OPEN ACCESS

Journal of Science & Cycling
Breakthroughs in Cycling & Triathlon Sciences

Conference Abstract

Science and Cycling Conference, Lille 2025

Applications of Language Modelling for a Cycling Aerodynamics' Coach

Callum Barnes [1,*] James Hopker [2] and Stuart Gibson [1]

Received: 28 February 2025 **Accepted:** 17 March 2025 **Published:** 19 November 2025

[1] School of Physics and Astronomy, Division of Natural Sciences, University of Kent

[2] School of Sports and Exercise Science, Division of Natural Sciences, University of Kent

Correspondence

Callum Barnes

Affiliation School of Physics and Astronomy, Division of Natural Sciences, University of Kent

cb835@kent.ac.uk

Abstract

This study investigates the application of Language Modelling in cycling aerodynamics. A novel ground truth is created through recruiting a cohort of experts in cycling aerodynamics, bike fit and biomechanics and taking that ground truth to be the collective expert consensus. Within this study 9 Large Language Models and 1 Large Reasoning Model were tested with 7 of the Large Language Models being open-source models from Google, Meta, Microsoft and Alibaba and the closed source models from OpenAI. This study tested these models without a system prompt, with a system prompt, with applied Retrieval Augmented Generation, with an enthusiast level knowledge base and Retrieval Augmented Generation with a more technical knowledgebase. The best performing model in this study was OpenAI's Chat-GPT 4o with an average mark of (90 $\pm$ 0.41)%. And the best performing opensource model was Alibaba's Qwen2.5:32b with a system prompt and the technical knowledge base providing an average score of (88.73 $\pm$ 0.29)%. The results from this study show that it is possible to develop a model which performs to a similar level of a human expert within the domain of aerodynamics, bike fit and biomechanics in cycling. Additionally, this study proposes a method to experimentally quantify the improvements an athlete can make through the assistance of a domain specific Large Language Model.

Keywords

positions; bike fit; aerodynamics; body rocket; machine learning, language modelling

1 Introduction

Aerodynamics is a topic widely debated within the field of cycling, and, for good reason. The aerodynamic resistance of a rider comprises of approximately 80% (Kyle & Burke, 1984) of the resistances a cyclist faces whilst cycling. Traditionally aerodynamics is tested in a wind tunnel or in the velodrome, both very controlled and consistent environments. However, wind tunnels and velodromes are not representative of the conditions athletes race in and can be very expensive facilities to hire with knowledgeable experts required to help interpret the test data. Within the cycling fraternity the pursuit and development of an in-field aerodynamic testing solution is advancing with more on-bike aerodynamic solutions appearing on the market, including Aerosensor (Aerosensor, 2025) , Streamlines (Streamlines, 2025) and Gibli (Gibli, 2025). One such sensor is the Body Rocket (Body Rocket, 2025) system which allows for the direct measurement of a riders aerodynamic drag in complete isolation of the bike and other resistive forces. The system comprises of 4 force sensors: one on the handlebar, one on the saddle and one on each pedal. The locations of the sensors can be seen in Figure 1. As each of these sensors are at the contact points of the bike, the drag of the rider can be determined through measuring the horizontal force at the sensor.

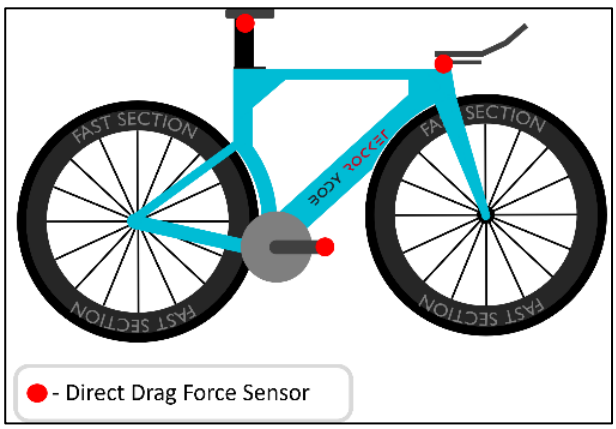

Figure 1. Diagram of where the four sensors of the Body Rocket system are located.

However, even though the rider now has access to on-bike in-field dynamic devices, there still is a key element missing from the full aerodynamic experience - the expert in the room. Whilst an experienced athlete can infer fast positions by looking at the peloton this is not, solely, what an aerodynamic coaches/experts would do. They would refer to their own knowledge and experiences, analyse the data then make iterative improvements with respect to that data providing a unique and prescribed aerodynamic position and equipment configuration.

Within recent years Language Modelling has seen a surge in popularity, most notably with Chat-GPT (OpenAI, 2025). The possibility of Large Language Modelling (LLM) based Artificial Intelligence (AI) being used to affectively share knowledge and analyse data within the domain of aerodynamics in cycling, is interesting as a solution to the limited access one may have to expert level knowledge.

Large Language Models are transformer based network architectures (Brown et al., 2020; Vaswani et al., 2017) which embed a given text into an n-dimensional vector representation then, with respect to the magnitude and direction of all the n-dimensional vectors, predicts the next vector which is then converted into text. This procedure is repeated until the full response is completed. LLM's have applications in automatic writing suggestions, summarising large bodies of text and in this application, a chatbot such as OpenAI's Chat-GPT (OpenAI, 2025) and googles notebookLM (Google, 2025). As these models are transformer-based models they rely on matrix multiplication across a number of layers, the number of weights in the matrices that make up the LLM is representative of the number of parameters of the model. It is expected that as the number of parameters increases the performance of the model will increase (C. Lee et al., 2021).

Retrieval Augmented Generation is a method where knowledge is embedded in a vector store and when the information is needed a retriever can retrieve the relevant information from the vector store including it in the response. This lowers the likely-hood of hallucinations (J. Li et al., 2024).

This study specifically focuses on general and un-finetuned models to determine if a relatively simple RAG pipeline has the capability to accurately provide knowledge within the domain of aerodynamics in cycling.

