Journal of SUSTAINABILITY OUTREACH

Vol 1 (1): Pages 30-36 (2025)
DOI: 10.37357/1068/JSO/5.1.04
ISSN: 2435-7243

Check for updates

Foundation models for chemical and biological informatics

A vertically aligned language paradigm

Md Zahir Ahmed
Department of Computer Science, College of Engineering, Technology and Aeronautics (CETA), Southern New Hampshire University, Hooksett, USA

Article
Open Access
Published

Keywords
- Foundation models
- Large language models (LLMs)
- Vertically aligned language paradigm
- Domain-specific NLP
- Chemical informatics

ABSTRACT

This paper introduces a novel vertically aligned language paradigm embodied in a suite of large language models (LLMs) designed for chemical and biological informatics. Addressing the limitations of general-purpose LLMs in specialized scientific domains, this research details the creation and training of models tailored to the intricacies of biopharmaceutical and chemical literature. The methodology encompasses a rigorous data curation process, advanced model architectures, and specialized training techniques to achieve state-of-the-art performance on domain-specific benchmarks. A key contribution is the demonstration of enhanced performance in tasks such as pharmaceutical knowledge assessment and biomedical text translation, even with models of smaller parameter sizes. The study provides a comprehensive analysis of the models' capabilities, emphasizing their potential to accelerate innovation in drug discovery, chemical synthesis, and related fields. Ethical considerations concerning data privacy, model reliability, and equitable access are also addressed, promoting responsible development and deployment of LLMs in the life sciences.

Received: October 13, 2025; Revised: November 27, 2025; Accepted: November 28, 2025; Published: December 01, 2025

1. Introduction

The advent of extensive neural models has profoundly reshaped the framework of natural language understanding and generation [1]. Recent state-of-the-art advancements have alleviated reliance on intricate manual feature crafting, simplifying the development of intelligent linguistic systems. Such models possess the capability to interpret and produce contextualized text with minimal human input [2].

Nevertheless, limitations persist when adapting these general-purpose systems to domain-focused sectors, notably in biopharmaceutical contexts [1]. The majority of top-tier models are built with a broad scope, primarily in English, and lack the specificity required to perform effectively within technical disciplines.

To bridge this limitation, PharmaGPT was conceptualized as a multilingual framework of LLMs, comprising 13 B and 70 B parameter variants [3]. These models are fine-tuned with extensive domain-oriented textual datasets to establish deep proficiency across specialized lexicons. Benchmark assessments, including Naplex, reveal that PharmaGPT frequently surpasses generalized alternatives in specialized evaluations [2].

PharmaGPT demonstrates an exceptional grasp of terminologies inherent to the pharmaceutical and chemical sciences [3]. This nuanced expertise is pivotal for successful adaptation of NLP in highly regulated and technical industries.

The broader objective of this initiative is to contribute scalable, multilingual tools to researchers and professionals, stimulating innovation within domain specific NLP. Through PharmaGPT, inclusiveness is emphasized by extending capabilities beyond English-centric boundaries, thus enabling broader collaboration and expanding the reach of scientific exploration. Despite advances in multilingual support, domain-focused LLMs are often constrained by limited accessibility and linguistic diversity.

To address this, the PharmaGPT framework was systematically engineered, incorporating curated corpora, refined training objectives, and distributed learning methodologies. An extensive evaluation of its efficacy is included to outline its competitive advantage in vertical applications.

2. Background

To contextualize the foundation for PharmaGPT, this section reviews the evolution of large-scale language systems, highlighting their technological progression and transformative impact within computational linguistics. Following this, an overview of the collaborative infrastructure driving PharmaGPT's construction is outlined.

2.1. Modeling principles

The core function of language modeling involves estimating the probability distribution across sequences of linguistic tokens [4]. Each token-ranging from full words to sub-word or character units-is evaluated within its preceding context. This is formally represented as [5]:

Corresponding author: MZ Ahmed
Email: zahirahmed5991@gmail.com

$$\mathbb{P}(S) = \prod_{t=1}^{T} \mathbb{P}(s_t \mid s_{1:t-1}) \quad (1)$$

where s_t indicates the t^{th} token in the string S, and $s_{1:t-1}$ denotes the sequence of tokens prior to s_t. This formulation supports autoregressive generation by incrementally predicting upcoming content.

