Advances in neural text generation: A systematic review (2022-2024)

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]*Kryvyi Rih National University, 11 Vitalii Matusevych Str., Kryvyi Rih, 50027, Ukraine*
[5]*Academy of Cognitive and Natural Sciences, 54 Universytetskyi Ave., Kryvyi Rih, 50086, Ukraine*

Abstract

Recent years have witnessed significant advancements in neural text generation driven by the emergence of large language models and growing interest in this field. This systematic review aims to identify and summarize current trends, approaches, and methods in neural text generation from 2022 to 2024, complementing the findings of a previous review covering 2015-2021. Following the PRISMA methodology, 43 articles were selected from the Scopus database for analysis. The review reveals a shift towards innovative model architectures like Transformer-based models (GPT-2, GPT-3, BERT), attention mechanisms, and controllable text generation. While BLEU, ROUGE, and human evaluation remain the most popular evaluation metrics, new metrics like BERTScore have emerged. Datasets span diverse domains and data types, with growing interest in unlabeled data. Applications have expanded to areas such as table-to-text generation, knowledge graph-based generation, and medical text generation. Although English dominates, there is increasing research on low-resource languages. The findings highlight the rapid evolution of neural text generation methods, the broadening of application areas, and promising avenues for future research.

Keywords

neural text generation, deep learning, systematic review, natural language processing, evaluation metrics, datasets, applications, low-resource languages

1. Introduction

1.1. Problem statement

Natural Language Processing (NLP) is an interdisciplinary field of computer science and linguistics [1, p. 1], the classification of the main tasks of which is shown in figure 1.

Text content generation is a branch of NLP that combines computational linguistics and artificial intelligence to generate text [2, p. 53490].

In 2022, OpenAI [3] introduced ChatGPT, a chatbot based on the GPT model that provided a natural language interface to the user. In most systematic reviews, similar questions are considered, which explains the choice of our review type.

The previous review "A Systematic Literature Review on Text Generation Using Deep Neural Network Models" [2] covered 90 sources from 2015 to 2021. The emergence of access to large language models in 2022-2023 [4] led to an increase in interest in them (figure 2), so there was a need to supplement the previous review, the main result of which is a classification (figure 3):

1) *by neural network architecture*:
   - traditional:
     - RNN – Recurrent Neural Network, used for sequential data;

CS&SE@SW 2024: 7th Workshop for Young Scientists in Computer Science & Software Engineering, December 27, 2024, Kryvyi Rih, Ukraine
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)

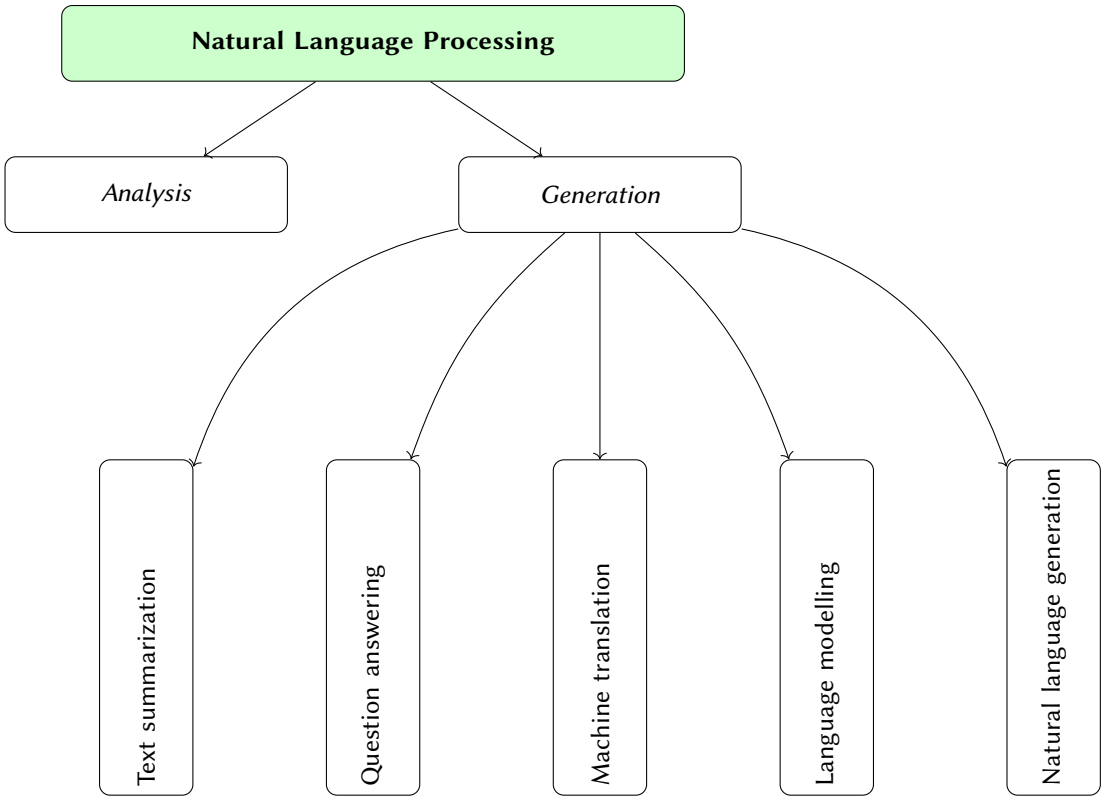

Figure 1: Taxonomy of popular NLP tasks for text generation (based on [1, p. 4]).

  - LSTM – Long Short-Term Memory network, works better than RNN for larger data volumes;
  - GRU – Gated Recurrent Unit (simplified version of LSTM);
  - CNN – Convolutional Neural Network.
- innovative:
  - Attention Based – networks that use an attention mechanism to increase the importance of input data;
  - Transformer – networks that use an attention mechanism without recurrent or convolutional layers;
  - BERT – a neural network developed by Google that combines attention mechanisms without recurrent or convolutional layers with bidirectional encoders

2) *by quality metrics*:

- human-centered:
  - Domain-Expert – involving a person who is an expert in the given field to validate the results.
- machine-centered (automatic):
  - BLEU (bilingual evaluation understudy) – compares the number and value of tokens (lexemes) of machine and human translation; the meaning of words is not taken into account;
  - ROUGE (Recall-Oriented Understudy for Gisting Evaluation) – compares machine-generated and human-generated summaries/translations;
  - Cosine Similarity – comparison of the cosines of the angle of two non-zero vectors: a value of +1 corresponds to unidirectional proportional vectors, -1 corresponds to oppositely directed proportional vectors;

  - Content Selection – a metric similar to ROUGE that uses an attention mechanism for a given task;
  - Diversity Score – a metric for evaluating diversity.

3) *by application of the neural network*:
   - AMR (Abstract Meaning Representation) – extracting semantic relationships from text;
   - Language Generation – generating human-like text;
   - Speech-to-text – converting speech to text;
   - Script Generation – generating scripts based on given words;
   - Machine Translation – generating machine translation of text from one language to another;
   - Text Summarization – generating a summary for a given text;
   - Image Captioning – generating a description for a given image;
   - Shopping Guide – generating an advertising description for a given product image;
   - Weather Forecast – generating a weather forecast text.
4) *by generation language*:
   - well-resourced: English, Chinese;
   - low-resourced: Bengali, Korean, Balinese, Spanish, Hindi, Slovak, Macedonian.
5) *by dataset for training the neural network*:
   - by annotation type:
     - Labeled (labeled data);
     - Unlabeled (unlabeled data);
   - by type:
     - Sentence – sentence;
     - Paragraph – paragraph;
     - Question/answer – question and answer type data;
     - Document – document type data.

1.2. Research tasks and questions

To obtain the results presented in figure 3, Fatima et al. [2] set the following tasks:

1. To investigate the existing traditional and advanced deep learning-based text generation approaches/techniques.
2. To explore various performance metrics used for evaluating text generation models.
3. To investigate various evaluation methods for measuring the quality of generated text.
4. To review the recent application domains where text generation is being applied.
5. To discuss the major challenges and future research directions in the text generation domain.

To supplement the results obtained by Fatima et al. [2], these *research tasks* were refined:

1. To explore deep learning methods (approaches, architectures) for text generation that have appeared or were mentioned in the works of 2022-2024.
2. To consider metrics for evaluating the effectiveness of text generation models that have appeared or were mentioned in the works of 2022-2024.
3. To identify text generation datasets described in the works of 2022-2024.
4. To explore new text generation applications described in the works of 2022-2024.
5. To determine which natural languages were used for text generation in the works of 2022-2024.

Similarly, the *research questions* were refined:

RQ1. What advanced deep learning methods are used for text generation in the literature of 2022-2024?
RQ2. What new metrics for evaluating the effectiveness of text generation models are there in the literature of 2022-2024?
RQ3. What text generation datasets are described in the literature of 2022-2024?
RQ4. What new text generation applications are described in the literature of 2022-2024?
RQ5. What natural languages are used for text generation in the literature of 2022-2024?

