OnlySportsLM: Optimizing Sports-Domain Language Models with SOTA Performance under Billion Parameters

Zexin Chen[1] **Chengxi Li**[2] **Xiangyu Xie**[3] **Parijat Dube**[4]
[1]New York University [2]Carnegie Mellon University [3]Cornell University [4]IBM Research
zc2404@nyu.edu
chengxil@andrew.cmu.edu
xx358@cornell.edu
pdube@us.ibm.com

Abstract

This paper explores the potential of a small, domain-specific language model trained exclusively on sports-related data. We investigate whether extensive training data with specially designed small model structures can overcome model size constraints. The study introduces the `OnlySports` collection, comprising `OnlySportsLM`, `OnlySports Dataset`, and `OnlySports Benchmark`. Our approach involves: 1) creating a massive 600 billion tokens `OnlySports Dataset` from FineWeb, 2) optimizing the RWKV-v6 architecture for sports-related tasks, resulting in a 196M parameters model with 20-layer, 640-dimension structure, 3) training the `OnlySportsLM` on part of `OnlySports Dataset`, and 4) testing the resultant model on `OnlySports Benchmark`. `OnlySportsLM` achieves a 37.62%/34.08% accuracy improvements over previous 135M/360M state-of-the-art models and matches the performance of larger models such as SomlLM 1.7B and Qwen 1.5B in the sports domain. Additionally, the `OnlySports` collection presents a comprehensive workflow for building high-quality, domain-specific language models, providing a replicable blueprint for efficient AI development across various specialized fields.

1 Introduction

General-purpose large language models (LLMs) have demonstrated remarkable capabilities across various tasks [17]. However, such performance comes at the cost of excessive computational resources and sometimes inefficiencies in domain-specific applications. Domain-specific language models offer a promising alternative, potentially achieving comparable or superior performance in targeted areas while significantly reducing model size.

Despite their potential, recent domain-specific models face several challenges. Large models such as BloombergGPT [23], while powerful, requires extensive computational resources (e.g., 64×8 A100 40GB with a total of 1.3 million GPU hours), making them infeasible for most research institutions. Additionally, many domain models suffer from a lack of high-quality domain-specific text data, with models like BioMedLM [4] trained on only 34.6 billion tokens and SportsBert [16] on merely 1-2 billion tokens. Furthermore, most domain models follow the model structure of general models, leaving room for optimization, especially for smaller model sizes.

In light of these challenges, recent research on small general-purpose language models, such as MobileLLM [13] and SmolLM [1], has provided valuable insights into efficient model structures. However, their effectiveness in domain-specific modeling remains unproven. To address these

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

challenges and leverage recent insights, we propose a new approach for small domain-specific language models, utilizing specialized model structures and a collection pipeline for large in-domain corpus for efficient and cost-effective training.

To verify this approach, we choose sports as the target domain due to its unique combination of broad public interest, rich content, and a constant influx of new data through ongoing events and competitions. Moreover, sports language often contains domain-specific jargon, statistics, and contextual nuances that general-purpose models may struggle to capture accurately. By focusing on sports, we can demonstrate the potential of domain-specific models in a field that is both widely accessible and technically challenging. Additionally, the sports domain provides an excellent testbed for evaluating a model's ability to handle real-time information processing and generation, skills that are crucial in many real-world applications. Based on this approach and domain selection, we present `OnlySports`[1], a novel framework for developing high-performance, small-scale sports language models.

1.1 Contributions

1. `OnlySports Dataset`: A large-scale, high-quality sports-specific text corpus of 600 billion tokens, extracted from the FineWeb dataset [19].

2. `OnlySports Benchmark`: A novel evaluation method for assessing sports knowledge generation, using 1000 diverse prompts and state-of-the-art (SOTA) language models for evaluation.

3. `OnlySportsLM`: A 196 million parameter RWKV-v6[2] [20] based sports language model trained on half of the OnlySports Dataset. In our OnlySports Benchmark, OnlySportsLM outperforms the preceding SOTA general purpose 135M/360M language model by 37.62%/34.08%.

2 Collection of Domain Data

In this section, we present the path to building `OnlySports Dataset`, a comprehensive collection of English sports documents. This dataset comprises a diverse range of content including news articles, blogs, match reports, interviews, and tutorials, all extracted from the FineWeb dataset. FineWeb is a thoroughly cleaned and deduplicated subset of CommonCrawl, spanning from 2013 to present. It represents one of the best open-source datasets for LLM training. Our extraction process involved two key steps: first, we applied URL filtering to identify potentially relevant content, and second, we developed a custom sports text classifier to accurately identify and extract sports-related documents from the filtered data. The resulting `OnlySports Dataset` encompasses 1.2 TB of text, equivalent to approximately 600 billion RWKV tokens. This makes it the largest sport domain dataset to date, significantly surpassing previous collections in both scale and comprehensiveness.

2.1 URL Filtering

To efficiently identify potentially sports-related content within the FineWeb dataset, we implemented a preliminary URL filtering step. We carefully select a list of sports-related terms, encompassing various sports, leagues, brands, and media. This approach allows us to rapidly narrow down the dataset to documents likely to contain sports content.