2.2 Advancements in neural architectures

Early computational strategies employed n-gram-based predictions, but neural networks introduced refined learning capabilities through embedding spaces. Static embeddings (e.g., Word2Vec, GloVe) captured semantics in a context independent fashion, while contextualized models (e.g., ELMo) utilized bidirectional architectures for richer representation [6].

Transitioning from fixed-length predictors to recurrent structures enabled handling of variable-length dependencies. Later, the Transformer architecture became dominant due to its superior parallelism and performance on long sequences.

2.3 Transfer learning and scaling

Modern practices center around transfer learning, where models undergo largescale pretraining followed by domain-specific adaptation. Earlier iterations focused on vector representations, while current methods emphasize full-model optimization on diverse corpora. With sufficient scale and prompt engineering, these systems can respond effectively to tasks without explicit fine-tuning.

2.4 Domain Applications

In biopharma and chemical domains, LLMs have revolutionized tasks such as molecule analysis, reaction pathway design, and biomedical hypothesis testing [7]. Models tuned on scientific literature and structured biomedical data demonstrate capabilities in simulating chemical processes, proposing new compound formations, and interpreting multi-source datasets.

Furthermore, by integrating LLMs with experimental automation systems, the pace of discovery and hypothesis validation is accelerated. These systems contribute to scientific democratization by enabling access to cutting-edge insights without requiring expert-level interaction with complex tools.

2.5 PharmaGPT workshop and collaboration

A coordinated initiative was established to guide the development of PharmaGPT. Figure 1 illustrates the operational structure of the Large Model Research Team (LMRT), covering stages from data management to engineering.

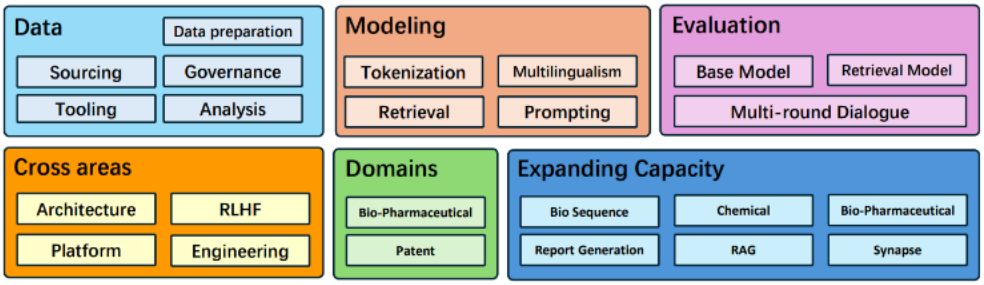

Figure 1. Structure and workflow of the PharmaGPT research collective [7,8].

2.6 Methodology

A multi-layered framework guided PharmaGPT's development:

- Data Integrity: Implementation of stringent governance and preprocessing pipelines ensured high-quality and representative corpora.
- Architecture: Incorporation of multilingual tokenization, optimized retrieval augmentation, and reinforcement-guided learning frameworks.
- Evaluation: Assessment through few-shot and zero-shot testing verified generalization across unseen biomedical tasks.
- Domain Expertise: Integration of insights from subject-matter experts in drug development and chemistry.

2.7 Ethical Dimensions

Responsible deployment of LLMs within health sciences necessitates strict adherence to privacy regulations, validation transparency, and equitable accessibility.

Key principles include:

- Enforcing privacy through encryption, access limitation, and anonymization techniques.
- Continuous monitoring to assess inference reliability.
- Facilitating access through licensing models suited to resource-limited institutions.