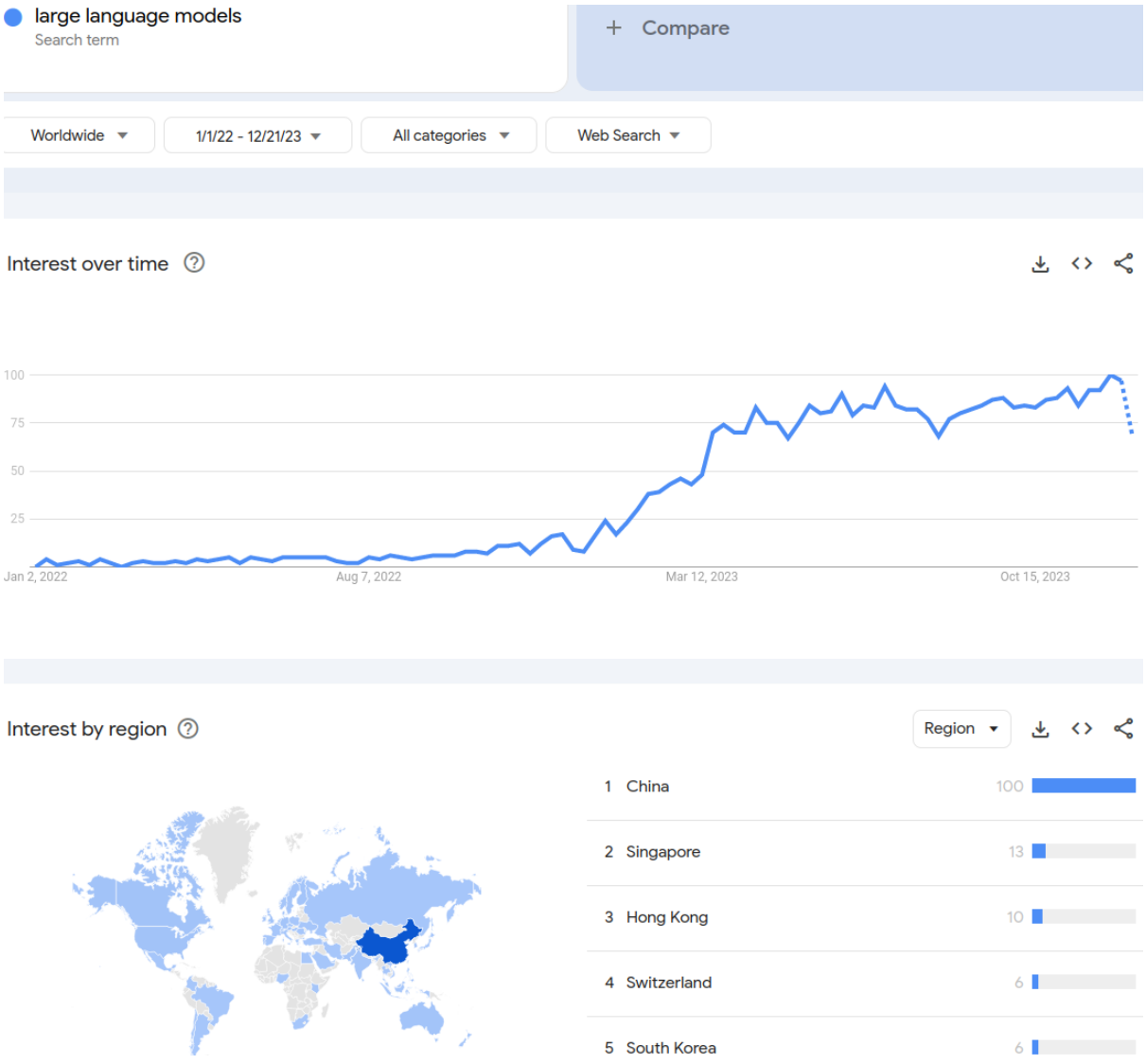

Figure 2: Dynamics of search queries for the term “large language models” [4].

2. Methodology

Systematic literature analysis is the main **method of this research**, which allows generalising and synthesising information from a large number of scientific publications (secondary sources) according to a clearly defined methodology. The PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) methodology, which is a generally recognised standard for systematic reviews and meta-analyses in various fields of science, was chosen for conducting the review [5]. Systematic analysis according to the PRISMA methodology involves clear research planning, defining criteria for the selection of publications, conducting a thorough literature search in leading scientific databases, selecting relevant studies, extracting and synthesising data. This approach ensures the completeness, reliability and reproducibility of the obtained results.

The chosen method fully corresponds to the aim and objectives of the research, allowing to obtain a generalised picture of the current state of research in the field of text content generation based on the analysis of a significant array of scientific publications in recent years.

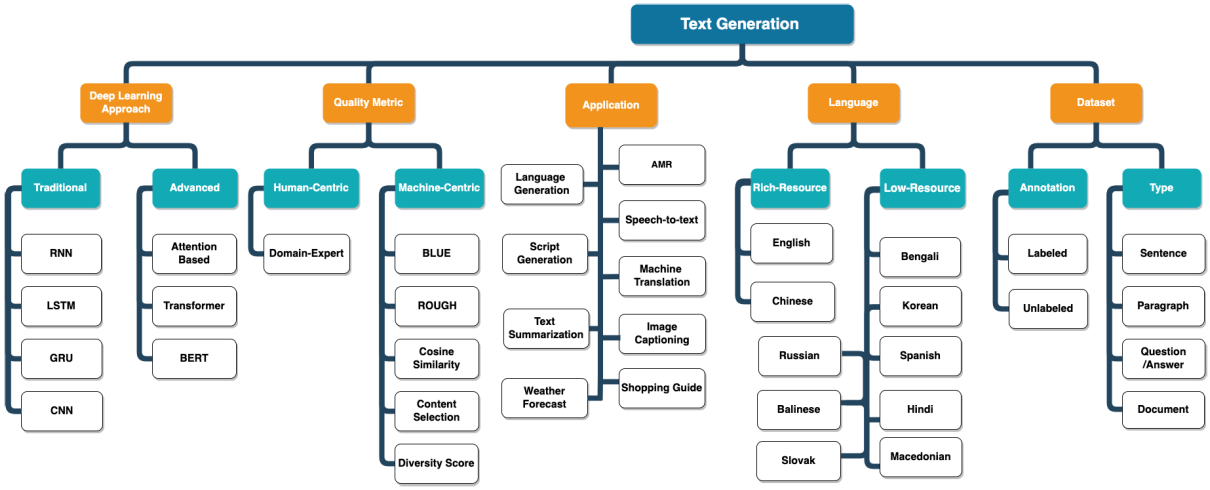

Figure 3: Taxonomy of the text generation [2, p. 53493].

2.1. Information sources and search strategy

Fatima et al. [2] in the previous review used 2 scientometric databases (Web of Science and Scopus) and 4 libraries (IEEE Xplore, SpringerLink, ScienceDirect and ACM Digital Library) as reliable data sources. The search query for article titles, abstracts and keywords used by Fatima et al. [2] is presented in table 1.

Table 1
Groups of selected keywords [2, p. 53494].

Group 1: Words related to deep learning	deep learning OR natural language processing OR NLP OR neural network OR RNN OR Recurrent OR Recursive OR LSTM OR GAN OR GPT-2 OR generative adversarial network
Group 2: Words related to text generation	text generation OR language generation OR language modelling OR natural language generation OR neural language generation
Search query	(Group 1) and (Group 2)

Currently, Scopus covers about 90% of IEEE Xplore and ACM Digital Library, Web of Science – about 50%; ScienceDirect and Scopus have the same owner – Elsevier. Given that Scopus includes a significant part of these libraries, only one database – Scopus – was used instead of 2 databases and 4 libraries. Applying the search query from the previous review (Table 1) yields 2580 documents for 2015–2020 (versus 100 documents specified in [2, p. 53494]). When searching only in article titles, the number of documents decreases to 109 and there is a partial match with the list of sources [2, p. 53500-53503]).

The inability to reproduce the previous results for the query from table 1 prompted the creation of a new query:

```
(
    TITLE-ABS-KEY(neural network)
    OR
    TITLE-ABS-KEY(machine learning)
    OR
    TITLE-ABS-KEY(deep learning)
```

```
)
AND
TITLE("text generation")
```

The first part of the query was simplified to three key phrases, two of which ("neural network" and "deep learning") match the first group of table 1, and the third ("machine learning") generalizes all other keywords of the first group, including those that did not exist at the time of the previous review. The second part of the query included only the key phrase "text generation", the search for which is performed in document titles (TITLE), and not in titles, abstracts and author keywords (TITLE-ABS-KEY).

2.2. Document inclusion and exclusion criteria

Inclusion criteria:

1. Documents published between 2022 and 2024.
2. Documents related to text generation using artificial neural networks.
3. Documents describing approaches, architectures, quality metrics, languages, datasets or applications of text generation.

Exclusion criteria:

1. Documents published before 2022 or those that do not contain data for 2022-2024.
2. Documents that are not related to text generation or do not use artificial neural networks.
3. Documents that do not contain relevant information regarding the posed research questions (new methods, metrics, datasets, applications, natural languages).

2.3. Document selection process

The Scopus query on 04.03.2024 returned 248 documents, the distribution of which by year is shown in figure 4. Of these, 2 were duplicates and 157 were dated before 2022, so they were excluded from the list for obtaining.

Figure 5 presents a scheme of data selection for the systematic review.

An attempt was made to obtain 89 documents from publishers' websites, the scientific social network ResearchGate, and preprint servers (primarily arXiv). 41 documents (primarily from the ACM Digital Library and IEEE Xplore) could not be obtained. Thus, 48 documents were selected for evaluation, the review of which revealed 1 document that did not contain data for 2022-2024, and 4 documents that did not contain relevant information regarding the posed research questions.

43 documents were selected for review: [6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 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, 48]. The review of each document was performed according to the review map (appendix A). To automate data extraction for the map questions, a large language model (LLM) Claude 3 Sonnet [49] was used, to which the document file in PDF format was fed with the following prompt:

```
Describe the article according to the following characteristics:
Document type:
journal article (ARTICLE) or conference proceedings article (CONFERENCE)
Title
Year of publication
Countries represented by the authors
Article purpose
Used neural network architectures
Used quality metrics
Characteristics of the used datasets - name
```