Our keywords include:

- General sports terms: *sport, athletic, athlete, fitness, workout, gym, league, team, champion, football, soccer, basketball, baseball, tennis, cricket, rugby, golf, volleyball, hockey, cycling, swimming, wrestling, running, boxing, racing, swim, goal*
- Major leagues and organizations: *NFL, NBA, MLB, NHL, FIFA, UEFA, NCAA, MMA, UFC, WWE, Premier League, LaLiga, Bundesliga, SerieA, Ligue1, EPL, NASCAR, MotoGP, Formula1, F1*
- Sports events, brands, and media: *Olympic, cup, playoff, marathon, copa, Nike, Adidas, ESPN, BleacherReport, SI.com, news*

[1]Our Huggingface collection is available at: `https://huggingface.co/collections/Chrisneverdie/onlysports-66b3e5cf595eb81220cc27a6`

[2]Our training code is available at: `https://github.com/chrischenhub/OnlySportsLM`

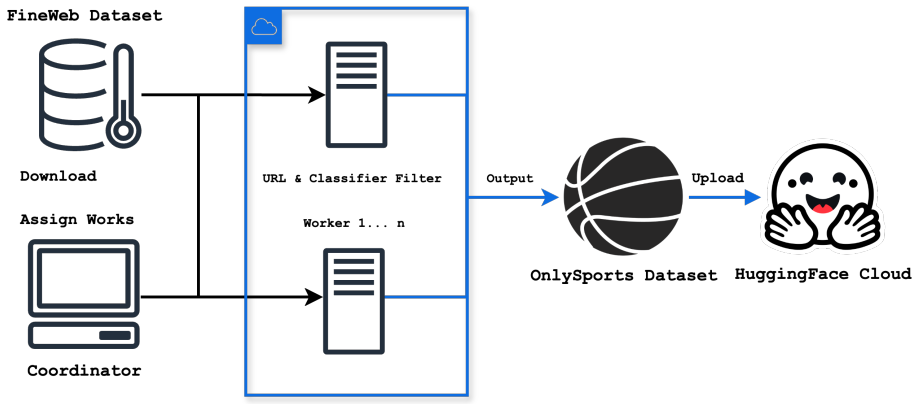

Figure 1: Data pipeline to create `OnlySports Dataset`

We applied these keywords in both their standard and capitalized forms where appropriate (e.g., NBA/nba, FIFA/fifa). This keyword list ensured a high recall in identifying potential sports content, which was then further refined by our classification model. Although the list does not exhaustively cover all sports, the nature of sports websites often includes the word *sport* in their URL, ensuring broad coverage of sports-related content.

Table 1: Sports text classifier performance in the test set, correctly classifying most labels

Class	Precision	Recall	F1-Score	Support
0	0.98	0.98	0.98	3631
1	0.99	0.99	0.99	6429
Accuracy			0.99	10060
Macro Avg	0.99	0.99	0.99	10060
Weighted Avg	0.99	0.99	0.99	10060

2.2 Sports Text Classifier

To develop our sports text classifier, we first created a balanced dataset of sports and non-sports content. We manually scraped 64k samples from seven prestigious sports websites, selected to cover a wide range of sports topics. To balance this, we classified 36k non-sports text documents from a subset of FineWeb using GPT-3.5, ensuring diversity in the non-sports content. We then labeled this combined dataset, designating sports-related text as class 1 and non-sports text as class 0.

For the classification model, we chose *Snowflake-arctic-embed-xs* [15] as our base due to its efficient performance on text classification tasks. We then add a binary classification layer to this model and train it for 10 epochs with a learning rate of 3e-4.

Table 1 presents the performance metrics of our classifier, demonstrating its exceptional accuracy in distinguishing between sports and non-sports documents. The model achieves near-perfect precision, recall, and F1-scores for both classes, with an overall accuracy of 0.99.

2.3 Data Filtering and Conversion

Figure 1 presents a scalable MapReduce architecture [8] to filter sports-related content from the 90TB FineWeb dataset for model training. This approach allows us to overcome limitations in CPU resources and disk space.

In the map phase, we use a Golang-based coordinator with the Gin Web framework to distribute tasks across eight Python-powered worker servers. The filtering process occurs in two steps: 1. URL keyword filtering, which reduced the dataset size by 85%. 2. Application of our sports text classifier for further curation.

The resulting filtered data is stored in parquet format and uploaded to HuggingFace. For the reduce phase, we utilized a high-capacity cloud server to tokenize the parquet files using an open-source Rust script. This streamlined pipeline enabled us to efficiently process the massive FineWeb dataset, extracting a high-quality sports-specific corpus for training `OnlySportsLM`.

3 Optimizing Model Structure for Sports Domain

We explore the potential for model structural optimization before training with the `OnlySports Dataset`. A previous study [13] suggests that general-purpose sub-billion parameter models perform better when using more layers than the traditional 12-layer design while having fewer embedding dimensions. Inspired by this depth versus width study for general models, we hypothesize that domain-specific small models would also follow this deep and thin rule. We explore models with approximately 190M parameters and find results that partially support this principle

3.1 Training Setup

Our experiments are conducted on 8 H100 GPUs. We perform exploratory experiments on a 4.5B tokens subset of `OnlySports Dataset`.