Therefore, this study aims to examine whether LLM's could be used to provide accurate knowledge in response to questions posed about cycling aerodynamics. Secondly, the study aimed to assess the capability of a relatively small in-house developed LLM with a simple Retrieval Augmented Generation pipeline Figure 2 compared to commercial based models such as ChatGPT.

2 Material and Methods

Following institutional ethical approval from the University of Kent, this study recruited 13 experienced coaches, bike fitters and aerodynamics specialists to help validate the domain specific knowledge from the Language Models.

Initially created through the use to numerous generative AI techniques a 50-question multiple choice online questionnaire (MCQ) was iteratively developed with expert assistance until suitably challenging. This MCQ was then shared with the wider cohort of experts to gain their responses as a "ground truth" comparator for the LLM. Questions were compiled on a range of topics including bike fit, biomechanics, physiology and aerodynamics.

Once the questionnaires were completed the collective expert consensus was found. This was then used to test the various LLMs and quantify their abilities. On a high performance server equipped with the NVIDIA Tesla T4 (Nvidia, 2025) opensource models from Google, Meta, Microsoft and Alibaba, as well as closed source models from OpenAI were tested in their knowledge retrieval against the same set of 50 questions provided to the experts alongside an additional 13 specific to Body Rocket. The models tested in this study were purposefully small and easy run. The specific open-source models tested were: llama3.2:3b and llama3.1:8b, (Grattafiori et al., 2024), Gemma2:9b and Gemma2:27b (Team et al., 2024), Phi4:14b (Abdin et al., 2024), and Qwen2.5:7b and Qwen2.5:32b (Qwen et al., 2025). The closed source models tested were Chat-GPT 3.5 turbo, Chat-GPT-4o alongside the Large Reasoning Model o1-mini. The models were examined on these questions over 10 iterations to obtain a representative evaluation of the capabilities of each of the model.

Each model from each respective source was tested in 4 different configurations, one without a system prompt, one with a system prompt, one with a general knowledge base covering the knowledge of an enthusiastic cyclist. The final configuration involved the use of material covering that of the questionnaire alongside material not covered in the questionnaire, akin to a student receiving course material and being examined on it.

Figure 2 shows the RAG pipeline implemented in this study based upon a public git repository (AllAboutAI-YT, 2024). This pipeline re-writes the users query, embeds it with the mxbai-embed-large embedding model (Lee et al., 2024; Li & Li, 2023) computes the similarity with the stored knowledge, retrieves the relevant context, and then generates a response from the LLM providing a final response with the user.

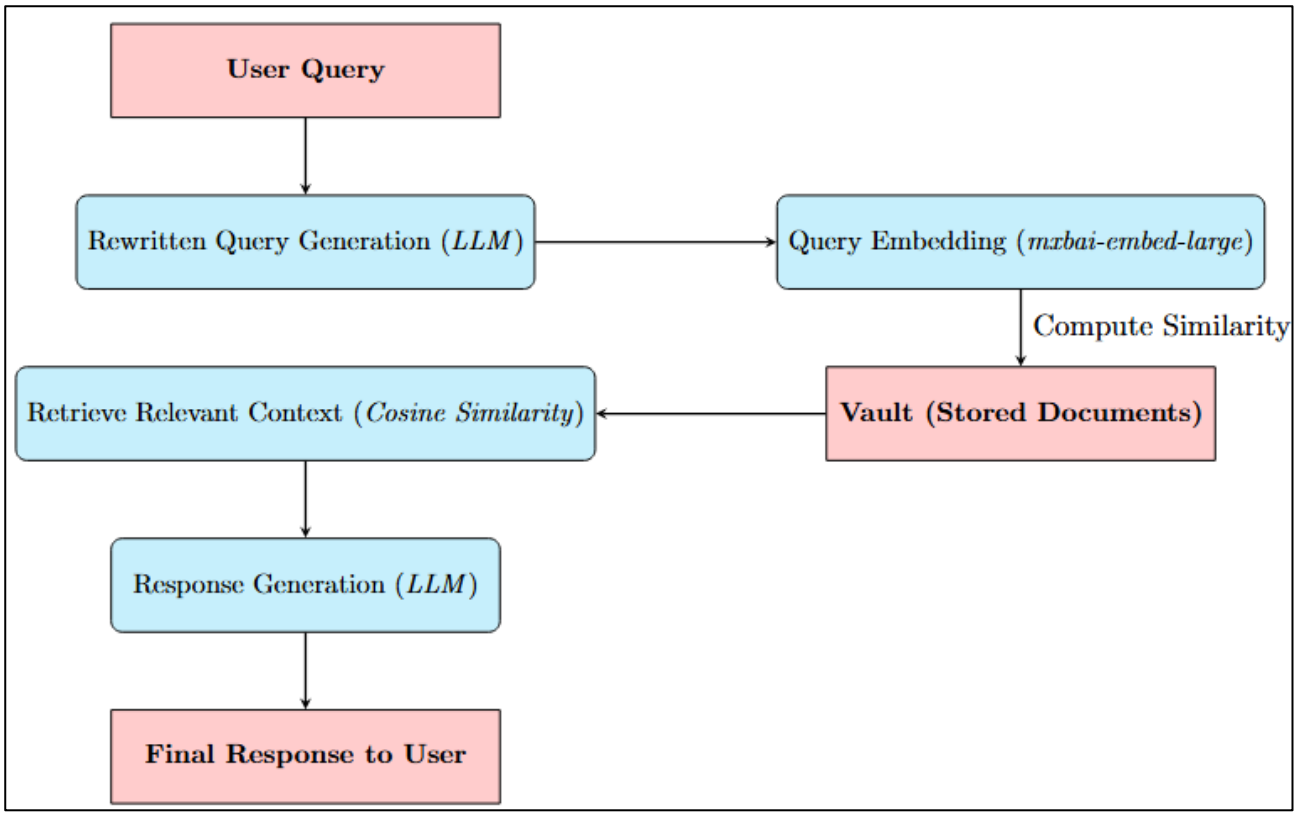

Figure 2. Diagram of the Retrieval Augmented Generation pipeline

2.1 Statistical Analysis

The standard error from each model was subsequently calculated as a measure of the total error rate for each model vs the collective expert consensus.

$$SE = \frac{\sigma}{\sqrt{n}}$$

Equation (1)

3 Results

3.1 Llama Models

In Figure 3, it can be seen that there is a general trend towards the larger model performing better. Specifically, the llama3.2:3b model obtained an average mark of (76.35 ± 1.07)% and the best performing llama3.1:8b obtained an average mark of (78.10 ± 0.62)%.