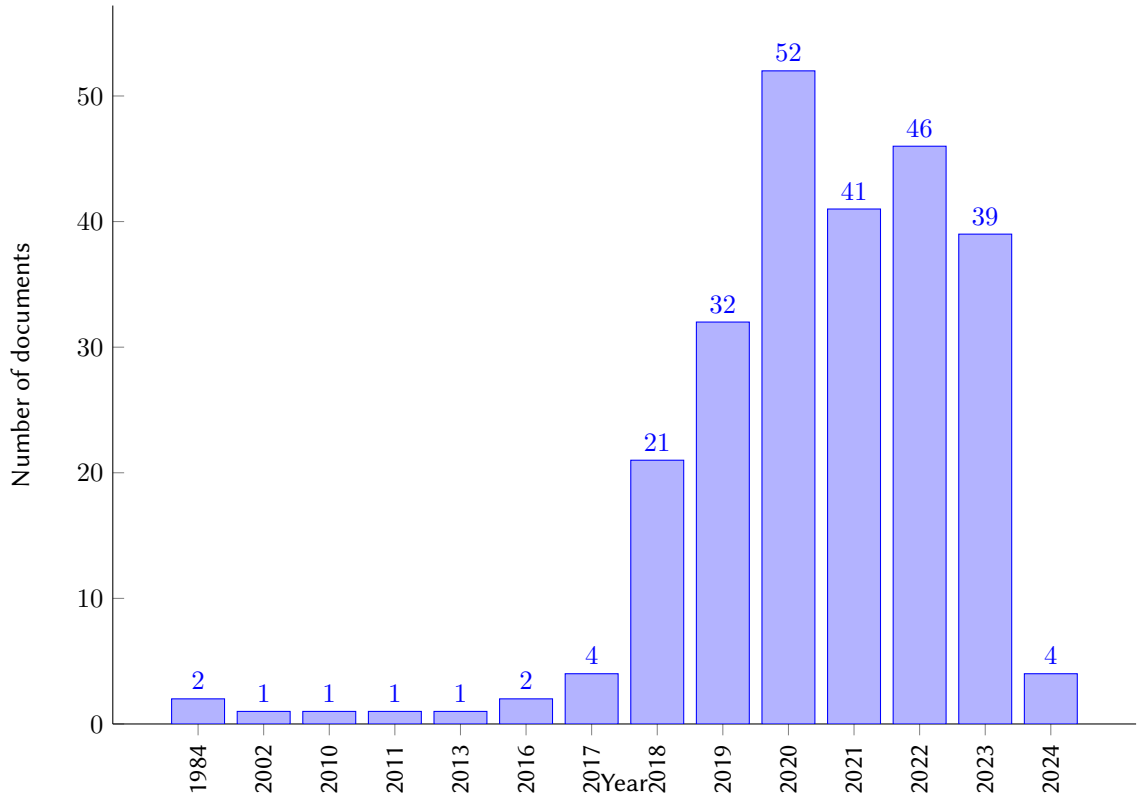

Figure 4: Distribution of search results by year.

```
Characteristics of the used datasets - data type:
sentence, paragraph, document, question-answer, not specified
Characteristics of the used datasets - size
Characteristics of the used datasets - format:
CSV, JSON, XML, files, not specified
Characteristics of the used datasets - by annotation type:
labeled data, unlabeled data
Characteristics of the used datasets - data quality:
raw (unprocessed), preprocessed
Characteristics of the used datasets - by availability:
publicly available, private, not specified
Characteristics of the used datasets - link
Solved text generation task (what was the neural network used for)
Language of text generation
```

An example of a response is shown in figure 6.

For external verification, the LLM GPT 4 was used through the Scholar GPT [50] extension. Table 2 compares the results of the automated review of the article [47].

Comparison of the description of the article [47], made with the help of another LLM and verified by a human, with the results in table 2 shows that the description is well consistent with the results of both Claude 3 Sonnet and GPT-4. Both models accurately identified the document type, title, year of publication, countries of authors, purpose of the article, neural network architectures used, quality metrics, and solved text generation tasks. Regarding dataset characteristics, both models indicated that details about specific datasets are not provided, with the exception of the Quora dataset for paraphrasing. They also noted that the article focuses on reducing dependence on labeled or publicly available datasets, although both labeled and unlabeled data are discussed in different contexts.

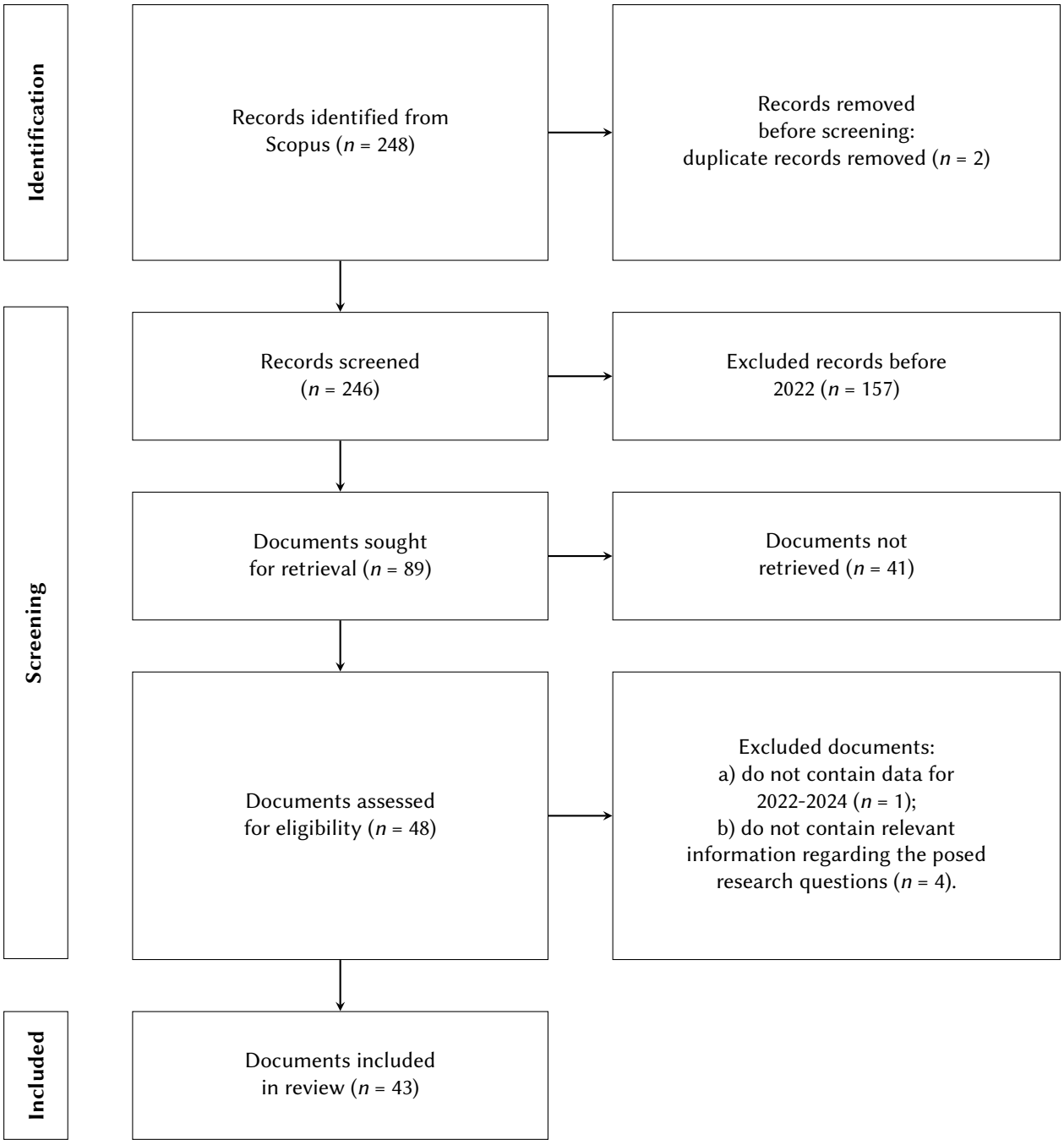

Figure 5: Scheme of data selection for the systematic review (according to the PRISMA methodology [5]).

Table 2: Comparison of the results of the large language models Claude 3 Sonnet and GPT 4.

Query	Claude 3 Sonnet Response	GPT 4 Response
Document type: journal article (ARTICLE) or conference proceedings article (CONFERENCE)	ARTICLE (journal article)	Article in a journal
Title	Search and learning for unsupervised text generation	Search and Learning for Unsupervised Text Generation
Year of publication	2022	2022

Continued on next page

Table 2 – continued from previous page

Query	Claude 3 Sonnet Response	GPT 4 Response
Countries represented by the authors	Canada	Canada (The author, Lili Mou, is affiliated with the University of Alberta in Edmonton, Canada)
Article purpose	To introduce a search and learning approach for unweighted text generation, where the task of text generation is formulated as a search problem with a heuristic function for evaluating the quality of a candidate sentence, and discrete local search algorithms are used to maximize this evaluation function. A machine learning model is also trained on the search results to smooth out noise and improve efficiency.	The article aims to introduce approaches to unsupervised text generation using search and learning methodologies. It focuses on overcoming the need for massive labeled datasets, which are required in supervised methods, thereby making text generation feasible for low-resource languages and reducing the human labor required for data annotation.
Used neural network architectures	Recurrent Neural Networks, Transformers, in particular GPT-2	Neural architectures mentioned include recurrent neural networks (RNNs), Transformer architecture, and language models such as RoBERTa and pretrained language models by Radford et al., 2019.
Used quality metrics	iBLEU (for evaluating paraphrasing)	Metrics such as language fluency, semantic coherency, and task-specific constraints (like sentence simplicity scores) are used to evaluate the quality of generated texts.
Characteristics of the used datasets		
- name	Quora dataset (for the task of paraphrasing). No names specified for other tasks	Not specified in detail for individual datasets.
- data type: sentence, paragraph, document, question-answer, not specified	sentence, paragraph (for the task of summarization)	Sentences, as the text generation involves tasks like paraphrasing, summarization, etc.
- size	Not specified	Specific sizes are not mentioned, but the use of large-scale generic corpora is implied.
- format: CSV, JSON, XML, files, not specified	Not specified	Not explicitly mentioned, but it includes labeled input-output pairs in various formats typical for NLP tasks.
- by annotation type: labeled data, unlabeled data	unlabeled data	Primarily deals with labeled data in supervised contexts, though the focus here is on reducing dependence on such datasets.
- data quality: raw (unprocessed), preprocessed	Not specified	The discussion includes the use of both raw and preprocessed data in different contexts.
- publicly available, private, not specified	Not specified	The text suggests a shift towards methods that do not require publicly available or parallel datasets.
- link	Not specified	No specific dataset URLs provided.
Solved text generation task (what was the neural network used for)	Paraphrasing, summarization, text simplification, sentence generation from keywords, grammar error correction, style transfer	The neural networks are applied to generate text for various applications including paraphrasing, summarization, text simplification, and creating sentences from keywords, all in an unsupervised manner.
Language of text generation	English	The tasks are primarily demonstrated with English examples, though the methodologies are applicable to multiple languages.