We evaluated the pre-trained model on zero-shot commonsense reasoning tasks, including ARC-easy, ARC-challenge [7], PIQA [3], HellaSwag [27], as well as sports text generation task using `OnlySports Benchmark`.

3.2 OnlySports Benchmark

We introduce a novel evaluation method inspired by the Hellaswag benchmark but targeted specifically for sports knowledge generation. Instead of asking multiple choice questions, our benchmark directly assesses a model's ability to complete sports-related prompts without fine-tuning, providing insight into sports-specific language understanding and generation capabilities. To ensure a comprehensive and relatively unbiased assessment, we employ multiple state-of-the-art language models as evaluators, assessing generated responses across two key criteria: accuracy and factuality, and continuity and relevancy. This approach allows for an evaluation of sports-related text generation capabilities across various models.

3.2.1 Tag and Partial Sentence Generation

To construct our evaluation dataset, we generated 50 diverse sports-related tags encompassing popular sports, major leagues, prominent athletes, and game strategies using GPT-4 API. These tags serve as the foundation for creating a comprehensive set of prompts. For each tag, we craft 20 incomplete sentences, resulting in a total of 1,000 prompts. Each prompt is intentionally designed to end abruptly, providing an ideal context for models to complete. The prompts incorporate well-known sports facts, statistics, or narratives, allowing assessment of a wide range of sports-related knowledge and generation capabilities. For instance, for the tag *#BasketballTeams*, the following partial sentence prompt is generated: *Spurred on by the superstar duo of Shaquille O'Neal and Kobe Bryant, the L.A Lakers clinched three consecutive*. This abrupt ending sets the stage for models to complete the narrative. A well-trained model would likely continue the sentence with "*NBA championships from 2000 to 2002*" or a similar factual completion, demonstrating its ability to maintain contextual coherence and accuracy.

3.2.2 Model Inference and Evaluation Using SOTA LLMs

In our inference process, each prompt is separately fed to the models. We employed consistent hyperparameter settings across all models, with temperature set to 1 and top-p value to 0.3, to ensure the generation of consistent, high-probability outputs. Each response is limited to 80 tokens.

To evaluate the model-generated responses, we adopt an approach inspired by LLM-as-a-judge [29], which approximates human preferences in assessing open-ended text. We utilize two state-of-the-art language models, GPT-4o and Claude 3.5 Sonnet, as evaluators. The assessment is conducted across two distinct criteria at a scale of 1-5, adhering to the principle of multi-dimensional evaluation as recommended by [29]. To mitigate potential biases inherent in large language model judges, we

implement several measures: 1. Deployment of multiple LLM judges to enhance reliability and reduce individual model biases. 2. Standardization of prompts and evaluation criteria to ensure consistency across assessments. After scores are generated by each model, we take the average of them to be the final score.

The input prompt format for evaluation is defined as follows:

- *prompt: (partial sentence fed to the models)*
- *response: [SEP] Answer1 [SEP] Answer2 [SEP] Answer3...*

Where *[SEP]* is a separator token used to distinguish between different model responses.

The two evaluation criteria are defined as follows:

- **Accuracy and Factuality:** Evaluates the model's ability to generate accurate and fact-based continuations, ensuring that the information aligns with well-known sports facts and data. The score is denoted as `OS-acc` on a scale from 1 (mostly inaccurate with significant factual errors) to 5 (fully accurate and factually impeccable).
- **Continuity and Relevancy:** Assesses the relevance of the generated text to the given prompt, ensuring that the continuation is contextually appropriate and directly related to the previous sentence. This criterion, denoted as `OS-rel`, is scored from 1 (poor continuation that diverges significantly from the prompt's context) to 5 (excellent continuation that seamlessly extends the prompt's narrative, context, and style).

For each criterion, a system message with a detailed grading rubric is provided in the appendix for reference.

Table 2: Model performance across varying architectures. Compares models with different layer counts and dimensions on `OnlySports Benchmark` and general zero-shot tasks (ARC-e, ARC-c, PIQA, Hellaswag).

#Layer	#Dim	#Param	final loss	OS-acc	OS-rel	ARC-e	ARC-c	PIQA	HS
12	768	196M	2.344	**1.88**	**2.42**	28.6	22.5	54.5	27.6
16	704	200M	2.360	1.70	2.19	28.9	22.0	53.6	27.8
20	640	196M	**2.335**	1.84	**2.42**	29.7	**23.5**	53.9	**27.9**
24	576	185M	2.338	1.86	2.38	**30.1**	22.3	53.1	27.7
28	512	169M	2.364	1.79	2.38	29.7	22.4	**54.6**	27.8

3.3 Depth and Width Experiments

Our experimental results presented in Table 2 reveal interesting insights about the relationship between model depth and performance. We conduct a study involving models ranging from 169M to 200M parameters, varying in depth from 12 to 28 layers and width from 512 to 768 dimensions. We observe that both traditional wider models and moderately deeper architectures perform well on `OnlySports Benchmark`. While the 12-layer wider model has the highest `OS-acc` (1.88) and `OS-rel` (2.42) scores, the 20 layers model shows comparable results in relevancy score and slightly less `OS-acc` (1.84). This finding, contrary to conclusion by [13] and [1], underscores the need for task-specific architectural experimentation.