3.2 Gemma2 Models

Similarly, Figure 4 demonstrates the larger the model the higher the accuracy when compared to the collective expert consensus. Here the best performing Gemma2:9b model obtained an average mark of (73.81 ± 0.79)%. The best performing Gemma2:27b model performed with an average mark of (87.30 ± 0.53)% .

3.3 Phi4 – 14b

Figure 5 shows the performance of the Phi4:14b model here the configuration with the highest accuracy has an average mark of (85.24 ± 0.53)%.

3.4 Qwen Models

As can be seen in Figure 6. Bar chart showing the performance of the Qwen2.5:7b and Qwen2.5:32b model from Alibaba against the cycling aerodynamics 6 Qwen2.5:32b performs better than Qwen2.5:7b likely due to being a larger 32 billion parameter model with a tuned knowledge base providing an average mark of (88.73 ± 0.29)%. The Qwen2.5:7b 7 billion parameter model, using the same techniques, obtains(82.7 ± 0.55)%.

3.5 Open AI Models

As can be seen in Figure 7 the best OpenAI model is Chat-GPT 4o with a system prompt providing an average mark of $(90 \pm 0.41)\%$.

3.6 The Best Performing Models

From Figure 8. Bar chart of the best performing models in this studythe best performing models in this study can be seen. Interestingly the best performing model here is Chat-GPT with a score of $(90.00 \pm 0.41)\%$ followed by Qwen2.5:32b with a mark of $(88.73 \pm 0.29)\%$. This shows that the best model tested in this study is a closed source model, however, with the best open-source model less than 1.5% behind the performance could be considered to be comparable.

3.7 Expert Performance

As can be seen in Figure 9 there are 8 marks between the best and worst performing expert, excluding the outlier with a mark of 36%. Interestingly the average expert mark is $(80.83 \pm 0.83)\%$ and the best expert obtained a score of 86%. This shows that the questions produced were suitably difficult even for the domain experts. To make this comparable the best performing LLM was assessed on the same 50 questions the experts were provided.

Figure 10 shows both LLMs outperform the domain experts in their knowledge retrieval when tested on the same 50 questions. Whilst LLMs performed better than the experts in our study, this may be partly attributable to their superior, generic question, answering capability in addition to specific domain knowledge and understanding. This is a potential confound in the work although we estimate the effect to be weak. An improvement on this would be to include more questions and make the questions more open as opposed to closed multiple choice questions, this would require the application of comparison techniques such as BLEU and ROUGE, which have been used in previous studies (Chatoui & Ata, 2021).

The results of this study show that there is a general trend where the performance of a model is proportional to the number of parameters incorporated within the model, and the quality of the dataset/knowledge base provided. This study demonstrates it is possible to create a high performance LLM through the creation of a unique MCQ sheet in the domain of aerodynamics in cycling. However, results would be improved by increasing the number of experts recruited and the number of questions within the MCQ sheet. Future work could include applying simple fine-tuning techniques such as LoRA (Qin et al., 2024) which could lead to an open-source based model obtaining a greater accuracy with minimal computational resources. As these models start to have a better understanding of the domains fundamentals, combining their outputs with traditional statistical methods (such as standard deviation or coefficient of variation) could support greater interpretation and understanding of aerodynamic performance. For example, some aerodynamic data collected may present high variability in CdA over the course of the testing session and also have a high number of shuffles on the saddle. Previous research indicates that shuffling may have an impact on performance (Barnes et al., 2023), with this in mind the LLM could look at the data and provide insights linking the two trends together.

For future work a domain specific LLM such as those shown in this study could be used in an investigation on the aerodynamic optimisation of cycling body position, equipment and clothing with and without the assistance of the LLM expert.