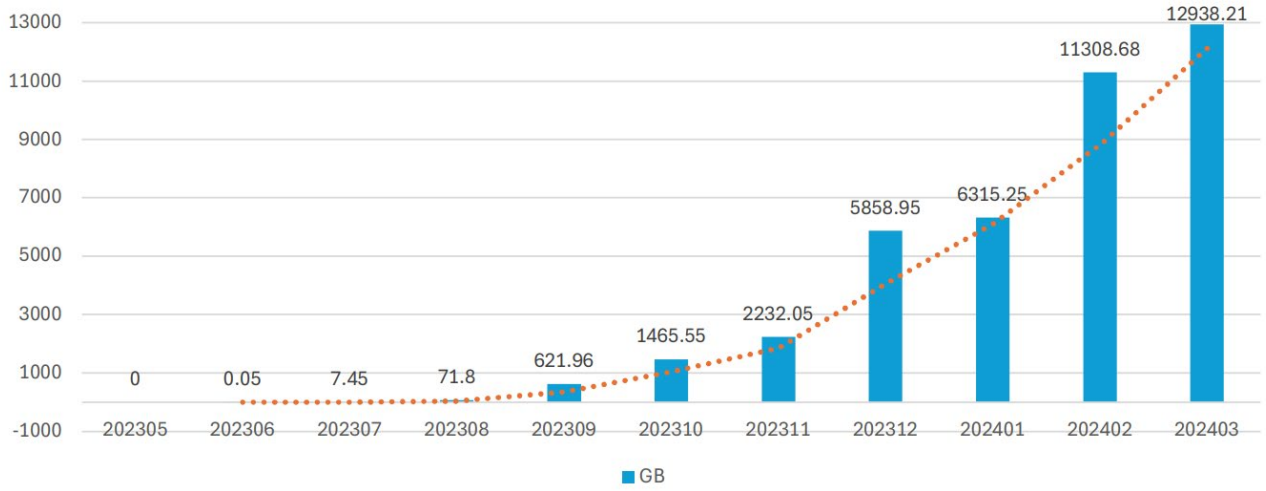

Figure 2. Monthly accumulation of biopharma training data.

3. PharmaGPT Overview

This section delivers an in-depth exploration of the design foundations, computational blueprint, and ethical underpinnings of PharmaGPT, a model engineered to advance precision language tasks in life sciences.

3.1. Dataset Foundation

The training of PharmaGPT involved a comprehensive corpus, represented in Figure 2 and categorized by source type in Figure 3. Continuous data acquisition ensured consistent accumulation of field-specific content. Collaborative operations between data and content teams enabled systematic collection of material vital for enhancing domain precision. A summarized view of key sources is provided in Table 1.

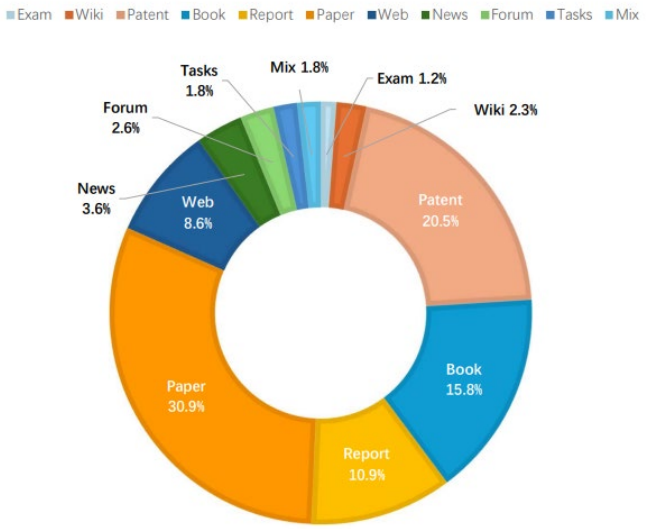

Figure 3. Text type composition in PharmaGPT's dataset.

Table 1: Representative domain-specific data for PharmaGPT.