General zero-shot tasks exhibit some benefits from increased depth, though less pronounced than in previous studies on general-purpose models. Models with 20 to 28 layers often outperform shallower configurations across various reasoning tasks.

Based on these findings, we selected the L20D640 (20 layers, 640 dimensions) model for further training, balancing strong performance across domain-specific and general tasks. We denote this model as `OnlySportsLM`

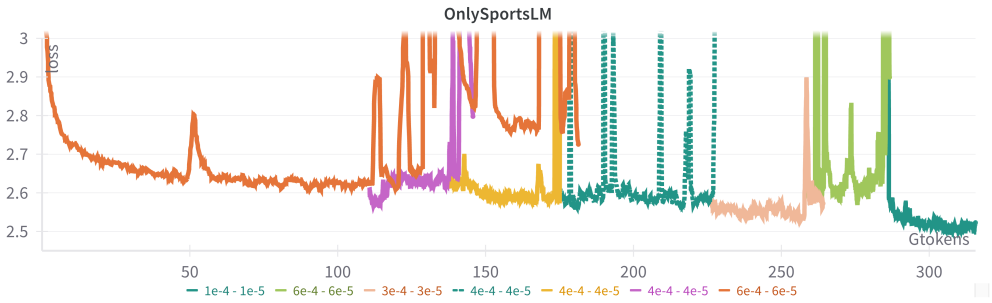

Figure 2: `OnlySportsLM` training loss over time with varying learning rates. The graph shows how loss fluctuates as we adjust the learning rate, starting from higher rates and gradually decreasing to stabilize training and reduce loss spikes. This insight is shared by the author of RWKV [20].

4 Experiments

4.1 Experimental Settings

We train `OnlySportsLM` from scratch utilizing the AdamW optimizer [14] with a weight decay of 0.1 and a context length of 1024 tokens. Our experiments are performed on a cluster of 8 H100 GPUs using Lambda Lab[3], with a per-GPU batch size of 40. Following a cosine decay schedule, the initial learning rate is set to 6e-4. However, due to observed loss spikes during training, the learning rate is subsequently adjusted, ultimately being reduced to 1e-4 (detailed in Figure 2). The ideal training time for all 600B tokens is estimated at 223 hours. However, due to loss spikes, we had to restart the training multiple times, resulting in a longer overall training process. Ultimately, due to constraints on available funding, we were able to train on 315B tokens in 142 hours, completing 7500 steps. This represents approximately half of the `OnlySports Dataset`.

4.2 Main Results

We compare the final `OnlySportsLM` checkpoint on `OnlySports Benchmark` and zero-shot commonsense reasoning tasks (Hellaswag, PIQA, ARC-challenge, and ARC-easy) with previous training checkpoints and recent open-source models. To ensure consistency in evaluation procedures, all models were assessed using their publicly available implementations from the HuggingFace model repository. General benchmark scores are retrieved from their corresponding paper.

4.2.1 Sports Domain Generation

Table 3 compares our `OnlySportsLM` and two recent state-of-the-art general-purpose models, ranging from 137M to 1.7B parameters. We focused on two sets of models: 1. The SmolLM series [1], with 137M, 360M, and 1.7B parameter models, reportedly surpasses the performance of all comparable small language models on general benchmarks. 2. The Qwen2 collection [26], with 500M and 1.5B parameter models, also claims top performance on major benchmarks, even though they were trained on multilingual datasets. These model collections, released in June 2024 and July 2024 respectively, represent the latest development in small model research. For models under 1B parameters, `OnlySportsLM` outperforms all models by a significant margin. Notably, our model gains 34.44% accuracy over Qwen2-0.5B while being 61% smaller in size. Moreover, even when comparing to models over 1B parameter, our model performs only slightly worse (-5.23%) than Qwen2-1.5B and marginally better (0.40%) than SmolLM-1.7B in average score. This is a surprising result considering our model is only 12% the size of SmolLM-1.7B.