Thus, LLMs can accurately extract key information from articles, although they may sometimes omit details that are not explicitly stated in the text. To minimize the risk of incorrect information extraction, a human review of Claude 3 Sonnet's results was performed. To avoid issues related to terminology

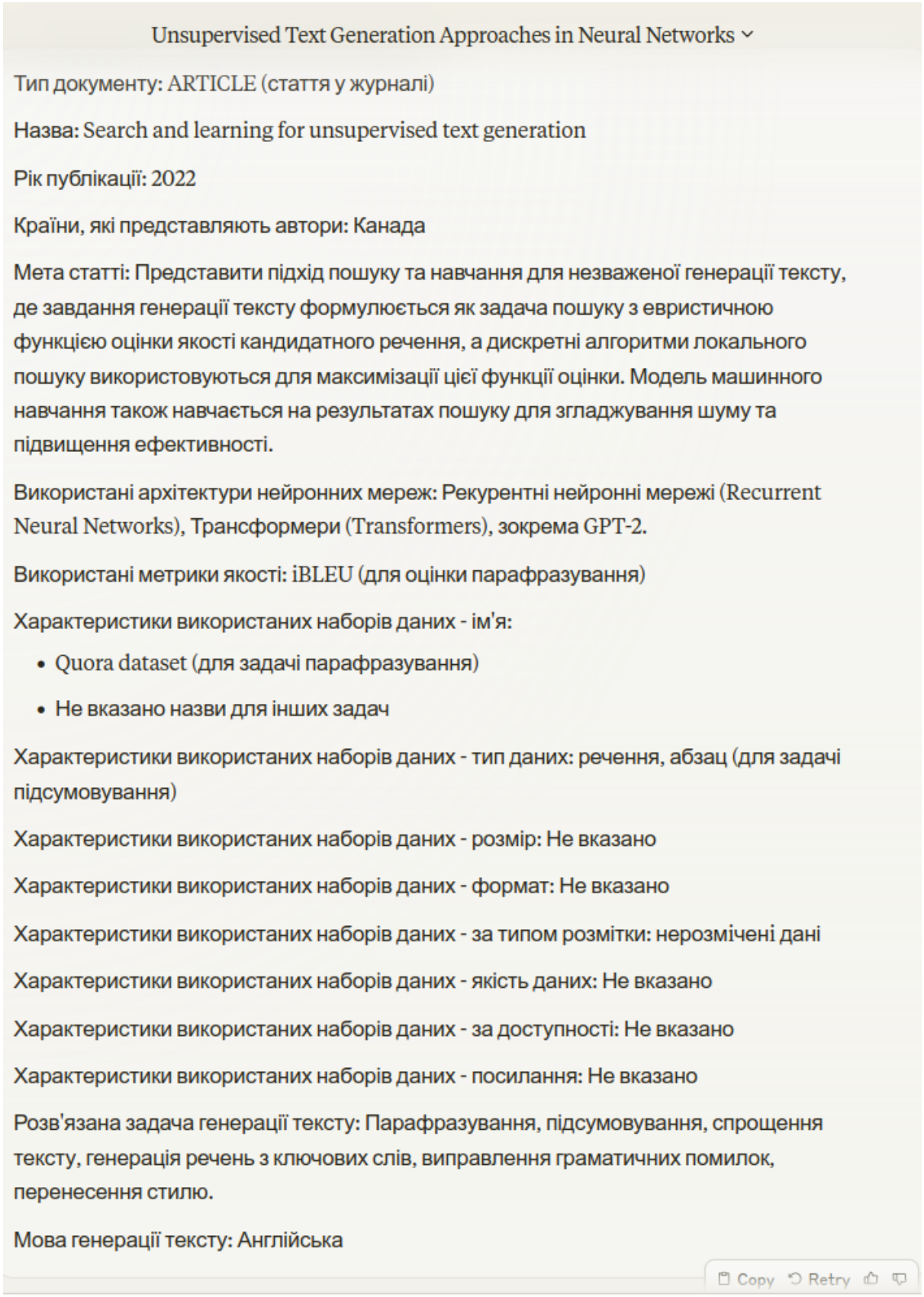

Figure 6: Claude 3 Sonnet's response to the request to compile a review map of the article [47].

translation, LLM responses were additionally requested in the language of the selected documents (English).

2.4. Quality assessment

To assess the quality of the research selection and analysis process in this review, the following criteria were applied:

1. Clarity and relevance of the research inclusion and exclusion criteria to the purpose of the review.
2. Completeness and systematic nature of the search for relevant research in the selected databases.
3. Consistency and reproducibility of the research selection process according to the inclusion and exclusion criteria.
4. Application of a standardized review map for collecting and systematizing data from selected studies.
5. Involvement of at least two independent researchers in the process of data selection, analysis, and synthesis to minimize the risk of bias.
6. Consideration and description of any discrepancies or uncertainties in the process of research selection and analysis.
7. Ensuring transparency and reproducibility of the review process by detailed description of each stage in the report.

Adherence to these quality criteria made it possible to ensure the reliability and validity of the results and conclusions of this systematic review.

PRISMA provides for the presence of the following additional components in the research methodology:

- *assessment of the risk of bias in the selected studies* is not relevant because this review considers different approaches and methods of text generation, and does not compare the results of individual studies;
- *determination of the effect size for each outcome (or type of outcome)* is not performed because this review does not aim to conduct a meta-analysis or quantitative synthesis of the results;
- *description of the methods of synthesizing research results*, such as meta-analysis, is not performed because the review does not involve a quantitative synthesis of the results;
- *assessment of the risk of bias due to incomplete presentation of the results in publications* is not performed because this review focuses on describing and classifying existing approaches and methods.
- *assessments of the reliability and trustworthiness of the results* obtained from publications are not performed due to the use of reliable sources: publications selected by Scopus.

3. Results

3.1. Distribution of selected documents by year

In [51], the completed review maps for each article are presented. The results of individual studies are not provided because this review does not aim to conduct a meta-analysis or quantitative synthesis of the results.

As can be seen from figure 7, the number of articles in journals (ARTICLE) prevails over the number of conference proceedings articles (CONFERENCE) during 2022-2024. In 2022, the number of conference proceedings documents (15) was significantly higher than the number of articles in journals (4), but in 2023 there is an increase in the number of articles in journals (16) compared to conference proceedings articles (6). For January and February 2024, there are only articles in journals (2), and conference proceedings articles are absent. In total, for the period 2022-2024, the number of articles in journals

(22) is equal to the number of conference proceedings articles (21). The increase may indicate a more thorough coverage of the issue in scientific journals compared to conference proceedings in recent years.

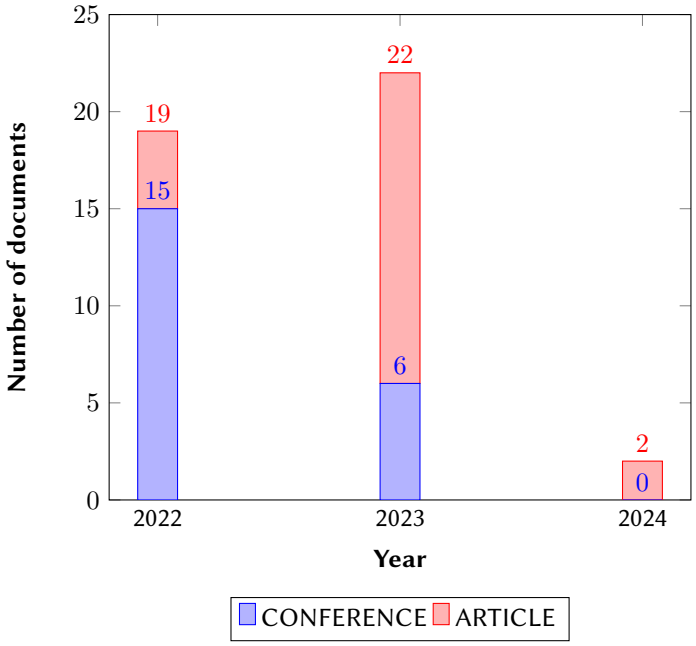

Figure 7: Number of CONFERENCE and ARTICLE document types by year.

3.2. RQ1: What advanced deep learning methods are used for text generation in the literature of 2022-2024?

Table 3 presents an overview of neural network architectures used for text generation, according to data from 2022-2024 studies.

Table 3: Neural network architectures for text generation.

Architecture	Description	Representatives	Articles
Traditional approaches			
RNN (Recurrent Neural Networks)	Recurrent neural networks used for processing sequential data.	–	[6, 9, 47, 10, 11, 29, 30]
LSTM (Long Short-Term Memory)	A variant of RNN that better remembers long-term dependencies.	–	[6, 10, 13, 14, 15, 29, 30, 33, 11, 41, 42]
GRU (Gated Recurrent Unit)	A simplified variant of LSTM with fewer parameters.	–	-
CNN (Convolutional Neural Networks)	Convolutional neural networks, often used for image processing.	YOLOv5	[6, 9, 38, 16]
Graph Neural Networks	Models that work with graph data structures.	GraphWriter, CGE-LW	[7, 9]
Innovative approaches			
Autoencoders	Networks used for learning efficient encodings of unlabeled data	AE, VAE, iVAE, clVAE+ MI, β0.4 VAE, SaVAE, LagVAE	[17, 15, 29]

Continued on next page

Table 3 – continued from previous page

Architecture	Description	Representatives	Articles
Transformer	Architecture that uses an attention mechanism for processing sequential data.	T5, CodeT5, TrICY, DETR	[7, 9, 18, 19, 31, 47, 8, 20, 32, 22, 27, 41, 39, 43, 34, 44, 19, 48]
BERT (Bidirectional Encoder Representations from Transformers)	A Transformer-based model trained on large amounts of unlabeled text.	PubmedBERT, BioLinkBERT, RoBERTa, XLM-RoBERTa	[8, 37, 13, 18, 19, 20, 26, 28, 30, 35, 32, 12, 9, 39, 40, 45]
GPT-2, GPT-3 (Generative Pre-trained Transformer)	Transformer-based models used for text generation.	OPT, Llama, CodeBERT	[6, 8, 10, 11, 13, 47, 15, 18, 21, 22, 12, 23, 24, 26, 19, 33, 25, 32, 37, 34, 45, 36, 39, 43, 44]
Attention-based models	Models that use an attention mechanism to improve the quality of generated text.	–	[47, 8, 20, 26, 43, 44]
Seq2Seq (Sequence-to-Sequence)	Architecture that uses an encoder and decoder to generate sequences.	S2ST, S2SL, S2SG, S2ST+, D+ Full, DSG	[39, 15, 42, 28, 31, 46, 43]
GAN (Generative Adversarial Networks)	Generative adversarial networks consisting of a generator and discriminator.	EGAN, TILGAN, DoubAN-Full, WRGAN, CatGAN, SeqGAN, DGSAN	[6, 29, 25]
Memory Networks	Models that use external memory for storing and accessing information.	DM-NLG (with memory), MemNNs, Mem2Seq, GLMP	[34, 9]
Diffusion Models	Models that use a diffusion process to generate text.	GENIE, NAT, iNAT, ELMER, MASS, ProphetNet, InsT, CMLM, LevT, BANG, ConstLeven	[41]
Prompt-based models	Models that use prompt-engineering fine-tuning to control text generation.	–	[23]

Table 4 presents a summary of text generation approaches based on the data from table 3.

Table 4
Approaches to text generation.

