104] Y. Lu, J. Cheng, Z. Zhang, S. Cui, C. Wang, X. Gu, Y. Dong, J. Tang, H. Wang, M. Huang, LongSafety: Evaluating Long-Context Safety of Large Language Models, in: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, 2025, pp. 31705–31725. doi:10.18653/v1/2025.acl-long.1530.

[105] S. Kim, Y. Lee, Y. Song, K. Lee, What Really Matters in Many-Shot Attacks? An Empirical Study of Long-Context Vulnerabilities in LLMs, in: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, 2025, pp. 2043–2063. doi:10.18653/v1/2025.acl-long.101.

[106] S. Mille, J. Sedoc, Y. Liu, E. Clark, A. J. Axelsson, M. A. Clinciu, Y. Hou, S. Mahamood, I. N. Obonyo, L. Zhang, The 2024 GEM Shared Task on Multilingual Data-to-Text Generation and Summarization: Overview and Preliminary Results, in: Proceedings of the 17th International

Natural Language Generation Conference: Generation Challenges, Association for Computational Linguistics, 2024, pp. 17–38. doi:10.18653/v1/2024.inlg-genchal.2.

[107] E. Emirtekin, Large Language Model-Powered Automated Assessment: A Systematic Review, Applied Sciences 15 (2025) 5683. doi:10.3390/app15105683.

[108] S. Doddapaneni, M. S. U. R. Khan, S. Verma, M. M. Khapra, Finding Blind Spots in Evaluator LLMs with Interpretable Checklists, in: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2024, pp. 16279–16309. doi:10.18653/v1/2024.emnlp-main.911.

[109] M. Di Pumpo, L. Villani, M. R. Gualano, D. Buonsenso, F. Raffaelli, D. Donà, P. Laurenti, V. Maio, S. Boccia, W. Ricciardi, LLM-as-a-judge for infection prevention and control and antimicrobial resistance impact: comparing three main LLMs vs. human experts' assessment, Frontiers in Public Health 14 (2026). doi:10.3389/fpubh.2026.1874389.

[110] N. Li, H. Zhou, M. Xu, From Text to Insight: Leveraging Large Language Models for Performance Evaluation in Management, Personnel Psychology 79 (2026) 251–267. doi:10.1111/peps.70020.

[111] K. Qian, S. Liu, T. Li, M. Raković, X. Li, R. Guan, I. Molenaar, S. Nawaz, Z. Swiecki, L. Yan, D. Gašević, Towards reliable generative AI-driven scaffolding: Reducing hallucinations and enhancing quality in self-regulated learning support, Computers & Education 240 (2026) 105448. doi:10.1016/j.compedu.2025.105448.

[112] T. Woelfle, J. Hirt, P. Janiaud, L. Kappos, J. P. Ioannidis, L. G. Hemkens, Benchmarking Human–AI collaboration for common evidence appraisal tools, Journal of Clinical Epidemiology 175 (2024) 111533. doi:10.1016/j.jclinepi.2024.111533.

[113] E. Sánchez Salido, A. Ghajari, G. Marco, J. Gonzalo, J. Abizanda, R. Morante, A. Benito-Santos, L. Plaza, J. Carrillo-de Albornoz, V. Fresno, E. Amigó, A. Fernández García, To What Extent Is LLM Performance on Multiple-Choice Questions Driven by Data Leakage? A Case Study with Contamination-Controlled Spanish Undergraduate Exams, Inteligencia Artificial 29 (2026) 131–151. doi:10.4114/intartif.vol29iss77pp131-151.

[114] I. Koleiev, R. Stratiichuk, N. Shevchuk, M. Melnychenko, A. Nyporko, D. Todoryshyn, V. Husak, S. Starosyla, S. Yesylevskyy, A. Nafiiev, An End-User Audit of Reproducibility, Data Leakage, and Overfitting of the Top-Ranked ADMET Prediction Models in TDC Leaderboards, Journal of Chemical Information and Modeling 66 (2026) 8045–8058. doi:10.1021/acs.jcim.6c00819.

[115] R. Gómez, C. E. Miranda, J.-A. Romero-González, D.-M. Córdova-Esparza, G. Alfonso-Francia, E.-A. Chávez-Urbiola, A. Ramirez-Pedraza, J. Terven, Large Language Model Benchmarks: A Taxonomy of Capabilities, Scientific Quality Assessment, and Saturation Analysis, Machine Learning and Knowledge Extraction 8 (2026) 141. doi:10.3390/make8060141.