4.2.2 Zero-shot General Benchmarks

Table 3 also presents the comparison in zero-shot commonsense reasoning benchmark between our model and the two other model collections detailed in the previous section. As expected,

[3]https://lambdalabs.com

Table 3: Performance comparison of `OnlySportsLM` against state-of-the-art models. Our model outperforms larger sub-1B models on sports tasks and competes with 1B+ models, raw scores provided in Appendix A.3

Model	#Params	OS-acc	OS-rel	OS-Avg.	ARC-e	ARC-c	PIQA	HS
number of parameters $<$ 1B								
OnlySportsLM	196M	**2.157**	**2.847**	**2.502**	37.2	23.5	59.6	37.8
SmolLM-135M	135M	1.684	1.951	1.818	43.9	-	69.9	42.3
SmolLM-360M	360M	1.705	2.027	1.866	**51.1**	-	**72.0**	**53.8**
Qwen2-0.5B	500M	1.645	2.077	1.861	39.7	**31.5**	69.3	49.3
number of parameters $\geq$ 1B								
Qwen2-1.5B	1.5B	**2.327**	**2.952**	**2.640**	48.2	**43.9**	75.5	**66.6**
SmolLM-1.7B	1.7B	2.261	2.723	2.492	**61.5**	-	**77.3**	64.1

`OnlySportsLM performs` the worst in all benchmarks, which is understandable given that it is only trained on sports-related text. For general-purpose models, we observe a positive correlation between their performance on sports domain tasks and their scores on commonsense reasoning benchmarks.

4.2.3 Performance Across Training Steps

In addition to cross-model comparison, we evaluate our model every 1000 checkpoints for `OnlySports Benchmark` and every 500 checkpoints for general benchmarks throughout the training process. This evaluation allows us to track the model's learning progression and identify any critical points or plateaus in performance

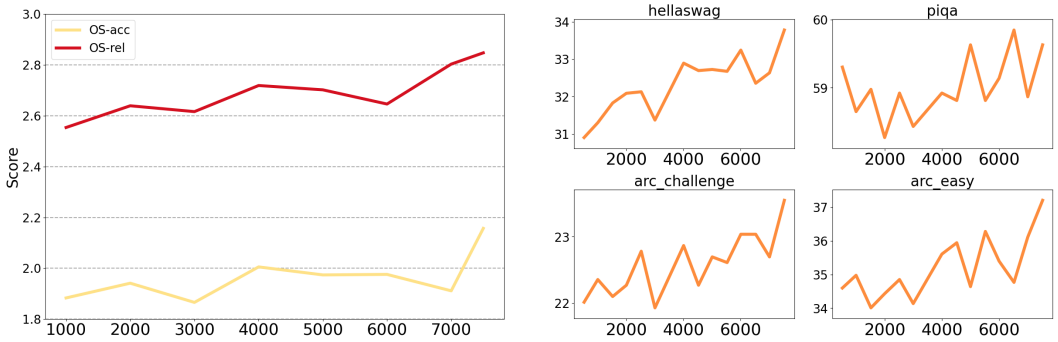

Figure 3: Evolution of `OnlySportsLM` performance across training steps. Left graph shows `OnlySports Benchmark` improving steadily. Right graphs display progress on general tasks, exhibiting upward trends despite fluctuations.

Figure 3 presents our model's performance on various Benchmarks throughout the training process. We observe a consistent improvement in both `OS-acc` and `OS-rel` scores for `OnlySports Benchmarks` as training progressed. Surprisingly, we also notice performance improvements across all general benchmarks. This unexpected trend suggests that domain-specific training on sports-related text may enhance the model's general language understanding and commonsense reasoning capabilities. While the overall trend is positive, some fluctuations in performance were observed, particularly in the general benchmarks, which could be attributed to the complexities of the training process and the diverse nature of the evaluation tasks.

4.3 Future Work

Building upon the promising results achieved with `OnlySportsLM`, future work will focus on exploring the model's full potential. We aim to complete training on the entire 600B token `OnlySports Dataset` when more funding is available, which may yield further improvements in both domain-specific and general language understanding. We also plan to explore instruction tuning techniques like instruction pre-training [6] and LAB [21] to improve performance of our model. Additionally, we plan to investigate fine-tuning approaches for `OnlySportsLM`, potentially enhancing its performance

on specific sports-related tasks. We are also interested in examining how domain improvements scale with increased model size, given that the performance of our model is comparable to other models with 1B parameters.

5 Related Work

Foundation models like GPT-4 [18] and Llama 3 [10] have demonstrated impressive performance on general-purpose language related tasks. These models are huge, with parameter ranges in hundreds of billions, and demand excessive computational resources to train. However, these general-purpose models fail to capture domain-specific nuances and context when generating content [2, 12]. Though techniques like fine-tuning [28, 19] and prompt engineering [5] can help in customization of general purpose LLMs for specific domains, the model size still remains an issue.

In parallel, efforts around developing domain-specific language models with models trained on in-domain data are also underway. Models like BloombergGPT [23] for finance, BioMedLM [4] for medical, and Galactica [22] for scientific research are LLMs trained on domain-specific data. These models also have billions of parameters and demand large-scale domain-specific dataset for training. The scale of training data and the computational cost has constrained wide-scale development of domain-specific LLMs. Further, excessive computational resources and energy requirement of such LLMs makes their deployment challenging on mobile devices thereby necessitating model compression through techniques like quantization [25] and pruning [11].

Recently [13] developed MobileLLM, a sub-billion parameter family of LLMs achieving SOTA performance on standard language benchmarks. Through model architecture search they identified that deep and thin architectures achieve better performance for compact LLMs. Within less than a month of the release of MobileLLM family, two new family of LLMs with sub-billion models, Qwen2 [26] and SmolLM [1], were introduced. SmolLM-360M is claimed to beat performance of existing models with less than 500M parameters. The performance gains in SmolLM family are attributed to training using a well curated, high quality dataset. These work, though focused on developing general-purpose models, highlight the importance of data quality and model architecture optimization in developing high performing compact LLMs. Our `OnlySports` framework incorporates these insights when developing `OnlySportsLM`.