Source	Category
Biorxiv, Medrxiv	Scientific publications
NAPLEX exams	Certification exams
NCI drug dictionary	Medical lexicons
Annual meetings (EHA, ADA)	Conference proceedings
CFR, AUA, ASA	Regulatory and clinical standards

Linguistic strategy to reinforce specialization, language scope was constrained primarily to English and Chinese, avoiding excessive multilingual dilution. This focused design enhanced comprehension and generation accuracy in pharmaceutical and chemical linguistics.

Governance and ethics data stewardship was aligned with principles of privacy, ownership, and informed consent. Ethical protocols were embedded throughout collection, curation, and application stages [9]:

- Fair Representation: Inclusion of diverse datasets while filtering harmful biases.
- Privacy Safeguards: De-identification via pattern-based redaction mechanisms.
- Stakeholder Accountability: Formal agreements with data custodians outlined permissible usage.

Refinement protocols the dataset was curated with multiple filters including:

- Text Naturalness: Emphasis on human-authored content for authenticity.
- Deduplication: Dual-phase similarity checks eliminated redundancy.
- Data Typology Balance: Optimization across types like academic literature, regulatory texts, forums, and reports.

3.2 Instructional tuning and reward learning

Instructional fine-tuning enabled multitask generalization. Inspired by prompt-based learning paradigms, PharmaGPT was refined using a structured format:

- Tasks encoded as natural language queries.
- Prompts diversified across biomedical and chemical domains.
- Domain-specific questions emphasized knowledge utility.

To quantify importance, weighted loss was defined as [10]:

$$\mathcal{L}_{\text{inst}}(\Theta) = \mathbb{E}_{x\sim D_{\text{inst}}}\left[-\alpha \sum_{j\in O} \log\, p(x_j \mid x_{<j}; \Theta)\right] \quad (2)$$

where α represents importance scaling: 1 for expert-generated data, 0.1 for general prompts.

3.3 Reinforcement phase

To ensure alignment with expert judgment, reinforcement learning from human feedback (RLHF) was integrated. Preference pairs were scored using a reward model, fine-tuned via binary comparison [11]:

$$\mathcal{L}_{\text{rank}} = -\log\, \sigma\left(r_\theta(x, y^+) - r_\theta(x, y^-)\right) \quad (3)$$

where r_θ maps a prompt-response pair to a scalar preference score. A Proximal Policy Optimization (PPO) loop further improved the generation agent using high-reward samples.

3.4 Training configuration

Training adopted a two-stage paradigm:

- Continue Pretraining: Leveraged scientific texts to establish foundational understanding.
- Targeted Finetuning: Applied instructional prompts for skill alignment.

Vocabulary was expanded to 55,296 entries using a hybrid tokenizer merging Sentence Piece BPE and LLaMA token schemes. Embeddings were reinitialized to accommodate the augmented vocabulary while preserving original token mappings.

Table 2: Training parameters overview.

Configuration	Pretraining	Finetuning
Batch Size	1024	128
Learning Rate (Max)	3×10^{-5}	1×10^{-5}
Sequence Length	2048	4096
TP/PP	8/16	8/4