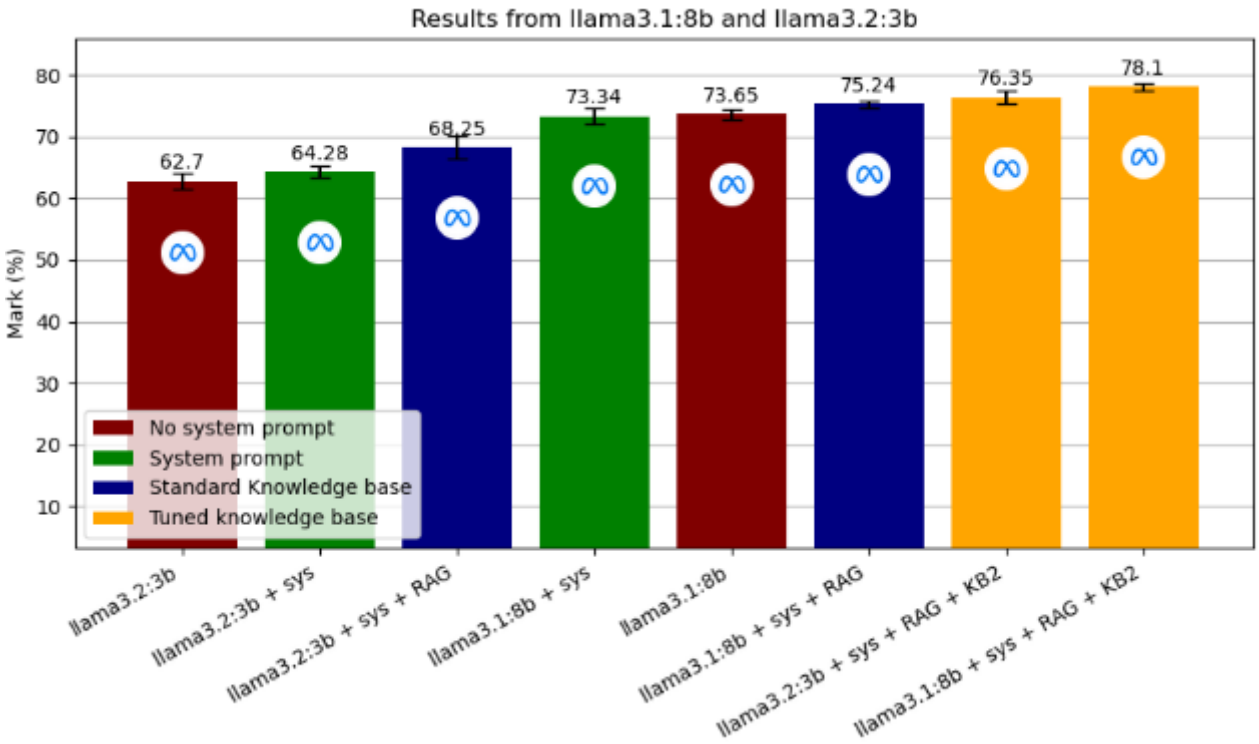

Figure 3. Bar chart comparing the performance of two llama models, llama3.1:8b and llama3.2:3b on the cycling aerodynamics assessment.

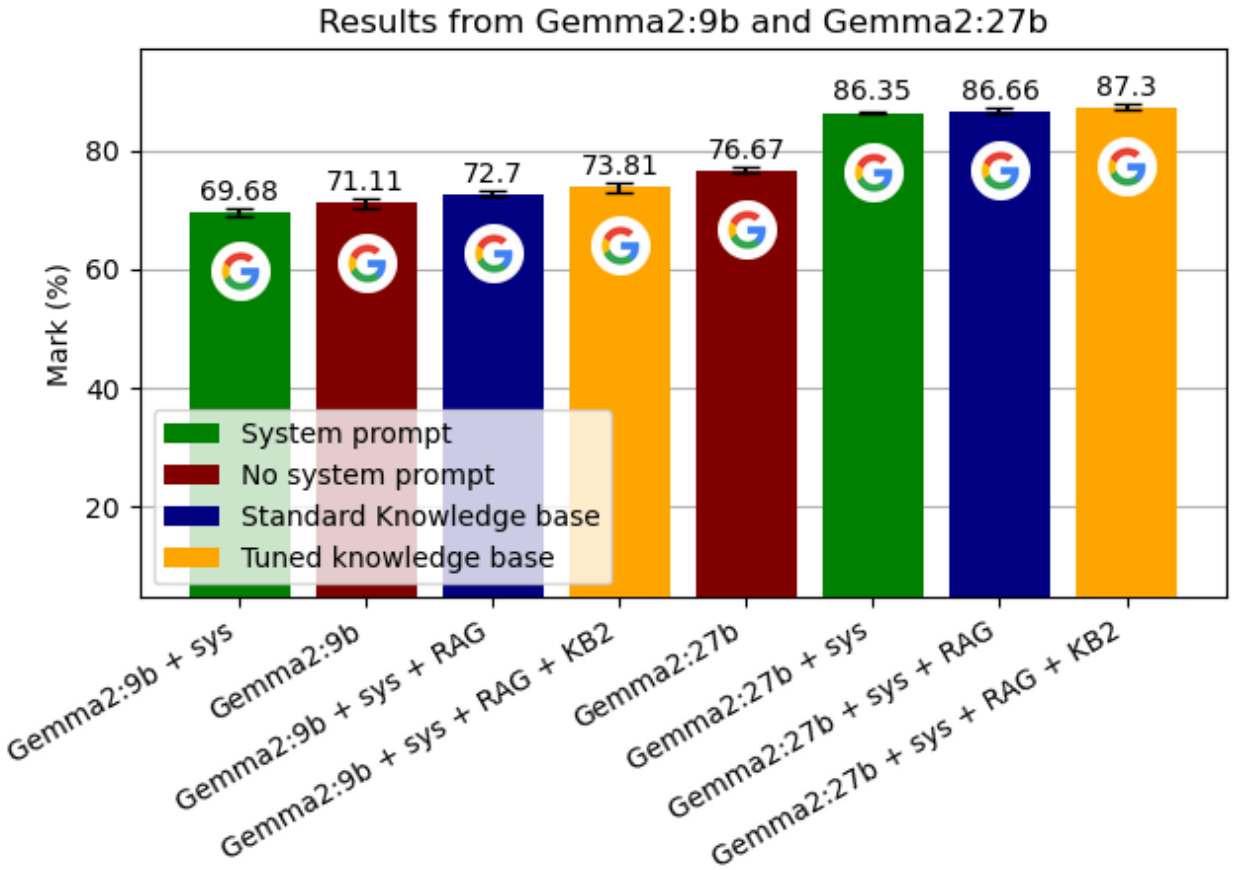

Figure 4. Bar chart comparing the performance of two Gemma2 models, Gemma2:9b and Gemma2:27b