In sports domain, most related works focus on video and image analysis. We identified only one other sports-specific language model, SportsBERT [16], a BERT-based [9] model trained on sports articles. However, SportsBERT was trained on a relatively small dataset of 8 million samples (1–2 billion tokens) and is limited to performing mask-filling tasks, preventing further evaluation in broader contexts. Additionally, we examined SportsQA [24], a multiple-choice sports comprehension benchmark. While it offers a useful benchmark for evaluating sports-specific language models, the study primarily employs general-purpose models with more than 13B parameters. We tested `OnlySportsLM` alongside other small models on this benchmark and found that such models struggle with this format, given their size constraints.

6 Limitation

While `OnlySports` collection presents promising results, we acknowledge its limitations. The primary limitation lies in the creation criteria of `OnlySports Dataset`. Due to the complicated nature of sports URLs and concerns about processing efficiency, we could not include sports beyond the mainstream ones such as lacrosse, yoga, and archery, and their representations in general sports websites such as ESPN can be limited. We are dedicated to discovering better low-cost extraction techniques for domain-specific texts and planning to expand `OnlySports Dataset` when more resources are available.

Another limitation is the strategy used in `OnlySports Benchmark`. A primary limitation in our benchmarking approach relies on LLMs as both content generators and evaluators. This dual role introduces the potential for error propagation, where biases or inaccuracies in generated outputs may influence evaluation judgments, possibly inflating model performance or concealing deficiencies. This concern is particularly relevant in the sports domain, where accurate handling of statistical data, historical records, and specialized terminology is essential. Although we have implemented a multi-model consensus to minimize bias, integrating human evaluators would enhance the reliability

of accuracy claims. However, employing human evaluators poses significant challenges, especially for under-resourced researchers, given the substantial time, cost, and domain-specific expertise required.

7 Conclusion

This study focuses on optimizing sports domain language models with sub-billion parameters. Our findings demonstrate that for sports-related tasks, a carefully designed small model can outperform larger general-purpose models. By leveraging `OnlySports Dataset` and a carefully designed model architecture, we achieved significant improvements in sports knowledge generation and understanding. Our `OnlySportsLM`, a 196M parameter model, exhibits substantial advancements in sports-related text generation compared to previous state-of-the-art methods. The model's performance on `OnlySports Benchmark` underscores its effectiveness in continuing sports-related text. Furthermore, we demonstrate the potential of our approach in creating high-quality, domain-specific large datasets and evaluation methods. The `OnlySports Dataset` and `Benchmark` can provide valuable resources for future research in sports-related NLP tasks. Our study contributes to the ongoing research in developing efficient, domain-specific language models. While our approach shows promise in the sports domain, further investigation is needed to determine its adaptability to other specialized fields. We believe this work may offer insights that could be valuable for researchers exploring resource-efficient AI solutions across various domains.

References

[1] Loubna Ben Allal, Anton Lozhkov, and Elie Bakouch. Smollm - blazingly fast and remarkably powerful. Hugging Face Blog, 2024. Published July 16, 2024.

[2] Joshua Au Yeung, Zeljko Kraljevic, Akish Luintel, Alfred Balston, Esther Idowu, Richard J. Dobson, and James T. Teo. Ai chatbots not yet ready for clinical use. *Frontiers in Digital Health*, 5, 2023.

[3] Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. Piqa: Reasoning about physical commonsense in natural language, 2019.

[4] Elliot Bolton, Abhinav Venigalla, Michihiro Yasunaga, David Hall, Betty Xiong, Tony Lee, Roxana Daneshjou, Jonathan Frankle, Percy Liang, Michael Carbin, and Christopher D. Manning. Biomedlm: A 2.7b parameter language model trained on biomedical text, 2024.

[5] Banghao Chen, Zhaofeng Zhang, Nicolas Langrené, and Shengxin Zhu. Unleashing the potential of prompt engineering in large language models: a comprehensive review, 2024.

[6] Daixuan Cheng, Yuxian Gu, Shaohan Huang, Junyu Bi, Minlie Huang, and Furu Wei. Instruction pre-training: Language models are supervised multitask learners, 2024.

[7] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge, 2018.

[8] Jeffrey Dean and Sanjay Ghemawat. Mapreduce: simplified data processing on large clusters. *Commun. ACM*, 51(1):107–113, jan 2008.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[10] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily

Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko,

Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vítor Albiero, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. The llama 3 herd of models, 2024.

[11] Elias Frantar and Dan Alistarh. Sparsegpt: massive language models can be accurately pruned in one-shot. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

[12] Yalan Lin, Meng Chen, Yuhan Hu, Hongyu Zhang, Chengcheng Wan, Zhao Wei, Yong Xu, Juhong Wang, and Xiaodong Gu. On the effectiveness of large language models in domain-specific code generation, 2024.