Category	Articles
Traditional approaches	[14, 38, 16]
Innovative approaches	[18, 19, 31, 8, 20, 32, 22, 27, 39, 43, 34, 44, 48, 37, 26, 28, 35, 12, 40, 45, 21, 23, 24, 25, 36, 46]
Combination of traditional and innovative approaches	[6, 9, 47, 10, 11, 29, 30, 13, 15, 33, 41, 42, 7]

Among the innovative approaches, the most popular are the use of models based on the Transformer architecture, in particular 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 are also gaining popularity.

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

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

Figure 8 shows that in 2022 and 2023, innovative approaches to text generation prevail, while traditional approaches and a combination of approaches are less common. In 2024, there are articles that use innovative and combined approaches in equal numbers, but the sample for this year is incomplete, since data were collected only for part of the year. In general, there is a trend towards an increase in the number of studies applying innovative approaches, such as models based on the Transformer architecture and attention mechanisms.

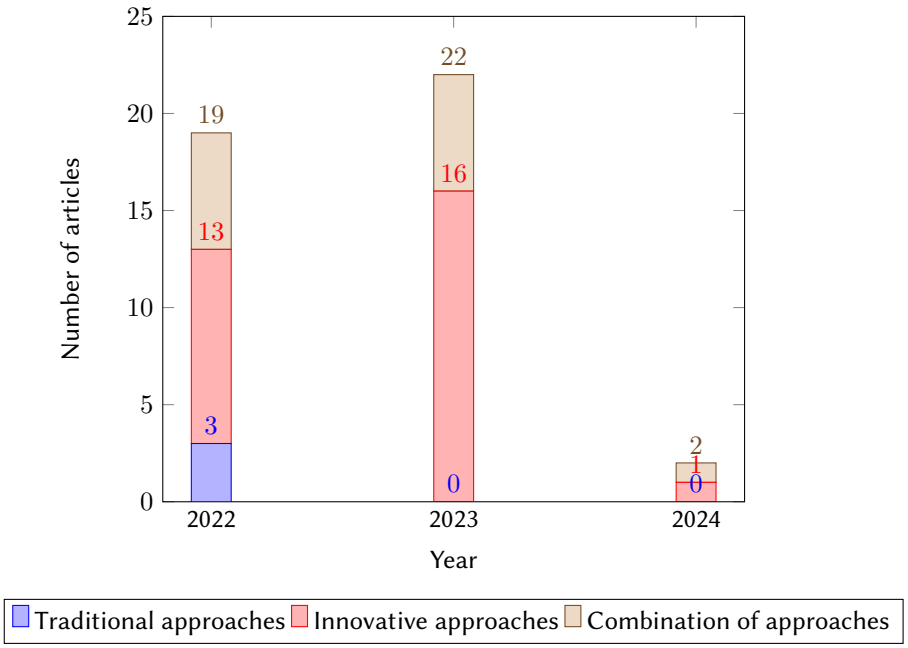

Figure 8: Distribution of articles by year according to categories of text generation approaches.

Comparing the obtained results with the data from the previous systematic review [2], the following conclusions can be drawn:

- Traditional approaches, such as RNN, LSTM, 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.
- Considerable attention is paid to models that use 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 transition from traditional approaches to more innovative and efficient models based on the Transformer architecture and attention mechanisms, which allows improving the quality of generated text and expanding the scope of application of these technologies.

Thus, comparing the results of the two reviews demonstrates that although traditional metrics, such as BLEU and ROUGE, remain widely used, new metrics appear in 2022-2024 that take into account various aspects of generated text quality. This indicates the active development of quality assessment methods and the search for more effective and informative approaches to evaluating text generation models.

3.3. RQ2: What new metrics for evaluating the effectiveness of text generation models are there in the literature of 2022-2024?

Table 5 presents an overview of quality metrics used to evaluate text generation. The metrics are divided into two categories: human-centred and machine-centred. Human-centred metrics include Human Evaluation and Turing Test, which involve evaluating the quality of generated text by human experts or testing a model's ability to generate text similar to that written by a human. Machine-centred metrics include a wide range of automatic metrics such as BLEU, ROUGE, METEOR, Perplexity, Distinct-n, BERTScore, and others. These metrics evaluate different aspects of generated text quality, such as similarity to the reference text, fluency, meaningfulness, lexical and syntactic diversity, etc.

Table 5: Main quality metrics for evaluating text generation.

Quality metric	Description	Representatives	Articles
Human-centred metrics			
Human Evaluation	Evaluation of the quality of generated text by human experts.	–	[9, 10, 11, 25, 30, 32, 33, 36, 31, 37]
Turing Test	A test of a model's ability to generate text indistinguishable from that written by a human.	–	[33]
Machine-centred metrics			
BLEU	A metric that evaluates the quality of generated text by comparing it with reference text.	BLEU-1, BLEU-2, BLEU-3, BLEU-4, BLEU-5	[7, 8, 18, 9, 10, 13, 15, 18, 19, 23, 27, 29, 31, 32, 33, 34, 36, 37, 41, 42, 43, 44, 45, 46, 47]
ROUGE	A metric that evaluates the quality of automatic text summarization.	ROUGE-1, ROUGE-2, ROUGE-3, ROUGE-L	[7, 8, 9, 10, 13, 18, 19, 23, 27, 28, 33, 34, 35, 36, 41, 42, 43, 44, 45, 46, 48]
METEOR	A metric that evaluates the quality of machine translation.	–	[18, 27, 32, 34, 36, 42, 43, 44, 46, 48]
BERTScore	A metric that evaluates the quality of generated text using a pre-trained BERT model.	–	[8, 13, 18, 19, 26, 34, 32, 37]
CIDEr	A metric that evaluates the quality of automatic image captioning by comparing machine-generated captions with sets of reference captions.	–	[14, 18, 23, 36, 37, 41, 42, 46]
Perplexity	A metric that evaluates the quality of a language model.	–	[8, 9, 15, 17, 26, 29, 36, 39]
F1-score	A metric that evaluates the quality of classification, particularly in binary classification tasks.	–	[13, 20, 21, 26, 34, 40]

Continued on next page

Table 5 – continued from previous page

Quality metric	Description	Representatives	Articles
CHRF++	A metric that evaluates the quality of machine translation based on character and n-gram matches.	–	[7, 32, 37, 48]
Distinct-n	A metric that evaluates the diversity of generated text.	Dist-1, Dist-2, Dist-3, Dist-4	[8, 9, 15]

Table 6 provides an overview of the quality evaluation metrics applied in the articles. Most studies use machine-centred metrics for automatic evaluation of generated text quality. A significantly smaller number of studies apply human-centred metrics, which may be due to the labour-intensive and subjective nature of human quality assessment. However, the use of human-centred metrics remains important for obtaining a more complete and reliable evaluation of text generation quality. Some studies do not apply any quality metrics, which may be related to the focus on other aspects of text generation, such as efficiency or speed of model operation.

Table 6
Overview of quality evaluation metrics applied in the articles.

Quality metrics	Articles
Machine-centred	[7, 8, 11, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 26, 27, 28, 29, 34, 35, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48]
Human-centred	[11, 30]
Both	[9, 10, 25, 32, 33, 36, 31, 37]
Not applied	[24, 12, 6]

The use of diverse quality metrics is important for a comprehensive evaluation of the effectiveness of models and approaches to text generation. Combining machine-centred and human-centred metrics allows obtaining more reliable and valid evaluation results.

The diagram in figure 9 shows that the most frequently used quality metrics are BLEU (55.8% of articles) and ROUGE (48.8% of articles). Human Evaluation is also quite common – it is applied in 23.3% of articles. Other metrics, such as Perplexity, METEOR, BERTScore, and Distinct-n, are used less frequently but still have a significant share of mentions in articles. The least common metrics are the Turing Test, Fluency, Coherence, Diversity, N-gram Overlap, and Embedding Similarity, each of which is mentioned in only one article (2.3%).

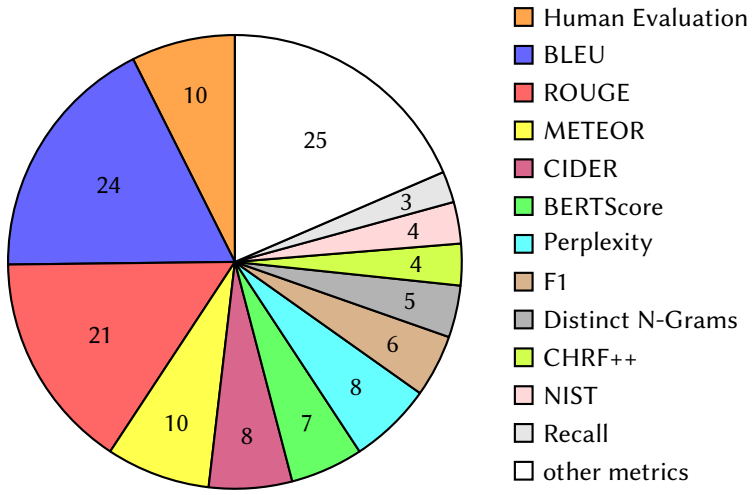

Figure 9: Distribution of quality metrics by the number of articles in which they are mentioned.

Automatic quality metrics, such as BLEU and ROUGE, are the most widely used for evaluating the effectiveness of text generation models, while human quality evaluation is used less frequently but remains an important component for obtaining a more complete and reliable assessment of generated text quality.

Comparing the obtained results with the data from the previous systematic review [2], the following observations can be made:

- BLEU and ROUGE remain the most popular metrics for evaluating the quality of generated text both in 2015-2021 and in 2022-2024.
- Human Evaluation is still widely used to obtain a more complete and reliable evaluation of text generation quality, despite the labour-intensiveness and subjectivity of this approach.
- In 2022-2024, new metrics appeared, such as BERTScore, Fluency, Coherence, Diversity, N-gram Overlap, and Embedding Similarity, which were not presented in the previous review. This indicates the active development of methods for evaluating the quality of generated text and the search for more effective and informative metrics.
- Perplexity has gained more popularity in 2022-2024 compared to the previous period, which may be related to its effectiveness in assessing the quality of language models.
- The METEOR metric, which evaluates the quality of machine translation, is also used more frequently in 2022-2024, which may indicate a growing interest in applying text generation to machine translation tasks.
- In general, there is a trend towards combining different types of metrics (machine-centred and human-centred) to obtain more reliable and valid results when evaluating the effectiveness of text generation models.

Thus, comparing the results of the two reviews demonstrates that while traditional metrics, such as BLEU and ROUGE, remain widely used, new metrics appear in 2022-2024 that take into account various aspects of generated text quality. This indicates the active development of quality assessment methods and the search for more effective and informative approaches to evaluating text generation models.