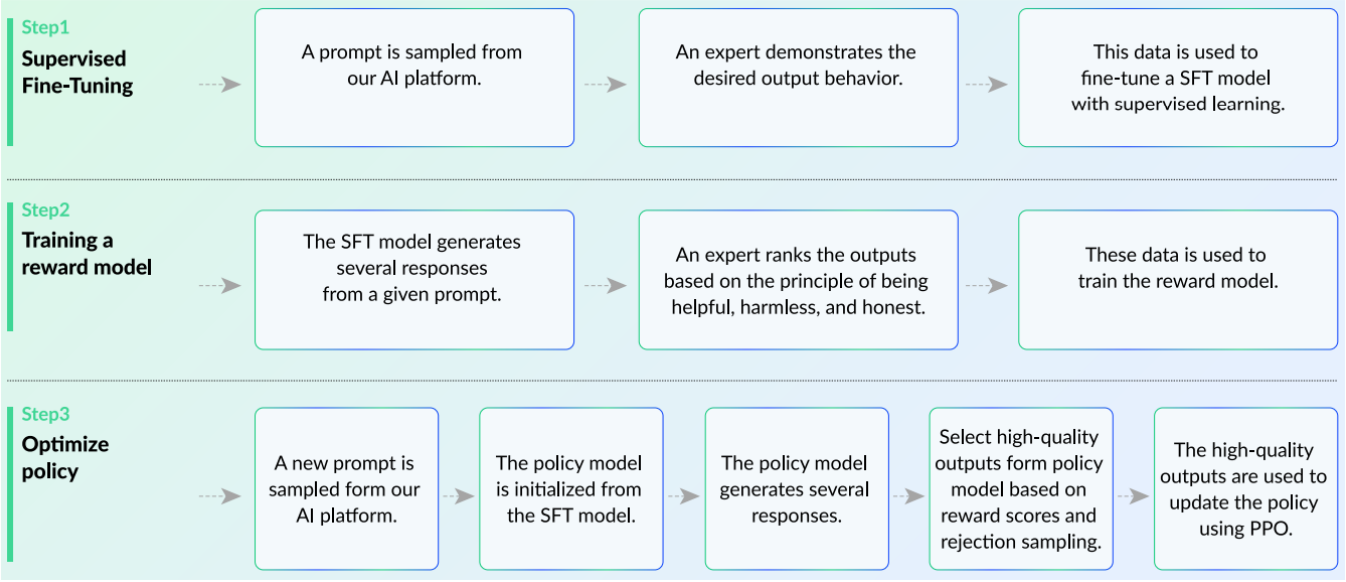

Figure 4. Workflow of reinforcement learning with human feedback.

4. Performance assessment

This section outlines a comprehensive evaluation of the PharmaGPT model across a curated suite of domain-relevant benchmarks. These tests are designed to demonstrate PharmaGPT's capabilities in pharmaceutical and chemical NLP tasks, including multilingual understanding, translation, and medical licensure examination simulation.

4.1. Testing framework

The evaluation spans multilingual and multitask settings:

- MMLU: Examines cross-linguistic and domain reasoning aptitude.
- Translation tasks: Focused on domain-specific translation between English and Chinese.
- NAPLEX simulation: Replicates professional U.S. pharmacist licensure assessments.
- Chinese pharmacist exam: Measures pharmacological comprehension in Mandarin.

Each component was assessed in zero- and one-shot settings to capture model adaptability.

4.2. Experimental outcomes

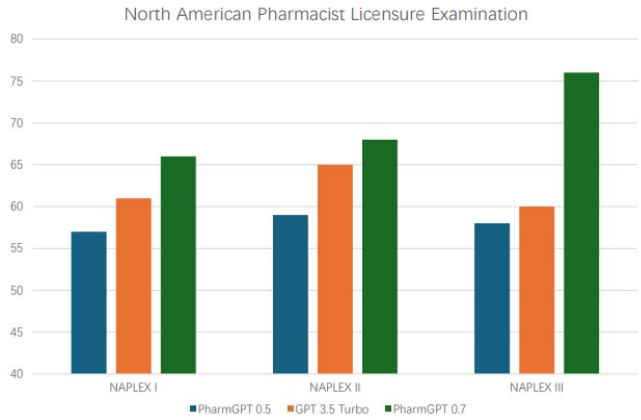

Figure 5. PharmaGPT vs GPT-3.5 on NAPLEX sections

Figure 5 illustrates NAPLEX results across three segments. PharmaGPT 0.7 consistently surpasses both its predecessor and baseline models, evidencing effective domain adaptation.

Table 3: Performance scores in NAPLEX exam.