[116] J. Xu, X. Zhou, Y. Yang, T. Liu, J. Xu, C. Zhang, L. Yang, L. Dong, Enhancing trustworthiness evaluation of Large Language Models through dataset refinement, Information and Software Technology 199 (2026) 108249. doi:10.1016/j.infsof.2026.108249.

[117] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, T. Liu, A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions, ACM Transactions on Information Systems 43 (2025) 1–55. doi:10.1145/3703155.

[118] M. P. Priola, Addressing hallucinations with RAG and NMISS in Italian healthcare LLM chatbots, Data & Knowledge Engineering 165 (2026) 102627. doi:10.1016/j.datak.2026.102627.

[119] Y. Tian, D. Li, S. Xue, A Survey of Jailbreaking Attacks on Large Language and Multimodal Models, in: Proceedings of the 2025 8th Artificial Intelligence and Cloud Computing Conference, AICCC 2025, ACM, 2025, pp. 142–150. doi:10.1145/3789982.3790001.

[120] B. Knowlton, J. Campa, D. S. Gallo, K. Dajani, N. Alzahrani, Prompt-Based Jailbreaking of Leading LLM Chatbots: A Survey of Attacks and Defenses, IEEE Transactions on Artificial Intelligence 7 (2026) 4237–4251. doi:10.1109/TAI.2026.3665656.

[121] J. Zheng, H. Liu, L. Zhou, H. Yan, P. Liu, Boosting Jailbreak Transferability for Large Language Models, Springer Nature Singapore, 2026, pp. 445–458. doi:10.1007/978-981-95-5761-5_31.

[122] Y. Xue, J. Wang, Z. Yin, Y. Ma, H. Qin, R. Tao, X. Liu, Dual Intention Escape: Penetrating and

Toxic Jailbreak Attack against Large Language Models, in: Proceedings of the ACM on Web Conference 2025, WWW '25, ACM, 2025, pp. 863–871. doi:10.1145/3696410.3714654.

[123] B. An, S. Zhang, M. Dredze, RAG LLMs are Not Safer: A Safety Analysis of Retrieval-Augmented Generation for Large Language Models, in: Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), Association for Computational Linguistics, 2025, pp. 5444–5474. doi:10.18653/v1/2025.naacl-long.281.

[124] B. Ma, H. Guo, P. Lv, M. Xu, X. Dai, Y. Zhang, Y. Yang, Y. Zhang, What breaks embodied AI security: LLM vulnerabilities, CPS flaws, or something else?, High-Confidence Computing 6 (2026) 100403. doi:10.1016/j.hcc.2026.100403.

[125] H. Alqahtani, G. Kumar, Large Language Models for Cybersecurity Intelligence: A Systematic Review of Emerging Threats, Defensive Capabilities, and Security Evaluation Frameworks, Computers, Materials & Continua 87 (2026) 1–10. doi:10.32604/cmc.2026.077367.

[126] S. Liu, C. Li, J. Qiu, F. Huang, Z. Chen, L. Zhang, Y. Hei, J. Zhu, X. Zhang, The scales of justitia: A comprehensive survey on safety evaluation of LLMs, Information Fusion 136 (2026) 104579. doi:10.1016/j.inffus.2026.104579.

[127] S. Dathathri, A. See, S. Ghaisas, P.-S. Huang, R. McAdam, J. Welbl, V. Bachani, A. Kaskasoli, R. Stanforth, T. Matejovicova, J. Hayes, N. Vyas, et al., Scalable watermarking for identifying large language model outputs, Nature 634 (2024) 818–823. doi:10.1038/s41586-024-08025-4.

[128] Z. Jiang, H. Wang, Q. Hou, R. Jiao, PGmark : Enhancing text quality in language model watermarking via probability guidance, Expert Systems with Applications 308 (2026) 131120. doi:10.1016/j.eswa.2026.131120.

[129] L. An, Y. Liu, Y. Liu, Y. Bu, Y. Zhang, S. Chang, A Reinforcement Learning Framework for Robust and Secure LLM Watermarking, in: Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, 2026, pp. 7181–7198. doi:10.18653/v1/2026.eacl-long.338.

[130] C. McMichael, D. Roussinov, S. Sharoff, Discourse-Level Watermarking using Rhetorical Structure Theory, in: 2026 8th International Conference on Natural Language Processing (ICNLP), IEEE, 2026, pp. 291–298. doi:10.1109/ICNLP69856.2026.11527712.