3.4. RQ3: What text generation datasets are described in the literature of 2022-2024?

Table 7 presents the datasets mentioned in the reviewed articles, sorted in descending order by the number of mentions and alphabetically in case of an equal number of mentions. The E2E dataset is mentioned most frequently – in 7 articles, followed by XSum (4 articles), CNN/DailyMail (4 articles), CommonGen (4 articles), ToTTo (4 articles), WebNLG (4 articles), WikiBio (3 articles), DDI (2 articles), NIST (2 articles), PubMed (2 articles), Quora (2 articles), RocStories (2 articles), Snips (2 articles), SST-2 (2 articles), WMT'14 English-German (2 articles), WMT'16 Romanian-English (2 articles), and Yelp (2 articles). Other datasets are mentioned once, sorted by appearance in the review.

Table 7: Datasets mentioned in the reviewed articles.

Dataset name	Articles
E2E	[19, 23, 30, 31, 34, 36, 44]
CNN/DailyMail (CNN/DM)	[9, 23, 41, 45]
Totto	[18, 31, 43, 46]
CommonGen	[9, 18, 36, 41]
WebNLG	[7, 31, 37, 44]
XSum	[9, 18, 23, 41]
WikiBio	[34, 18, 31]
Abstract Generation Dataset (AGENDA)	[7, 9]
DDI	[9, 12]

Continued on next page

Table 7 – continued from previous page

Dataset name	Articles
NIST	[9, 27]
PubMed	[12, 23]
Quora	[47, 9]
ROCStories	[36, 9]
Snips	[19, 39]
SST-2	[21, 45]
WMT'14 English-German	[18, 27]
WMT'16 Romanian-English	[18, 27]
Yelp	[17, 40]
Baidu Tieba	[9]
PersonaChat	[9]
Gigawords	[9]
Yahoo! Answers	[9]
NLPCC	[9]
Tencent	[9]
SQuAD	[9]
ComVE	[9]
αNLG-ART	[9]
EntDesc	[9]
VisualStory	[9]
PaperWriting	[9]
Reddit-10M	[9]
EMNLP dialog	[9]
ICLR dialog	[9]
NarrativeQA	[9]
Wizard of Wikipedia (WoW)	[9]
MS-MARCO	[9]
ELI5	[9]
ChangeMyView	[9]
Amazon books	[9]
Foursquare	[9]
Scratch online community comments	[11]
BC5-Chemical	[12]
BC5-Disease	[12]
NCBI-Disease	[12]
BC2GM	[12]
JNLPBA	[12]
EBM PICO	[12]
ChemProt	[12]
GAD	[12]
BIOSSES	[12]
HoC	[12]
PubMedQA	[12]
BioASQ	[12]
Logic2Text	[13]
Concadia	[14]
REDIAL	[15]

Continued on next page

Table 7 – continued from previous page

Dataset name	Articles
Custom dataset for Bangla word sign language	[16]
Synthetic dataset	[17]
Penn Treebank	[17]
IWSLT'14 De-En	[18]
WMT16 English-German	[45]
WMT17 English-German	[36]
WMT20	[37]
WMT21	[37]
WMT'14 German-English	[27]
Multi-News	[18]
Java	[18]
Python	[18]
English ATIS	[19]
ViGGO	[19]
TREC	[19]
Korean Weather	[19]
Rest	[19]
KLUE-TC	[19]
C4	[20]
M2D2	[20]
Political Slant	[20]
Layoff	[21]
MC	[21]
M&A	[21]
Flood	[21]
Wildfire	[21]
Boston Bombings	[21]
Bohol Earthquake	[21]
West Texas Explosion	[21]
Dublin	[21]
New York City	[21]
WSC	[22]
CBT-CN	[22]
CBT-NE	[22]
Wikihow	[23]
SAMSum	[23]
DART	[23]
Custom dataset composed of tweets labeled with emotions	[25]
AFQMC	[26]
CHIP-STS	[26]
QQP	[26]
MRPC	[26]
ParaNMT-small	[27]
NIST Chinese-English	[27]
GTZAN	[28]
Minions	[29]
Japanimation	[29]

Continued on next page

Table 7 – continued from previous page

Dataset name	Articles
WikiArt	[29]
Nottingham	[29]
Lakh MIDI	[29]
TheoryTab	[29]
Poem-5	[29]
Poem-7	[29]
Synthetic date generation dataset	[30]
LDC2020T02 (AMR 3.0 release)	[32]
One Million Urdu News Dataset	[33]
Australian Broadcasting Corporation (ABC) news dataset	[33]
DailyMed drug labels	[35]
COCO Image Captioning	[37]
German and French commercial datasets	[39]
MASSIVE	[39]
Gold-PMB	[42]
Silver-PMB	[42]
numericNLG	[43]
Custom dataset related to text messaging applications	[44]
TweetEval	[45]
AGnews	[45]
QNLI	[45]
IMDB	[45]
CC-News	[45]
WITA	[46]
XWIKIREF	[48]

The analysed studies use a wide range of datasets covering various domains and types of texts, from user reviews and news articles to medical and technical texts. This indicates the active development and application of text generation methods in diverse fields.

Table 8 presents the data types used in the reviewed articles, sorted in descending order by the number of mentions. Datasets containing sentences are used most often – they are mentioned in 26 articles. In 5 articles, the data type is not explicitly specified. Other data types, such as paragraphs (18 articles), documents (11 articles), question-answer (10 articles), descriptive tables (9 articles), translations (7 articles), stories (4 articles), images (4 articles), and others, are less common.

The prevalence of datasets with sentences may be due to the fact that many text generation tasks, such as machine translation, paraphrasing, question answering, etc., often work at the sentence level. At the same time, the presence of various data types, including paragraphs, documents, images, music, and others, indicates that text generation methods can be applied to a wide range of tasks and domains.

Table 9 presents the data annotation types used in the reviewed articles, sorted in descending order by the number of mentions. Labeled datasets are used most often – they are mentioned in 22 articles. In 20 articles, the annotation type is not explicitly specified. In 5 articles, unlabeled data are used. In 4 articles, both labeled and unlabeled data are used.

The prevalence of labeled datasets may be due to the fact that many text generation tasks, especially those that use controlled approaches or require compliance with certain templates or structures, require labeled data for training models. Annotation can include elements such as parts of speech, syntactic structures, semantic roles, tags for controlled generation, etc.

At the same time, the presence of studies that use unlabeled data or a combination of labeled and unlabeled data indicates the active development of unsupervised and semi-supervised learning methods in the field of text generation. These approaches allow using large volumes of unlabeled text data for

Table 8
Data types used in the reviewed articles.

Data type	Articles
Sentence	[7, 9, 11, 12, 14, 15, 17, 18, 19, 20, 22, 23, 25, 26, 29, 30, 31, 32, 33, 36, 39, 40, 41, 45, 46, 48]
Paragraph	[7, 9, 12, 15, 17, 18, 19, 20, 23, 29, 37, 14, 39, 40, 41, 45, 46, 48]
Document	[9, 12, 18, 19, 20, 29, 35, 40, 41, 42, 45]
Question-answer	[7, 9, 12, 17, 18, 19, 22, 26, 45, 47]
Descriptive tables	[13, 21, 30, 31, 33, 34, 36, 43, 44]
Translations	[18, 27, 31, 33, 36, 37, 45]
Stories	[9, 31, 33, 36]
Images	[14, 29, 37, 38]
Audio files	[29, 28]
Video clips	[16]
Computer programs	[18]
Not specified	[6, 8, 10, 24]

Table 9
Data annotation types used in the reviewed articles.

Annotation type	Articles
Labeled data	[12, 13, 14, 16, 17, 18, 21, 27, 28, 31, 33, 34, 35, 39, 40, 42, 43, 44, 46, 48, 47, 25]
Unlabeled data	[11, 12, 39, 40, 47]
Not specified	[6, 7, 8, 9, 10, 15, 19, 20, 22, 24, 23, 26, 29, 30, 32, 36, 37, 38, 41, 45]

Table 10
Data quality used in the reviewed articles.

Data quality	Articles
Preprocessed	[13, 14, 16, 17, 18, 44, 47, 48, 39, 34, 31, 42]
Raw	[7, 11, 28, 33, 35, 37, 39, 34, 31, 42]
Not specified	[6, 8, 9, 10, 12, 15, 20, 19, 21, 22, 24, 25, 23, 26, 27, 29, 30, 32, 36, 38, 40, 41, 43, 45, 46]

pre-training models and improving their ability to generate coherent and meaningful text.

Table 10 presents the data quality used in the reviewed articles, sorted in descending order by the number of mentions. In 28 articles, the data quality is not explicitly specified. In 12 articles, preprocessed data are used, while in 10 articles – raw data. In 4 articles, both preprocessed and raw data are used.

Preprocessed data usually go through the stages of cleaning, normalization, tokenization, and sometimes additional annotation before being used in model training. This improves the quality and consistency of the data, as well as facilitates the learning process. Examples of preprocessed data can be datasets obtained from existing corpora or databases that have already undergone some processing.

Raw data, on the other hand, are data obtained directly from real sources, such as web pages, social networks, unprocessed texts, etc. They can contain noise, incorrect formatting, errors, and other artifacts. Using raw data can be useful for training models that need to be robust to real conditions and able to process unstructured data.

The lack of information about data quality in a significant part of the analyzed articles may indicate that the authors do not pay enough attention to this aspect or consider it less important for the research. At the same time, data quality is a critical factor that affects the efficiency and generalizability of text generation models, so it is worth paying more attention to the description and analysis of the quality of the data used in future research.

Comparing the results of the 2022-2024 review with the previous review [2], the following conclusions can be drawn:

- In 2022-2024, new datasets appeared, such as XWIKIREF, DailyMed, numericNLG, WITA, DISTToTTo, which were not presented in the previous review. This indicates the active development of resources for research and application of text generation methods.

- The datasets E2E, WikiBio, ToTTo, CommonGen, CNN/DailyMail, and XSum remain popular and widely used in research both in 2015-2021 and in 2022-2024.
- There is a trend towards the use of more diverse data types, such as descriptive tables, images, music, translations, question-answer, video clips, and computer programs, in addition to traditional types such as sentences, paragraphs, and documents.
- Labeled data remain the most widely used, but there is a growing interest in using unlabeled data and a combination of labeled and unlabeled data for training text generation models.
- Although data quality is a critical factor affecting model efficiency, a significant part of the 2022-2024 research does not cover this aspect, which may indicate the need to pay more attention to the description and analysis of the quality of the data used in future research.