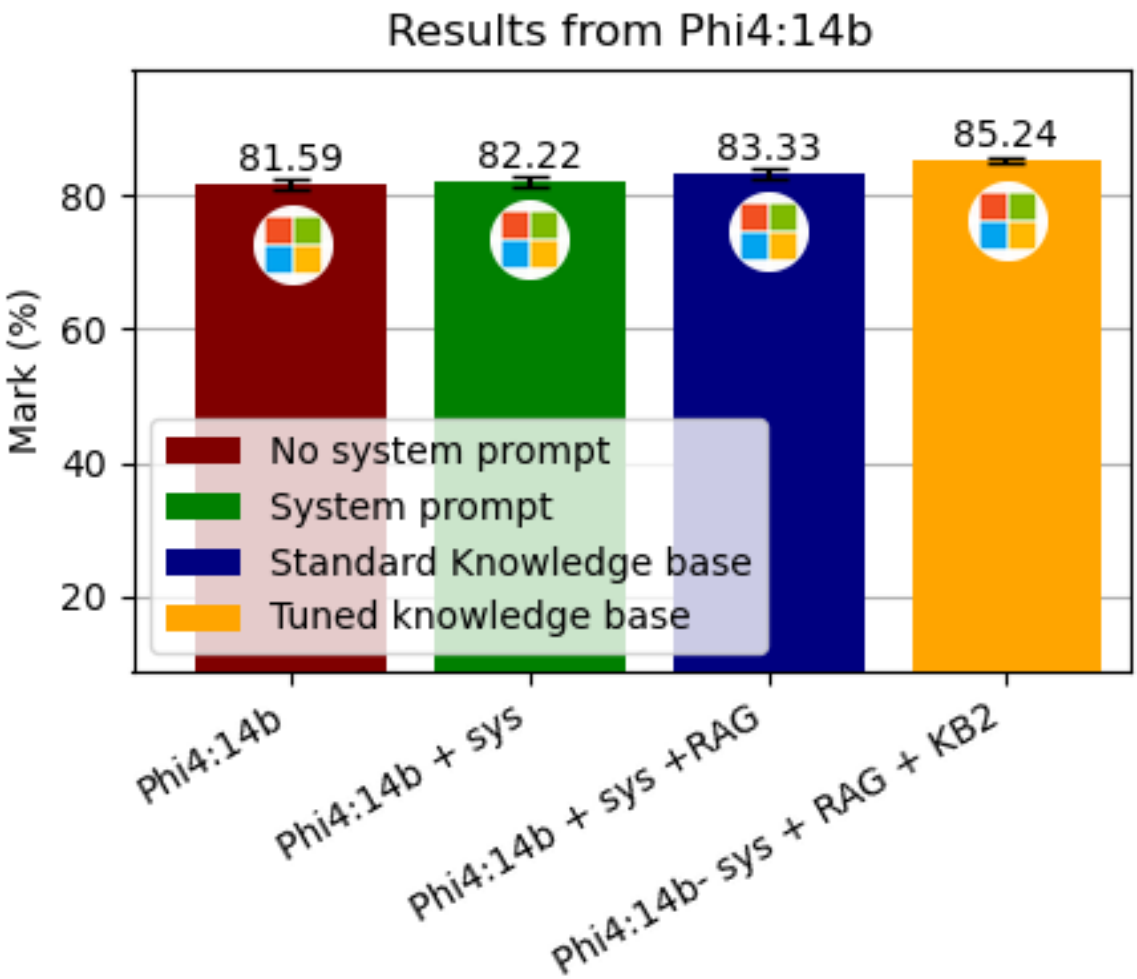

Figure 5. Bar chart showing the performance of the Microsoft Phi4:14b model against the cycling aerodynamics MCQ sheet

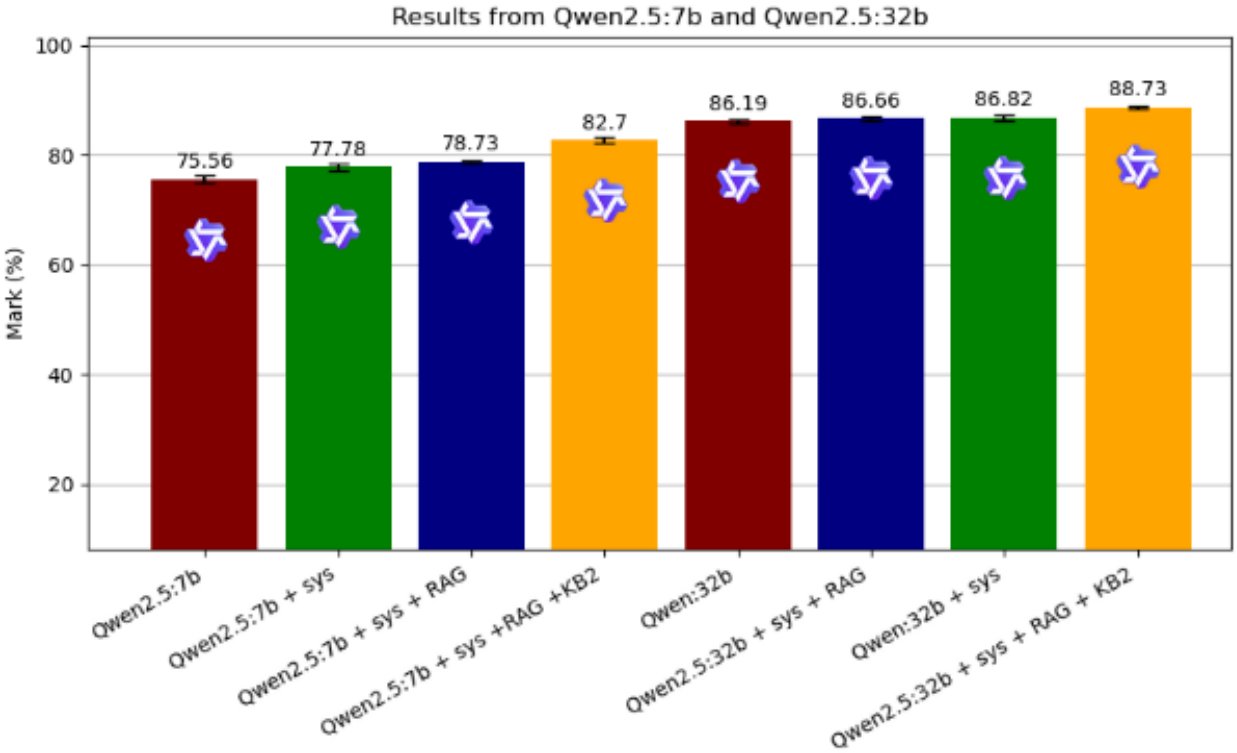

Figure 6. Bar chart showing the performance of the Qwen2.5:7b and Qwen2.5:32b model from Alibaba against the cycling aerodynamics MCQ sheet