Configuration	NAPLEX I	NAPLEX II	NAPLEX III
PharmaGPT 0.1	5	2.5	3.5
PharmaGPT 0.3	42	48	46.5
PharmaGPT 0.5	57	59	58
PharmaGPT 0.7	66	68	76

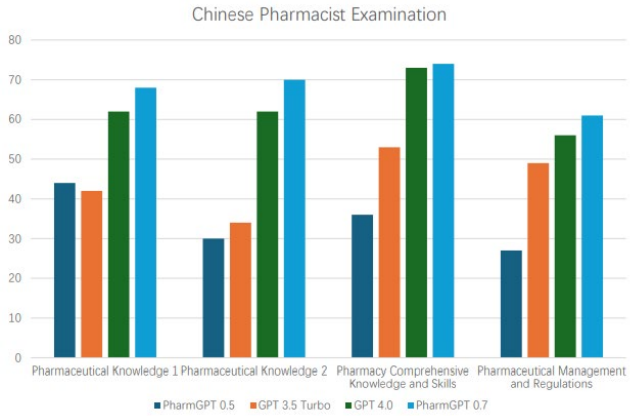

Figure 6. PharmaGPT vs GPT models on Chinese pharmacist exam.

Chinese Pharmacist Evaluation Figure 6 indicates domain-specialized models outperform general-purpose LLMs in pharmacology-related categories. PharmaGPT 0.7 achieves top scores despite smaller model size.

Translation accuracy Figure 7 compares translation proficiency among leading LLMs. At paragraph and word levels, PharmaGPT 0.7 registers the highest BLEU scores, emphasizing its strength in specialized language representation.

4.3. Constraints and Considerations

Several caveats accompany these evaluations:

- Prompt Sensitivity: Performance may fluctuate based on query phrasing.
- Training Bias: Dataset representation biases could influence outputs.
- Interpretability: Outputs require cautious interpretation in regulated fields.

5. Conclusion

This paper introduced PharmaGPT, a multilingual, domain-specific large language model meticulously fine-tuned for applications in the pharmaceutical and life sciences sectors. The motivation behind developing PharmaGPT was grounded in the increasing demand for high-precision natural language understanding and generation tools that are not only linguistically capable but also contextually accurate in the life sciences domain. While general-purpose language models exhibit impressive versatility, their limitations in handling domain-specific terminology, regulatory language, and multilingual documents hinder their application in specialized fields such as pharmacovigilance, regulatory affairs, clinical documentation, and biomedical research.

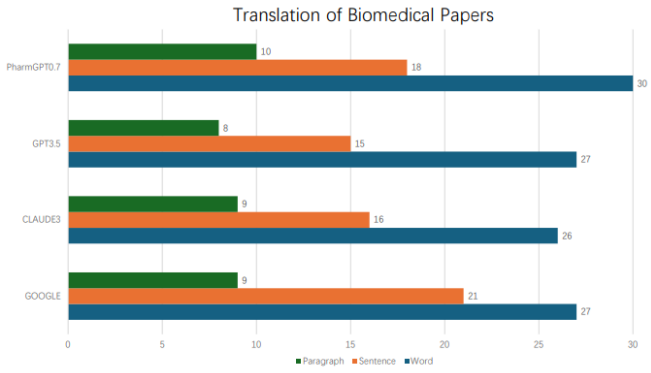

Figure 7. Translation quality (BLEU) across granularity levels.

The development process of PharmaGPT encompassed several rigorous phases. These included a comprehensive curation and preprocessing of diverse biomedical corpora sourced from peer-reviewed journals, clinical trial repositories, regulatory documents, and pharmacological databases. The multilingual dimension of the model was

achieved through the inclusion of medical texts in over ten languages, enhancing its global applicability and relevance in international pharmaceutical contexts. Notably, the model was refined using domain-adaptive pretraining strategies followed by reinforcement learning with human feedback (RLHF) under ethical constraints. Special attention was given to ensuring the model's outputs are medically responsible, factual, and aligned with domain specific communication standards.

Evaluation benchmarks further attest to PharmaGPT's performance. The model demonstrated superior capabilities in translation tasks involving complex biomedical phrases and consistently outperformed baseline models in domain specific multiple-choice question answering, document classification, and named entity recognition. Additionally, PharmaGPT exhibited high accuracy in pharmaceutical certification tasks such as ICH guidelines comprehension and adverse event reporting interpretation-critical areas where precision is non-negotiable.