Thus, comparing the results of the two reviews demonstrates that text generation datasets continue to actively develop, covering new domains and data types. At the same time, some popular datasets remain relevant and widely used in research. There is a trend towards the use of more diverse data types and a growing interest in unlabeled data and combined approaches. However, the description of data quality still requires more attention in future research to ensure the reliability and reproducibility of the results.

3.5. RQ4: What new text generation applications are described in the literature of 2022-2024?

Table 11 shows the text generation applications found in the analyzed articles, sorted in descending order by the number of references. The most common applications are text summarization (8 articles), machine translation (8 articles), table-to-text generation (5 articles), paraphrasing and data augmentation (4 articles each). Other applications, such as controllable text generation, image-based text generation, text generation from knowledge graphs, etc., are mentioned in a smaller number of articles.

Table 11
Text generation applications.

Application	Articles
Table-to-Text Generation	[43, 13, 36, 30, 34]
Text Generation from Knowledge Graphs	[7, 9]
Controllable Text Generation	[8, 19, 30]
Medical Text Generation	[12, 35]
Paraphrasing	[9, 47, 39, 27]
Image-based Text Generation	[29, 14, 37]
Text Summarization	[9, 18, 47, 23, 41, 37, 45, 48]
Emotional Text Generation	[11, 25]
Question Answering	[15, 9]
Music Text Generation	[28, 29]
Machine Translation	[9, 16, 17, 18, 27, 47, 36, 37]
Data Augmentation	[8, 21, 42, 40]
Script Generation	[29, 9]
News Headline Generation	[33]
Technical Documentation Generation	[10]
Cybersecurity	[45]
Encyclopedic Text Generation	[48]
Data-to-Text Generation	[8, 31, 34, 44, 46]
Sign Language to Text Translation	[16]

Figure 10 visualizes the text generation applications listed in table 11 as a diagram. The diagram clearly shows the prevalence of table-to-text generation, text generation from knowledge graphs, controllable text generation, and medical text generation applications compared to other areas.

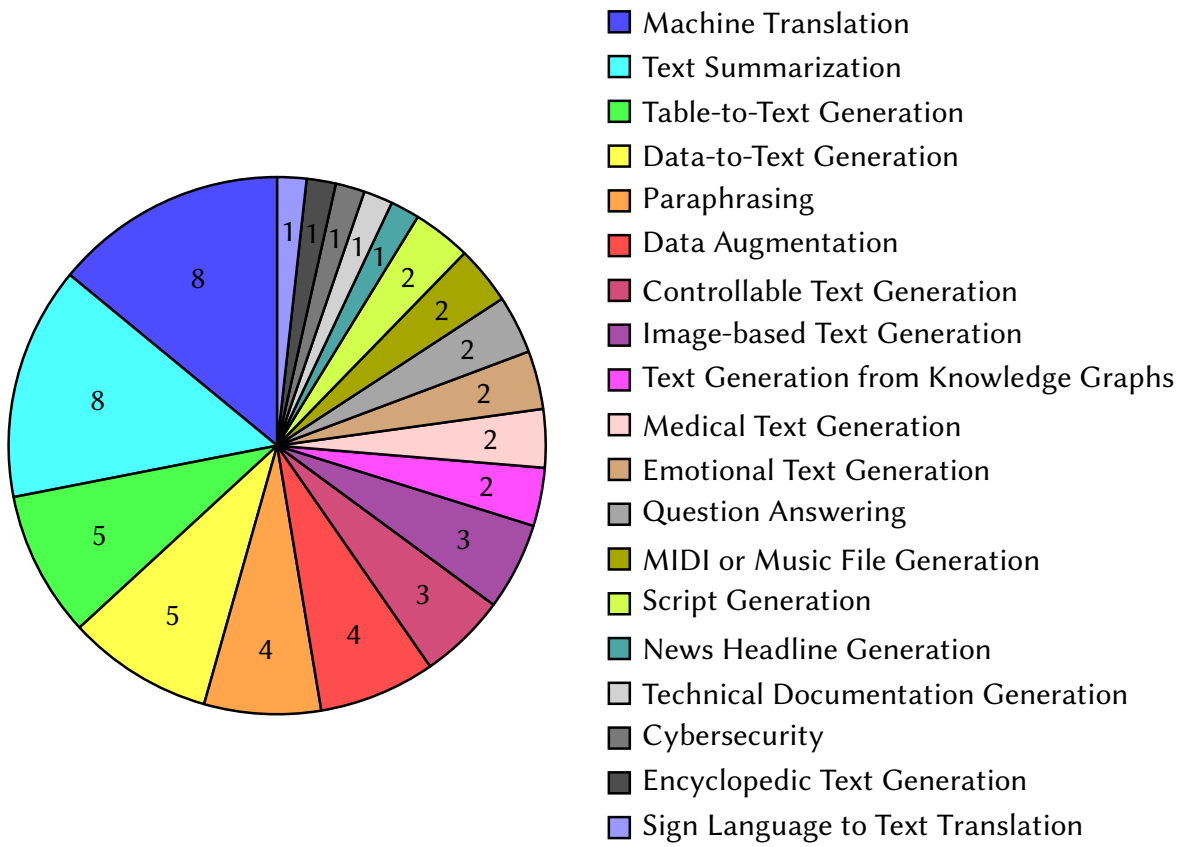

Figure 10: Text generation applications in the analyzed articles of 2022-2024.

The analysis of text generation applications demonstrates a wide range of possibilities for using this technology in various fields, from processing structured data to creating emotionally colored texts and translating sign language into text. The development of new methods and neural network architectures opens up new prospects for further expanding the areas of text generation application.

Comparing the results of the 2022-2024 review with the previous review [2], the following observations can be made:

- Machine Translation and Text Summarization have gained more popularity in 2022-2024 compared to the previous period. However, in 2022-2024, text generation from tables and structured data was added to them, which may indicate a growing interest in processing structured information using text generation methods.
- Controllable Text Generation has also become more common, indicating a growing interest in methods that allow controlling the text generation process and obtaining more relevant and high-quality results.
- Medical Text Generation has emerged as a new area of text generation application in 2022-2024, which may be related to the active development of methods for processing medical data and the need to automate the creation of medical documentation.
- New applications have emerged, such as Emotional Text Generation, Encyclopedic Text Generation, Technical Documentation Generation, and Sign Language to Text Translation, indicating an expansion of the areas of text generation use.
- Paraphrasing and Data Augmentation remain relevant text generation applications both in 2015-2021 and in 2022-2024.
- Some applications that were popular in the previous review, such as poetry generation, dialogue systems, text classification, topic modeling, do not appear among the most frequently mentioned

in the new review. This may be related to a change in research focus and the emergence of new promising directions.
- Overall, there is a trend towards increasing diversity of text generation applications compared to the previous one, which indicates the active development of this area of research and the expansion of the possibilities of using generative models to solve applied problems in various subject areas.

Thus, comparing the results of the two reviews demonstrates that the field of text generation application continues to actively expand, covering new areas and directions. The popularity of such applications as text generation from tables and knowledge graphs, controllable text generation, and medical text generation indicates a growing interest in methods that allow efficiently processing structured data and obtaining more relevant and high-quality results. At the same time, traditional applications, such as paraphrasing, text summarization, and machine translation, remain relevant and widely used in research.

3.6. RQ5: What natural languages are used for text generation in the literature of 2022-2024?

Table 12 presents an extended annual summary of the languages used for text generation in 2022-2024. English is the most widely used language, with 38 articles covering all three years. Various neural network architectures are used for generating English texts, including Transformer, BERT, GPT-2, GPT-3, RNN, LSTM, CNN, GAN, and Seq2Seq.

German is represented in 5 articles using GAN architectures (Conditional GAN, StyleGAN, DCGAN). Chinese is represented in 4 articles using Graph Neural Networks and B2T architecture. Bengali is represented in 2 articles (one in 2022 and one in 2023) dedicated to recognition using CNN and YOLO. Romanian is represented in 2 articles (one in 2022 and one in 2023) using DCGAN and BART architectures. French, Urdu, Shakespearean English, and Korean are each mentioned in one article, using various architectures such as Conditional GAN, StyleGAN, DCGAN, and GPT-2.

In 2023, a study by Taunk et al. [48] appears dedicated to generating texts in several Indian languages (Hindi, Malayalam, Marathi, Oriya, Punjabi, and Tamil) using HipoRank, mBART, and mT5 architectures.

Table 12: Extended annual summary of text generation languages.

Language	2022	2023	2024	Total	Architectures
English	17 [6, 14, 17, 18, 23, 28, 22, 25, 16, 42, 47, 31, 36, 37, 39, 30, 43]	19 [7, 8, 11, 12, 15, 19, 41, 20, 21, 26, 27, 32, 33, 34, 35, 40, 45, 46, 48]	2 [13, 44]	38	Transformer, BERT, GPT-2, GPT-3, RNN, LSTM, CNN, GAN, Seq2Seq
German	3 [39, 37, 18]	2 [45, 27]	–	5	Conditional GAN, StyleGAN, DCGAN
Chinese	1 [37]	3 [27, 29, 10]	–	4	Graph Neural Networks, B2T
French	1 [39]	–	–	1	Conditional GAN, StyleGAN, DCGAN
Bengali	1 [16]	1 [48]	–	2	CNN, YOLO, mBART
Urdu	–	1 [33]	–	1	GPT-2

Continued on next page

Table 12 – continued from previous page

Language	2022	2023	2024	Total	Architectures
Hindi, Malayalam, Marathi, Oriya, Punjabi, Tamil	–	1 [48]	–	1	HipoRank, mBART, mT5
Shakespearean English	1 [29]	–	–	1	Modified DCGAN
Romanian	1 [18]	1 [27]	–	2	DCGAN, BART
Korean	–	1 [19]	–	1	Modified DCGAN

Comparing the results of the 2022-2024 review with the previous review [2], the following observations can be made:

- English remains the most widely used language for text generation in both 2015-2021 and 2022-2024. However, there is a trend towards an increase in the number of studies dedicated to other languages, especially low-resource languages.
- In 2022-2024, studies appeared dedicated to generating texts in languages that were not represented in the previous review, such as Urdu, Hindi, Malayalam, Marathi, Oriya, Punjabi, and Tamil. This indicates a growing interest in developing text generation models for diverse languages.
- The study [48] demonstrates the possibility of generating texts in several Indian languages simultaneously using modern architectures such as HipoRank, mBART, and mT5, which was not presented in the previous review.
- Both traditional architectures (RNN, LSTM, CNN) and more modern approaches, such as Transformer, BERT, GPT-2, GPT-3, GAN, and Graph Neural Networks, are used for generating texts in different languages.
- Overall, there is a trend towards expanding the range of languages for which text generation models are being developed and using more diverse neural network architectures for this task.