[131] M. S. Al-Shaibani, M. Ahmed, Arabic machine-generated text detection: Stylometric analysis and cross-model evaluation, Expert Systems with Applications 305 (2026) 130644. doi:10.1016/j.eswa.2025.130644.

[132] F. Kang, N. Ardalani, M. Kuchnik, Y. Emad, M. Elhoushi, S. Sengupta, S.-W. Li, R. Raghavendra, R. Jia, C.-J. Wu, Demystifying Synthetic Data in LLM Pre-training: A Systematic Study of Scaling Laws, Benefits, and Pitfalls, in: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2025, pp. 10750–10769. doi:10.18653/v1/2025.emnlp-main.544.

[133] Z. Hu, M. Rostami, J. Thomason, Multimodal Synthetic Data Finetuning and Model Collapse: Insights from VLMs and Diffusion Models, in: Proceedings of the 27th International Conference on Multimodal Interaction, ICMI '25, ACM, 2025, pp. 588–599. doi:10.1145/3716553.3750806.

[134] R. Qamar, B. Appasani, N. Bizon, S. Verma, A. V. Jha, Capability Degradation in Small Language Models Under Low-Signal Data Corruption, in: 2026 18th International Conference on Electronics, Computers and Artificial Intelligence (ECAI), IEEE, 2026, pp. 1–8. doi:10.1109/ECAI69016.2026.11613657.

[135] Z. Ji, M. Jiang, A systematic review of electricity demand for large language models: evaluations, challenges, and solutions, Renewable and Sustainable Energy Reviews 225 (2026) 116159. doi:10.1016/j.rser.2025.116159.

[136] M.-K. Kim, T.-A. Yoo, J.-B. Chung, Toward Sustainable AI: A Scoping Review of Carbon Footprint and Environmental Impacts Across Training and Inference Stages, IEEE Access 14 (2026) 18881–18896. doi:10.1109/ACCESS.2026.3659894.

[137] S. B. M. Ali, S. M. Danish, Z. Sattar, A. Asad, Assessing the Sustainability of LLM Inference through Energy–Accuracy Analysis, in: Proceedings of the 17th ACM International Conference

on Future and Sustainable Energy Systems, E-Energy '26, ACM, 2026, pp. 627–635. doi:10.1145/3744255.3811741.

[138] J. Zschache, T. Hartwig, Comparing energy consumption and accuracy in text classification inference, Scientific Reports 16 (2026). doi:10.1038/s41598-026-45023-0.

[139] Y. He, Y. Bu, Academic journals' AI policies fail to curb the surge in AI-assisted academic writing, Proceedings of the National Academy of Sciences 123 (2026). doi:10.1073/pnas.2526734123.

[140] J. Oh, Publication ethics issues in research papers using generative artificial intelligence: a narrative review, Journal of the Korean Medical Association 69 (2026) 126–135. doi:10.5124/jkma.25.0157.

[141] Y. Fedorchenko, A. Usen, B. Seiil, O. Zimba, M. Korkosz, Editorial policies for ethical use of artificial intelligence in rheumatology journals, Seminars in Arthritis and Rheumatism 79 (2026) 152966. doi:10.1016/j.semarthrit.2026.152966.

[142] R. Giusti, M. Filetti, P. Lombardi, F. Lo Bianco, S. Sganga, T. Giovagnoli, G. M. Iannantuono, A. Spinazzola, D. Santini, M. Mariniello, G. Porzio, M. Ibrahim, et al., Artificial intelligence in oncology publishing: a systematic review and policy analysis of high-impact journals, Frontiers in Oncology 16 (2026). doi:10.3389/fonc.2026.1717048.

[143] T. Anikina, J. Cegin, J. Simko, S. Ostermann, A Rigorous Evaluation of LLM Data Generation Strategies for Low-Resource Languages, in: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2025, pp. 8293–8314. doi:10.18653/v1/2025.emnlp-main.418.

[144] A. Dhawan, C. Driggers-Ellis, C. Grant, D. Z. Wang, Improving Indigenous Language Machine Translation with Synthetic Data and Language-Specific Preprocessing, in: Proceedings for the Ninth Workshop on Technologies for Machine Translation of Low Resource Languages (LoResMT 2026), Association for Computational Linguistics, 2026, pp. 119–126. doi:10.18653/v1/2026.loresmt-1.10.

[145] L. Petersen, M. Beck, M. Andersen, J. Xu, F. Bruun, Prompt engineering enables open-source large language models to match proprietary models in diagnostic accuracy for annotation of radiology reports, Clinical Radiology 97 (2026) 107315. doi:10.1016/j.crad.2026.107315.