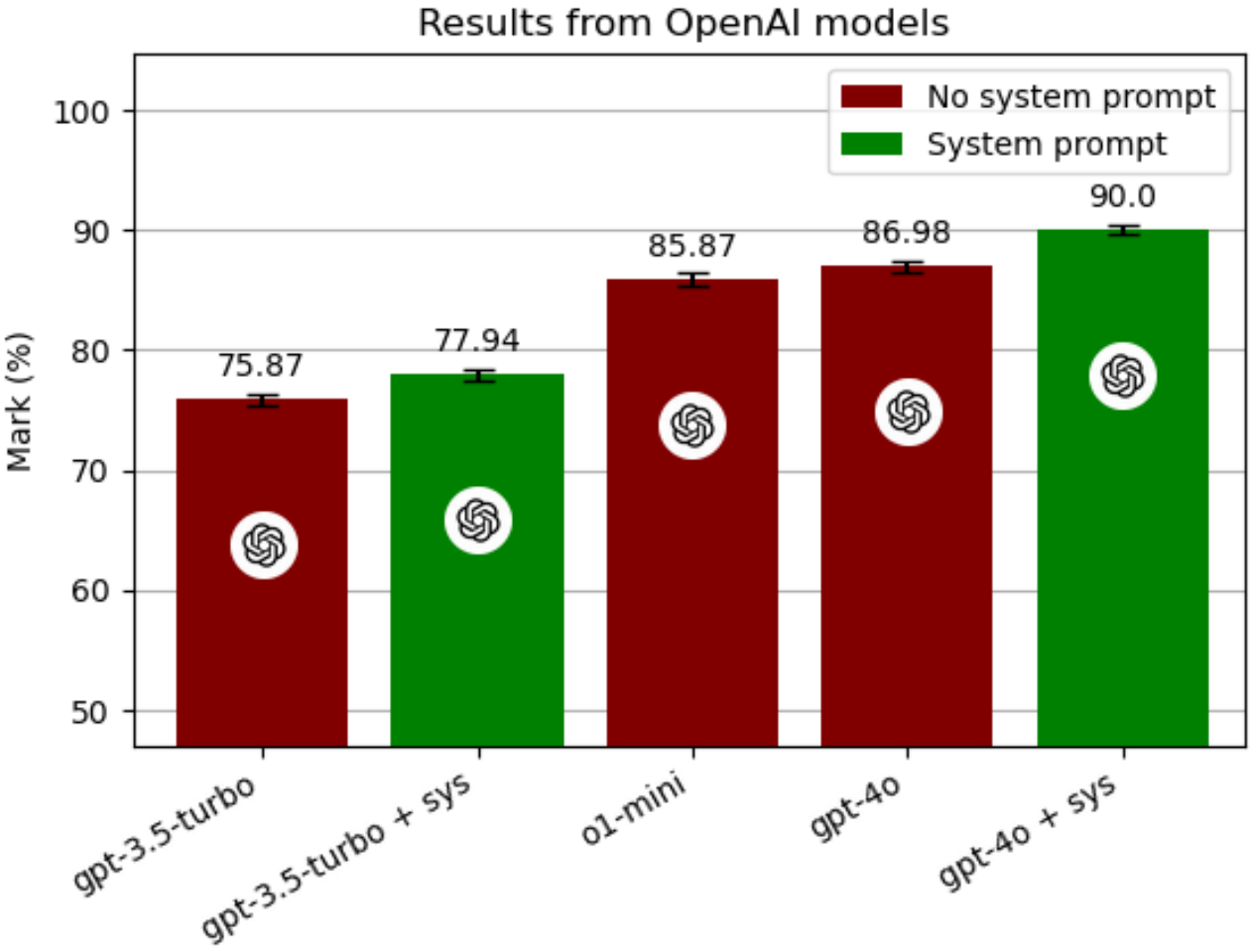

Figure 7. Bar chart showing the performance of three OpenAI models with and without a system prompt and their respective performance agaisnt the Cycling aerodynamics MCQ sheet