[13] Zechun Liu, Changsheng Zhao, Forrest Iandola, Chen Lai, Yuandong Tian, Igor Fedorov, Yunyang Xiong, Ernie Chang, Yangyang Shi, Raghuraman Krishnamoorthi, Liangzhen Lai, and Vikas Chandra. Mobilellm: Optimizing sub-billion parameter language models for on-device use cases, 2024.

[14] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019.

[15] Luke Merrick. Embedding and clustering your data can improve contrastive pretraining, 2024.

[16] Microsoft. "sportsbert". `https://huggingface.co/microsoft/SportsBERT`. Accessed: 2024-08-27.

[17] Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. Large language models: A survey, 2024.

[18] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke,

Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2024.

[19] Guilherme Penedo, Hynek Kydlíček, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. The fineweb datasets: Decanting the web for the finest text data at scale, 2024.

[20] Bo Peng, Daniel Goldstein, Quentin Anthony, Alon Albalak, Eric Alcaide, Stella Biderman, Eugene Cheah, Xingjian Du, Teddy Ferdinan, Haowen Hou, Przemysław Kazienko, Kranthi Kiran GV, Jan Kocoń, Bartłomiej Koptyra, Satyapriya Krishna, Ronald McClelland Jr. au2, Niklas Muennighoff, Fares Obeid, Atsushi Saito, Guangyu Song, Haoqin Tu, Stanisław Woźniak, Ruichong Zhang, Bingchen Zhao, Qihang Zhao, Peng Zhou, Jian Zhu, and Rui-Jie Zhu. Eagle and finch: Rwkv with matrix-valued states and dynamic recurrence, 2024.

[21] Shivchander Sudalairaj, Abhishek Bhandwaldar, Aldo Pareja, Kai Xu, David D. Cox, and Akash Srivastava. Lab: Large-scale alignment for chatbots, 2024.

[22] Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. Galactica: A large language model for science, 2022.

[23] Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. Bloomberggpt: A large language model for finance, 2023.

[24] Haotian Xia, Zhengbang Yang, Yuqing Wang, Rhys Tracy, Yun Zhao, Dongdong Huang, Zezhi Chen, Yan Zhu, Yuan fang Wang, and Weining Shen. Sportqa: A benchmark for sports understanding in large language models, 2024.

[25] Guangxuan Xiao, Ji Lin, Mickael Seznec, Hao Wu, Julien Demouth, and Song Han. Smoothquant: accurate and efficient post-training quantization for large language models. In

Proceedings of the 40th International Conference on Machine Learning, ICML'23. JMLR.org, 2023.

[26] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. Qwen2 technical report, 2024.

[27] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence?, 2019.

[28] Zheng Zhang, Chen Zheng, Da Tang, Ke Sun, Yukun Ma, Yingtong Bu, Xun Zhou, and Liang Zhao. Balancing specialized and general skills in llms: The impact of modern tuning and data strategy, 2023.

[29] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, 2024.

A Appendix

A.1 Evaluation Criteria

This appendix provides detailed grading rubrics for the two main evaluation criteria used in the `OnlySports Benchmark`: Accuracy and Factuality (OS-acc), and Continuity and Relevancy (OS-rel). These rubrics were provided to the GPT-4 and Claude 3.5 Sonnet models as part of their system messages when acting as evaluators. *{num}* specified the number of responses that will be in the prompt for evaluation.

A.1.1 Accuracy and Factuality

Prompt for evaluating accuracy and factuality:

You are a sports expert assigned to grade language models' generation performance on general sports-related text according to the provided rubric. 1 prompt and {num} responses will be presented, all attempting to complete the same given prompt. Each response is separated by [SEP] and limited to 80 tokens.

Evaluate responses using the following rubric for "Accuracy and Factuality":
"1": "Mostly inaccurate, significant factual errors.",
"2": "Partially accurate, mix of correct and incorrect information.",
"3": "Mostly accurate, minor factual errors.",
"4": "Highly accurate, negligible errors.",
"5": "Fully accurate and factually impeccable."

When evaluating, only consider the accuracy and factuality in the context of the given prompt. Score these generated responses on a scale of 1-5. Only output the scores! Output scores in the following format: 'X, X, X, X, X...', where X is a number between 1 and 5.

A.1.2 Continuity and Relevance

Prompt for evaluating continuity and relevance:

You are a sports expert assigned to grade language models' generation performance on general sports-related text according to the provided rubric. 1 prompt and {num} responses will be presented,

all attempting to complete the same given prompt. Each response is started by [SEP] and limited to 80 tokens.

Evaluate responses using the following rubric for "Continuity and Relevance":
"1": "Poor continuation, diverges significantly from the prompt's context or topic.",
"2": "Weak continuation, maintains some elements of the prompt but introduces unrelated content.",
"3": "Adequate continuation, generally follows the prompt's direction with some minor deviations.",
"4": "Strong continuation, closely follows the prompt's context and style with minimal inconsistencies.",
"5": "Excellent continuation, seamlessly extends the prompt's narrative, context, and style."