[146] Y. Zhuang, Q. Liu, Enhancing Chinese-to-English legal translation by generative artificial intelligence: a corpus-informed prompt engineering approach, Perspectives (2026) 1–22. doi:10.1080/0907676X.2026.2696986.

[147] A. Saini, A. Chernodub, V. Raheja, V. Kulkarni, Spivavtor: An Instruction Tuned Ukrainian Text Editing Model, in: Proceedings of the Third Ukrainian Natural Language Processing Workshop (UNLP) @ LREC-COLING 2024, LREC-COLING, European Language Resources Association (ELRA) and ICCL, 2024, pp. 95–108. doi:10.63317/59qw7bvuvej8.

[148] M. Romanyshyn, O. Syvokon, R. Kyslyi, The UNLP 2024 Shared Task on Fine-Tuning Large Language Models for Ukrainian, in: Proceedings of the Third Ukrainian Natural Language Processing Workshop (UNLP) @ LREC-COLING 2024, LREC-COLING, European Language Resources Association (ELRA) and ICCL, 2024, pp. 67–74. doi:10.63317/57pbivswb65o.

[149] A. Kiulian, A. Polishko, M. Khandoga, O. Chubych, J. Connor, R. Ravishankar, A. Shirawalmath, From Bytes to Borsch: Fine-Tuning Gemma and Mistral for the Ukrainian Language Representation, in: Proceedings of the Third Ukrainian Natural Language Processing Workshop (UNLP) @ LREC-COLING 2024, LREC-COLING, European Language Resources Association (ELRA) and ICCL, 2024, pp. 83–94. doi:10.63317/5dowr7hgrb4v.

[150] Y. Paniv, Isolating LLM Performance Gains in Pre-training versus Instruction-tuning for Mid-resource Languages: The Ukrainian Benchmark Study, in: Natural Language Processing in the Generative AI Era, volume 2025 of *RANLP 2025*, Incoma Ltd. Shoumen, BULGARIA, 2025, pp. 876–883. doi:10.26615/978-954-452-098-4-100.

[151] D. Maksymenko, O. Turuta, Tokenization efficiency of current foundational large language models for the Ukrainian language, Frontiers in Artificial Intelligence 8 (2025). doi:10.3389/frai.2025.1538165.

[152] S. F. Chen, A. Alyakin, A. Seas, E. Yang, J. J. Choi, J. V. Lee, A. L. Chen, P. I. Warman, R. T. Bitolas,

R. J. Steele, D. A. Alber, E. K. Oermann, LLM-assisted systematic review of large language models in clinical medicine, Nature Medicine 32 (2026) 1152–1159. doi:10.1038/s41591-026-04229-5.

[153] W. Zhang, B. Jiang, Y. Fu, H. Cheng, M. Hummel, V. Scotti, N. Hagel, J. Li, G. Grossmann, M. Stumptner, R. Hebig, D. Strüber, et al., Large language models in model-driven engineering: a systematic mapping study, Empirical Software Engineering 32 (2026). doi:10.1007/s10664-026-10921-4.

[154] Z. Zhou, W. Gao, Y. Li, J. Yu, Understanding the interaction between human and LLMs-based AI in creative tasks: Insights from UX professionals' collaboration and communication with ChatGPT, Design Studies 104 (2026) 101403. doi:10.1016/j.destud.2026.101403.

[155] S. Maier, M. Schneider, S. Feuerriegel, Partnering with Generative AI: Experimental Evaluation of Model-Led and Human-Led Interaction in Human-AI Co-Creation, in: Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems, CHI '26, ACM, 2026, pp. 1–28. doi:10.1145/3772318.3791185.

[156] Y. S. Park, N. A. P. Arvi, S. Kim, J. Kim, Authorship Drift: How Self-Efficacy and Trust Evolve During LLM-Assisted Writing, in: Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems, CHI '26, ACM, 2026, pp. 1–18. doi:10.1145/3772318.3790276.

[157] Y. Jiao, Q. Wang, Large language models for formative feedback in writing instruction: a systematic review of classroom interventions, feedback quality, and student outcomes, Frontiers in Education 11 (2026). doi:10.3389/feduc.2026.1834085.

[158] D.-W. Zhang, X. Hong, Y. Qi, When and how does pedagogically enriched LLM feedback outperform corrective automated writing evaluation? A learning trajectory analysis of writing improvement, Computers in Human Behavior Reports 22 (2026) 101121. doi:10.1016/j.chbr.2026.101121.