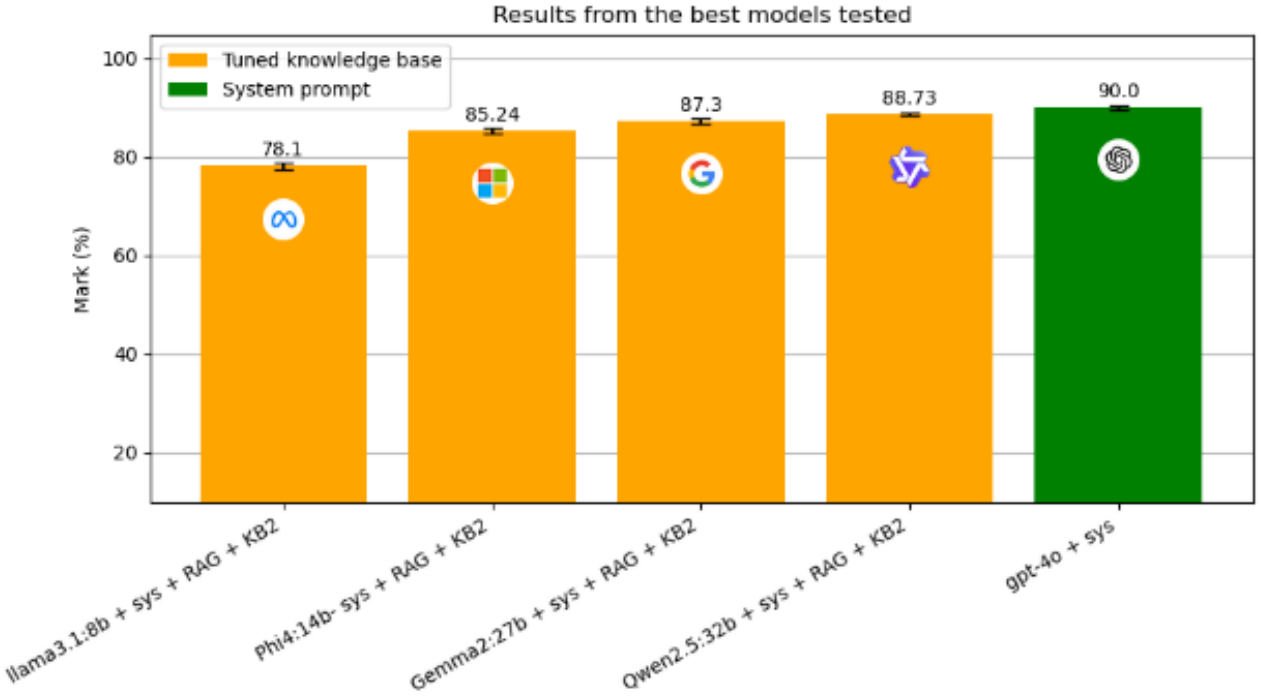

Figure 8. Bar chart of the best performing models in this study

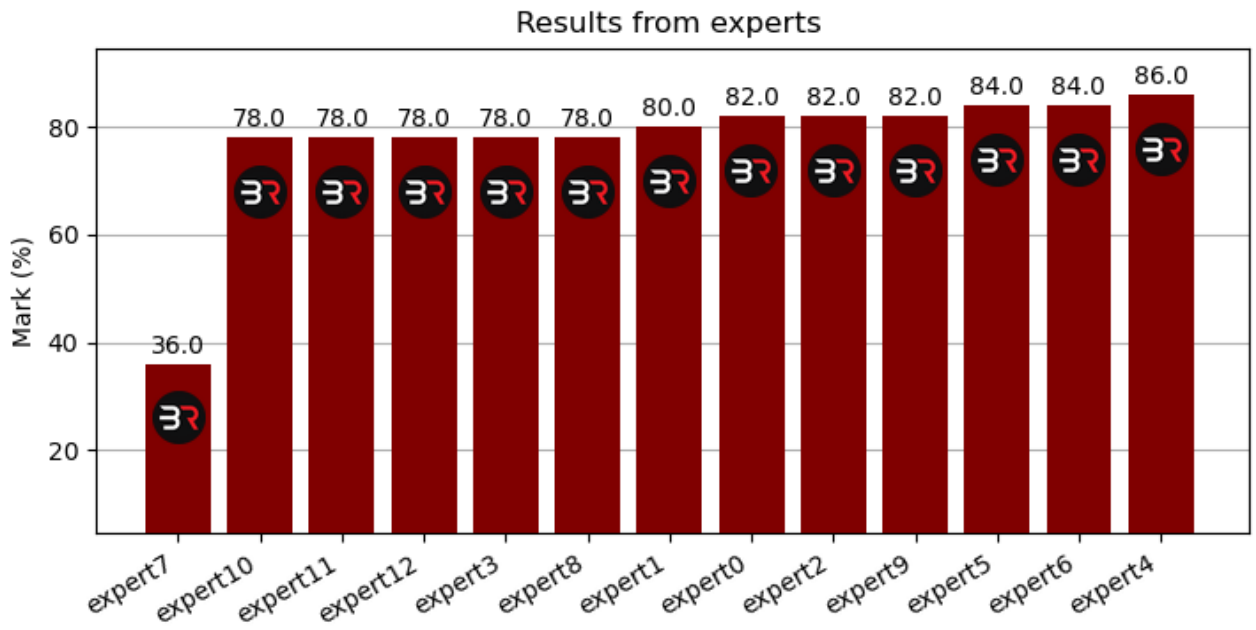

Figure 9. Bar chart showing the performance of all the experts in the study.