When evaluating, only consider how well the generated text continues or extends the given prompt in terms of context, topic, and style. Score these generated responses on a scale of 1-5. Only output the scores! Output scores in the following format: 'X, X, X, X, X...', where X is a number between 1 and 5.

A.2 Example Prompts and Responses for OnlySports Benchmark Scoring

This appendix provides concrete examples of prompts and `OnlySportsLM` responses to illustrate the scoring criteria used in the `OnlySports Benchmark`. We present four distinct scenarios that demonstrate various combinations of Accuracy and Factuality (OS-acc) and Continuity and Relevancy (OS-rel) scores:

1. High Accuracy, High Relevancy 2. High Accuracy, Low Relevancy 3. Low Accuracy, High Relevancy 4. Low Accuracy, Low Relevancy

For each scenario, we provide the original prompt, response, and a brief explanation of the scoring rationale. These examples are designed to offer clarity on how the evaluation criteria were applied and to aid in the reproducibility of our results.

1. Prompt: *Named after the famed coach, the Vince Lombardi Trophy is awarded to the team that wins*

Response: *the Super Bowl. The trophy is awarded to the team that wins the Super Bowl, and the trophy is named after the coach who won the most Super Bowls. The trophy is named after Vince Lombardi, who won the Super Bowl with the Green Bay Packers in 1967. The*

This response accurately states that the Vince Lombardi Trophy is awarded to the Super Bowl winner. It correctly mentions Vince Lombardi and his connection to the Green Bay Packers. The information is directly relevant to the prompt and provides additional context about the trophy's history.

2. Prompt: *With flippers and a flutter kick, scuba divers use a modified version of the freestyle stroke to*

Response: *keep their hands and feet on the ground. The dives are performed in a single-piece, high-tech suit. The diver is required to wear a mask and a cap. In the 2012 London Olympics, the dives were judged by a panel of experts. In 2013, the International Olympic Committee (IOC) banned the dives.*

While the response contains accurate information about diving suits and Olympic judging, it's not relevant to the prompt about scuba diving techniques. The mention of flippers and freestyle stroke is ignored, making the response off-topic despite its factual correctness.

3. Prompt: *"As the Lakers head toward another challenging season, they are hopeful about LeBron James's"*

Response:*' decision to return to the team. James, who will be a free agent this summer, has not made a decision on his future."I'm just happy that he's back," Lakers coach Frank Vogel said. "I think he's a big part of our team. He's a big part of our team. I think he's going to be a big part of our team'*

The response is relevant to the Lakers and LeBron James's future with the team, matching the prompt's topic. However, it contains inaccuracies: LeBron James was not a free agent at the time specified, and Frank Vogel was no longer the Lakers' coach. The content is on-topic but factually incorrect.

4. Prompt: *Hockey legend Wayne Gretzky holds the NHL record for the most points scored in a single season at*

Response: *100 points. NHL players are not the only ones who have the ability to score goals. Some of the most popular sports in the world include soccer, basketball, baseball, football, and hockey. In addition to the games themselves, the sport of hockey is also a popular pastime for many people. There are many ways to play hockey, and the rules of the game*

This response is both inaccurate and irrelevant. The stated record of 100 points is incorrect (Gretzky's record is 215 points). Most of the response diverges into general sports discussion unrelated to Gretzky's NHL record, making it largely irrelevant to the prompt.

A.3 Model Scores from Claude and GPT

In this section, we present the raw score from the two state-of-the-art LLMs: Claude 3.5 and GPT-4o. The table below (Table 4) showcases the scores for each model across OS-rel (Relevance) and OS-acc (Accuracy). These scores provide insight into how well each model performs in generating sports-related content, with higher scores indicating better performance. We observe that Claude 3.5 Sonnet generally gives higher scores than GPT4o, using the same prompt.

Table 4: Performance scores for different language models across two evaluators

Model	**Claude 3.5**		**GPT-4o**	
	OS-rel	**OS-acc**	**OS-rel**	**OS-acc**
OnlySportsLM	3.19	2.38	2.50	1.94
Qwen2-0.5B	2.34	1.93	1.82	1.36
Qwen2-1.5B	3.23	2.73	2.68	1.93
SmolLM-135M	2.25	1.96	1.66	1.41
SmolLM-360M	2.23	1.91	1.82	1.50
SmolLM-1.7B	2.97	2.55	2.48	1.97

A.4 Social Impact

Our work on `OnlySportsLM` has potential for both positive and negative societal impacts. On the positive side, a more efficient, domain-specific language model for sports could democratize access to sports information and analysis, enhancing fan engagement and potentially supporting smaller sports organizations with limited resources. It could also aid in sports journalism, making it easier to generate accurate, timely reports on sporting events. However, we acknowledge potential negative impacts as well. The model could be misused to generate false or misleading sports news, potentially spreading misinformation or manipulating betting markets. There's also a risk of perpetuating biases present in sports reporting, potentially reinforcing stereotypes or unfair representations of athletes or teams. Privacy concerns arise if the model is used to generate detailed profiles of athletes based on publicly available data.