Thus, comparing the results of the two reviews demonstrates that although English remains the dominant language in text generation research, there is a growing interest in developing models for other languages, especially low-resource languages. The emergence of studies dedicated to generating texts in languages such as Urdu, Hindi, Malayalam, Marathi, Oriya, Punjabi, and Tamil indicates an expansion of the possibilities for applying text generation to diverse languages. Furthermore, the use of modern neural network architectures such as Transformer, BERT, GPT-2, GPT-3, GAN, and Graph Neural Networks allows improving the quality and efficiency of text generation for various languages.

Comparing the language distributions in the old and new reviews with the distribution of languages by the number of models on Hugging Face [52], the following observations can be made:

- English dominates in all three distributions. In the old and new reviews, it is the most widely used for text generation, and on Hugging Face, the largest number of models (51738) are available for it. This indicates significant attention from researchers and developers to the English language and the availability of a large number of resources for it.
- Chinese ranks second in the number of models on Hugging Face (4546) and is mentioned in several articles in the new review. This points to a growing interest in Chinese text generation and the development of relevant resources.
- Languages such as French, Spanish, Russian, and German have a significant number of models on Hugging Face (from 2326 to 4049) but are less frequently mentioned in the reviews. This may indicate that despite the availability of resources for these languages, text generation research for them is not as widely represented in the literature.

- Low-resource languages such as Bengali, Urdu, Arabic, and Hindi are mentioned in the new review, indicating a growing interest in developing text generation models for these languages. However, the number of available models on Hugging Face for these languages is significantly lower compared to English (from 670 to 1674).
- Hugging Face represents significantly more languages (over 200) than are mentioned in the reviews. This indicates that text generation research covers only a portion of the languages for which models and resources are available.
- Some languages, such as Japanese, Korean, Indonesian, and Arabic, have a significant number of models on Hugging Face (from 1674 to 2920) but are rarely mentioned in the reviews. This may indicate the potential for further research on text generation in these languages.

Comparing the language distributions shows that despite the dominance of English in research and available resources, there is a growing interest in text generation in other languages, especially low-resource ones. However, the number of available models and resources for these languages is still significantly lower compared to English. Furthermore, the presence of a large number of models for some languages on Hugging Face that are rarely mentioned in the reviews indicates the potential for further research and development in this field.

4. Conclusions

The paper presented the results of a systematic review of the application of artificial neural networks for generating textual content in 2022-2024 and compared them with the results of the previous review [2] for 2015-2021. The main conclusions can be summarized as follows:

1. There is a trend towards an increase in the number of articles in scientific journals compared to conference proceedings, which may indicate a more thorough coverage of text generation issues in journals.
2. Among the advanced deep learning methods for text generation, the most popular are models based on the Transformer architecture, such as GPT-2, GPT-3, BERT, and their variations. Approaches using attention mechanisms and controlled text generation are also gaining popularity. Overall, there is a shift from traditional approaches to more innovative and efficient models.
3. Among the metrics for evaluating the effectiveness of text generation models, BLEU and ROUGE are the most widely used, along with human evaluation. In 2022-2024, new metrics such as BERTScore, Fluency, Coherence, Diversity, N-gram Overlap, and Embedding Similarity appeared, indicating active development of methods for assessing the quality of generated text.
4. Datasets for text generation continue to actively develop, covering new domains and types of data. There is a trend towards using more diverse types of data (tables with descriptions, images, music, translations, etc.) and a growing interest in unlabeled data and combined approaches.
5. The field of text generation applications continues to actively expand, covering new areas and directions. The popularity of applications such as text generation from tables and knowledge graphs, controlled text generation, and medical text generation indicates a growing interest in methods that allow efficient processing of structured data and obtaining more relevant and high-quality results.
6. Although English remains the dominant language in text generation research, there is a growing interest in developing models for other languages, especially low-resource languages. The use of modern neural network architectures allows improving the quality and efficiency of text generation for various languages.

The results of this review demonstrate the active development of the field of text generation in 2022-2024, characterized by the emergence of new approaches, metrics, datasets, and the expansion of application areas.

Despite significant progress in the development of text generation technologies, questions remain open regarding the assessment of the quality of generated text, the adaptation of models to different subject domains and languages, and the ethical aspects of using these technologies. Further research can be aimed at solving these problems and developing more effective, universal, and safe text generation models.

Declaration on Generative AI: During the preparation of this work, the authors used Claude 3 Opus in order to: Text Translation, Abstract drafting, Formatting assistance. After using this tool, the authors reviewed and edited the content as needed and takes full responsibility for the publication's content.

References

[1] T. Ganegedara, Natural Language Processing with TensorFlow: Teach language to machines using Python's deep learning library, Packt Publishing, Birmingham – Mumbai, 2018. URL: https://tinyurl.com/3xps3c5u.

[2] 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.

[3] OpenAI, Introducing ChatGPT, 2022. URL: https://openai.com/blog/chatgpt.

[4] 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.

[5] 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.

[6] A. Bas, M. O. Topal, Ç. 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, pp. 1–4. doi:10.1109/ICCA56443.2022.10039545.

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

[8] 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) 64. doi:10.1145/3617680.

[9] 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) 227. doi:10.1145/3512467.

[10] 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.

[11] 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.

[12] 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) btad557. doi:10.1093/bioinformatics/btad557.

[13] I. Alonso, E. Agirre, Automatic logical forms improve fidelity in table-to-text generation, Expert Systems with Applications 238 (2024). doi:10.1016/j.eswa.2023.121869.

[14] 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. doi:10.18653/v1/2022.emnlp-main.308.

[15] 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. URL: http://www.jatit.org/volumes/Vol101No7/30Vol101No7.pdf.

[16] 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, pp. 1–6. doi:10.1109/IPAS55744.2022.10052928.

[17] 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. URL: https://proceedings.mlr.press/v180/teng22a.html.

[18] C. An, J. Feng, K. Lv, L. Kong, X. Qiu, X. Huang, CONT: contrastive neural text generation, in: Proceedings of the 36th International Conference on Neural Information Processing Systems, NIPS '22, Curran Associates Inc., Red Hook, NY, USA, 2022, p. 160. URL: https://dl.acm.org/doi/10.5555/3600270.3600430.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] M. Ghazvininejad, V. Karpukhin, V. Gor, A. Celikyilmaz, Discourse-Aware Soft Prompting for Text Generation, 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. 4570–4589. doi:10.18653/v1/2022.emnlp-main.303.

[24] J. J. Koplin, Dual-use implications of AI text generation, Ethics and Information Technology 25 (2023) 32. doi:10.1007/s10676-023-09703-z.

[25] A. Pautrat-Lertora, R. Perez-Lozano, W. Ugarte, EGAN: Generatives Adversarial Networks for Text Generation with Sentiments, in: F. Coenen, A. Fred, J. Filipe (Eds.), International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, IC3K - Proceedings, volume 1, Science and Technology Publications, Lda, 2022, pp. 249–256. doi:10.5220/0011548100003335.

[26] 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.

[27] 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. doi:10.18653/v1/2023.acl-long.788.

[28] X. Chu, Feature extraction and intelligent text generation of digital music, Computational Intelligence and Neuroscience 2022 (2022). doi:10.1155/2022/7952259.

[29] 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.

[30] 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.

[31] 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. doi:10.18653/v1/2022.acl-long.531.

[32] 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, pp. 727–736. doi:10.18653/v1/2023.eacl-main.51.

[33] 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.

[34] E. Seifossadat, H. Sameti, Improving semantic coverage of data-to-text generation model using dynamic memory networks, Natural Language Engineering 30 (2024) 454–479. doi:10.1017/S1351324923000207.

[35] 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.

[36] 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. doi:10.18653/v1/2022.naacl-main.57.

[37] 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: Y. Goldberg, Z. Kozareva, Y. Zhang (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2022, Association for Computational Linguistics (ACL), 2022, pp. 6588–6603. doi:10.18653/v1/2022.findings-emnlp.489.

[38] 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. URL: http://www.jatit.org/volumes/Vol100No2/18Vol100No2.pdf.

[39] 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. doi:10.18653/v1/2022.emnlp-industry.26.

[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. doi:10.18653/v1/2023.acl-long.74.

[41] 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: Proceedings of the 40th International Conference on Machine Learning, ICML'23, JMLR.org, 2023. URL: https://dl.acm.org/doi/abs/10.5555/3618408.3619275.

[42] M. S. Amin, A. Mazzei, L. Anselma, Towards Data Augmentation for DRS-to-Text Generation, CEUR Workshop Proceedings 3287 (2022) 141–152. URL: https://ceur-ws.org/Vol-3287/paper14.pdf.

[43] 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. doi:10.18653/v1/2022.emnlp-main.562.

[44] 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.

[45] 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. URL: https://www.usenix.org/system/files/usenixsecurity23-si.pdf.

[46] 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.

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

[48] 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.

[49] Introducing the next generation of Claude, 2024. URL: https://www.anthropic.com/news/claude-3-family.

[50] awesomegpts.ai, Scholar GPT, 2024. URL: https://chatgpt.com/g/g-kZ0eYXlJe-scholar-gpt?oai-dm=1.

[51] A. V. Slobodianiuk, Ohliad statei [Papers' review], 2024. URL: https://docs.google.com/spreadsheets/d/e/2PACX-1vR6ZUaeeBjVgVl-do6QXm-Pua-HdztOxjC4DUqunrSDZ_-YSRz-Ng9xktYH9b0LDT502SiVy3YePx9F/pubhtml.

[52] Hugging Face, Languages, 2024. URL: https://huggingface.co/languages.

A. Review map for an article

1. Bibliographic reference
2. Document type: journal article or conference paper
3. Title
4. Year of publication
5. Countries represented by the authors
6. Purpose of the article
7. Neural network architectures used
8. Quality metrics used
9. Characteristics of the datasets used
    - name
    - data type: sentence, paragraph, document, question-answer, not specified
    - size
    - format: CSV, JSON, XML, files, not specified
    - labeling type: labeled data, unlabeled data
    - data quality: raw, pre-processed
    - accessibility: publicly available, private, not specified
    - link
10. Text generation task solved (what the neural network was used for)
11. Language of text generation