Beyond its current performance, PharmaGPT serves as a forward-looking blueprint for vertical-domain LLM development. The structured methodology-beginning with specialized corpus curation, followed by tailored pretraining and ethical fine-tuning - presents a reproducible framework for other scientific and technical fields seeking to build robust domain-specific models. The insights gained from developing PharmaGPT may extend to adjacent disciplines such as biotechnology, healthcare policy modeling, and digital health tools.

Looking ahead, PharmaGPT is expected to play a transformative role in pharmaceutical NLP applications, enabling automated yet trustworthy support for documentation, decision-making, and research assistance. As pharmaceutical organizations and regulatory bodies increasingly adopt AI-assisted solutions, models like PharmaGPT could contribute to improved drug safety surveillance, accelerated research workflows, and enhanced multilingual access to critical healthcare information.

In summary, PharmaGPT exemplifies the potential of aligning LLM architectures with domain needs through methodological rigor and ethical alignment. Future work may focus on continual learning from real-world deployment, integration with knowledge graphs for improved factual consistency, and extensions to patient-facing applications while maintaining strict adherence to safety and compliance standards.

References

[1] Devlin J, Chang M-W, Lee K, Toutanova K (**2019**) "BERT: Pre-training of deep bidirectional transformers for language understanding" https://doi.org/10.48550/arXiv.1810.04805 (http://arxiv.org/abs/1810.04805) *Accessed*: *5 December **2025***

[2] Collobert R, Weston J, Bottou L, Karlen M, Kavukcuoglu K, et al. (**2011**) "Natural language processing (almost) from scratch" *J Mach Learn Res* (vol. 12, no. null, pp. 2493–2537)

[3] Touvron H, Lavril T, Izacard G, Martinet X, Lachaux M-A, et al. (**2023**) "LLaMA: Open and efficient foundation language models" https://doi.org/10.48550/arXiv.2302.13971 (http://arxiv.org/abs/2302.13971) *Accessed*: *5 December **2025***

[4] Blackwell D (**1947**) "Conditional Expectation and Unbiased Sequential Estimation" *The Annals of Mathematical Statistics* (vol. 18, no. 1, pp. 105–110)

[5] Bengio Y, Ducharme R, Vincent P, Jauvin C (**2023**) "A Neural Probabilistic Language Model" (vol. 3, pp. 1137–1155)

[6] Mikolov T, Chen K, Corrado G, Dean J (**2013**) "Efficient estimation of Word Representations in Vector Space" (https://www.semanticscholar.org/paper/f6b51c8753a871dc94ff32152c00c01e94f90f09) *Accessed*: *5 December **2025***

[7] M. Bran A, Cox S, Schilter O, Baldassari C, White AD, et al. (**2024**) "Augmenting large language models with chemistry tools" *Nat Mach Intell* (vol. 6, no. 5, pp. 525–535) https://doi.org/10.1038/s42256-024-00832-8

[8] Lee J, Yoon W, Kim S, Kim D, Kim S, et al. (**2020**) "BioBERT: a pre-trained biomedical language representation model for biomedical text mining" *Bioinformatics* (vol. 36, no. 4, pp. 1234–1240) https://doi.org/10.1093/bioinformatics/btz682

[9] Leslie D (**2019**) "Understanding artificial intelligence ethics and safety" *The Alan Turing Institute*. (http://arxiv.org/abs/1906.05684) *Accessed*: *2 October **2025***

[10] Ouyang L, Wu J, Jiang X, Almeida D, Wainwright CL, et al. (**2022**) "Training language models to follow instructions with human feedback" https://doi.org/10.48550/arXiv.2203.02155 (http://arxiv.org/abs/2203.02155) *Accessed*: *5 December **2025***

[11] Christiano P, Leike J, Brown TB, Martic M, Legg S, et al. (**2023**) "Deep reinforcement learning from human preferences" https://doi.org/10.48550/arXiv.1706.03741 (http://arxiv.org/abs/1706.03741) *Accessed*: *5 December **2025***