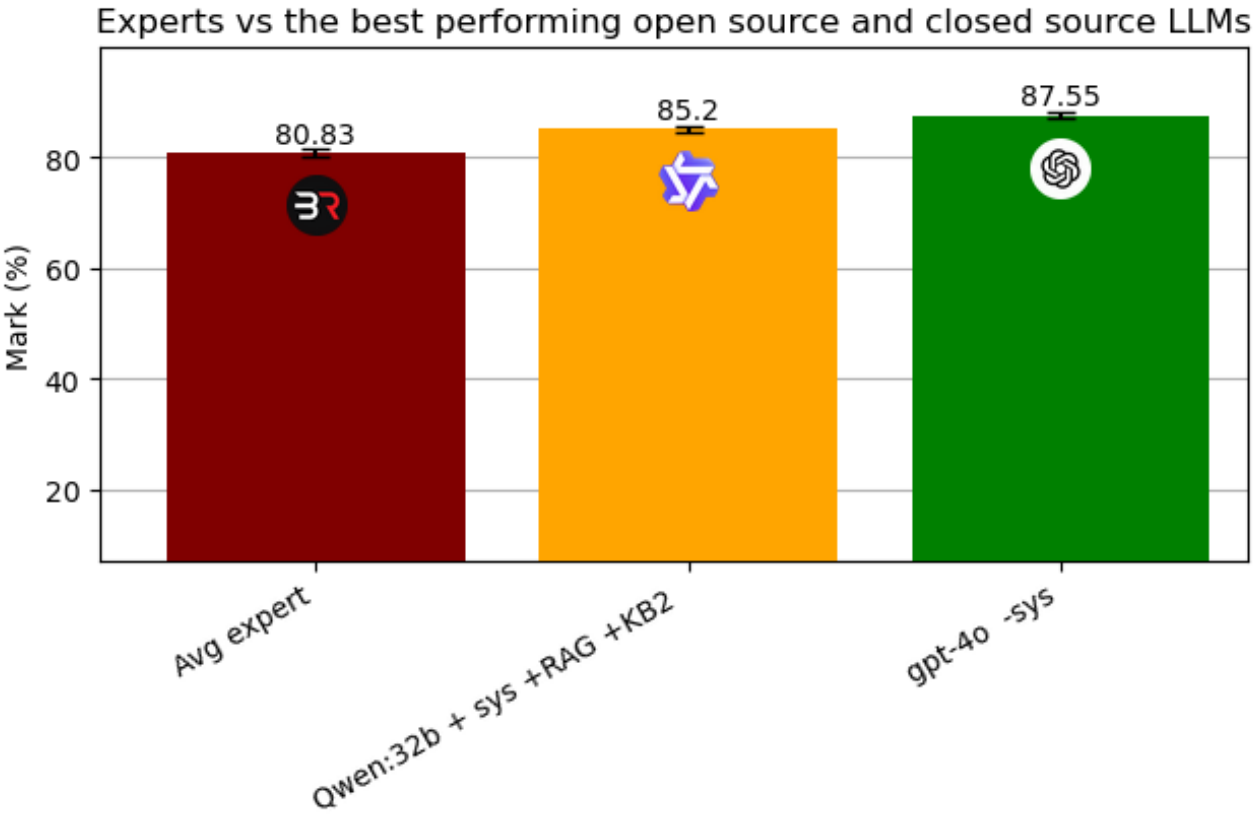

Figure 10. Bar chart comparing the average marks from the experts recruited in this study, the best performing opensource LLM , Qwen2.5:32b and the best performing closed source LLM Chat-GPT4

4 Conclusions

In this study it is clear that open source and closed source LLMs have the ability to retrieve information in the domain of cycling aerodynamics to a similar level to a human expert, and in this study indicating a greater performance. However, there could be features around the style of the question which provides a small advantage to the LLM. Additionally, the performance of the opensource LLMs in this study are similar to the closed source models indicating that there is only a marginal trade-off when using open source LLMs for an in-house system.

Funding: This work was supported by the SEPnet SME-DTN in partnership with Research England and Body Rocket Ltd.

Acknowledgments: Datasets provided by Body Rocket Ltd (Sussex, UK).

Conflicts of Interest: The analysis described in this paper was performed independently by the University of Kent.

References

Abdin, M., Aneja, J., Behl, H., Bubeck, S., Eldan, R., Gunasekar, S., Harrison, M., Hewett, R. J., Javaheripi, M., Kauffmann, P., Lee, J. R., Lee, Y. T., Li, Y., Liu, W., Mendes, C. C. T., Nguyen, A., Price, E., Rosa, G. de, Saarikivi, O., … Zhang, Y. (2024). *Phi-4 Technical Report*. https://arxiv.org/abs/2412.08905

Aerosensor. (2025). *Aerosensor*. https://aerosensor.tech/

AllAboutAI-YT. (2024). *Easy-local-rag*. https://github.com/AllAboutAI-YT/easy-local-rag/tree/main?tab=readme-ov-file

Barnes, C., Hopker, J., & Gibson, S. (2023). To shuffle or not to shuffle. *Journal of Science and Cycling*, *12*(2), 28–30.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., & others. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901.

Chatoui, H., & Ata, O. (2021). Automated Evaluation of the Virtual Assistant in Bleu and Rouge Scores. *2021 3rd International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)*, 1–6. doi: 10.1109/HORA52670.2021.9461351

Gibli. (2025). *Gibli*. https://giblitech.com/

Google. (2025). *Think Smater Not Harder*. https://notebooklm.google/

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Hinsvark, A., … Ma, Z. (2024). *The Llama 3 Herd of Models*. https://arxiv.org/abs/2407.21783

Kyle, C. R., & Burke, E. (1984). Improving the racing bicycle. *Mechanical Engineering*, *106*(9), 34–45.

Lee, C., Kim, Y.-B., Ji, H., Lee, Y., Hur, Y., & Lim, H. (2021). On the Redundancy in the Rank of Neural Network Parameters and Its Controllability. *Applied Sciences*, *11*, 725. doi: 10.3390/app11020725

Lee, S., Shakir, A., Koenig, D., & Lipp, J. (2024). *Open Source Strikes Bread—New Fluffy Embeddings Model*. https://www.mixedbread.ai/blog/mxbai-embed-large-v1

Li, J., Yuan, Y., & Zhang, Z. (2024). Enhancing LLM Factual Accuracy with RAG to Counter Hallucinations: A Case Study on Domain-Specific Queries in Private Knowledge-Bases. *arXiv Preprint arXiv:2403.10446*.

Li, X., & Li, J. (2023). AnglE-optimized Text Embeddings. *arXiv Preprint arXiv:2309.12871*.

Nvidia. (2025). *NVIDIA T4*. https://www.nvidia.com/en-gb/data-center/tesla-t4/

OpenAI. (2025). *Get answers. Find inspiration. Be more productive*. https://openai.com/chatgpt/overview/

Qin, H., Ma, X., Zheng, X., Li, X., Zhang, Y., Liu, S., Luo, J., Liu, X., & Magno, M. (2024). Accurate LoRA-Finetuning Quantization of LLMs via Information Retention. *arXiv Preprint arXiv:2402.05445*

Qwen, :, Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Wei, H., Lin, H., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., … Qiu, Z. (2025). *Qwen2.5 Technical Report*. https://arxiv.org/abs/2412.15115

Rocket, B. (2025). *Body Rocket*. https://www.bodyrocket.cc/

Streamlines. (2025). *Streamlines*. https://streamlines.aero/?srsltid=AfmBOopI1l2mJ9UQ67CIwXkjK1GmD1sEgkplt-LsuyhAmjHYXuDBwB64

Team, G., Riviere, M., Pathak, S., Sessa, P. G., Hardin, C., Bhupatiraju, S., Hussenot, L., Mesnard, T., Shahriari, B., Ramé, A., Ferret, J., Liu, P., Tafti, P., Friesen, A., Casbon, M., Ramos, S., Kumar, R., Lan, C. L., Jerome, S., … Andreev, A. (2024). *Gemma 2: Improving Open Language Models at a Practical Size*. https://arxiv.org/abs/2408.00118

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*.