Scaling Multimodal Search and Recommendation with Small Language Models via Upside-Down Reinforcement Learning

Yu-Chen Lin, Adobe
Sanat Sharma, Meta
Hari Manikandan, Adobe
Jayant Kumar, Adobe
Tracy Holloway King, Adobe
Jing Zheng, Adobe

***Abstract*—In this work, we investigate how small language models (SLMs) can be scaled to support multimodal search and recommendation use cases while remaining efficient enough for real-time, resource-constrained deployments. We present a framework that combines upside-down reinforcement learning with synthetic data distillation from a large language model (Llama-3 [1]) to train a 100M-parameter GPT-2 model [2] for multitask prompt generation. Despite being up to 80 times smaller than state-of-the-art large language models (LLMs), our SLM achieves relevance and diversity scores within 6% of competitive baselines such as Llama-3 8B, Qwen3 8B, and Ministral 8B. These results demonstrate that SLMs can effectively handle multimodal search and recommendation tasks, while dramatically reducing inference latency and memory overhead. Our study highlights the potential of lightweight models as practical engines for scalable multimodal discovery, bridging the gap between cutting-edge research and real-world multimodal applications such as media recommendations and creative content generation.**

***Index Terms*—small language models, multimodal search and recommendation, upside-down reinforcement learning**

I. Introduction

Multimodal search and recommendation has emerged as a central capability for applications such as e-commerce, media platforms, and interactive agents. With the revolution of large language models (LLMs), more powerful multimodal use cases have become possible, yet these gains come with substantial computational and memory costs, especially at inference time. These limitations hinder their deployment in high-frequency, real-time applications where efficiency is critical. To address this challenge, we propose a small language model (SLM) framework that serves as an efficient and effective multitask learner for multimodal prompt generation, leveraging upside-down reinforcement learning (UDRL) to control generation.

Fig. 1 provides an overview of our approach. Our multimodal system accepts both image and text inputs, which are processed by an in-house detector for intent understanding before passing to a prompt generator. The generated prompts, in turn, enable multiple downstream use cases. For example, they can power contextual search, content generation suggestions, and multimodal recommendations by producing relevant text or image outputs aligned with user intent. To train the prompt generator, we employ a compact SLM. Synthetic data is first generated from a large language model (Llama-3) using prompt engineering and few-shot examples, and then distilled into the SLM via UDRL [3], [4]. This produces a specialized model optimized for low latency, memory efficiency, targeted multitasking, and real-time multimodal generation.

Our approach is built on three core contributions:

- SLMs as Efficient Multitask Learners: We demonstrate that SLMs, trained with targeted synthetic data, can support diverse multimodal discovery tasks while bridging the gap between LLMs and deployable engines.
- A scalable training pipeline with UDRL: We introduce a novel pipeline that distills supervision from LLMs and applies UDRL to optimize SLMs for prompt generation across multiple tasks, achieving competitive relevance and diversity.
- Synthetic Dataset Distillation: Leveraging LLM-generated data, we curate high-quality training examples that enable SLMs to learn effectively from minimal resources.

Through systematic experiments, we show that our SLM can achieve performance comparable to much larger LLMs with an 80-fold reduction in model size and delivering substantial gains in latency and memory efficiency. This makes SLMs suitable for real-time multimodal discovery applications, where responsiveness and scalability are essential. Moreover, our framework is highly adaptable and can be integrated with commercial text-to-image generation systems (e.g., Adobe Firefly), enabling practical deployment in diverse real-world scenarios.

II. Methodology

As illustrated in Fig. 1, our task involves generating prompts for generative models based on multimodal inputs. The process begins with detecting user intents using an in-house intent detector. Once the intents are identified, we aim to generate prompts accordingly. Rather than relying solely on a large language model (LLM) or a multimodal LLM, we adopt a more efficient approach.

Specifically, our methodology focuses on training a small language model (SLM) for prompt generation with intents that emphasizes high efficiency and multitask capabilities. We achieve this by leveraging synthetic data distillation from a large language model (LLM) combined with upside-down reinforcement learning.

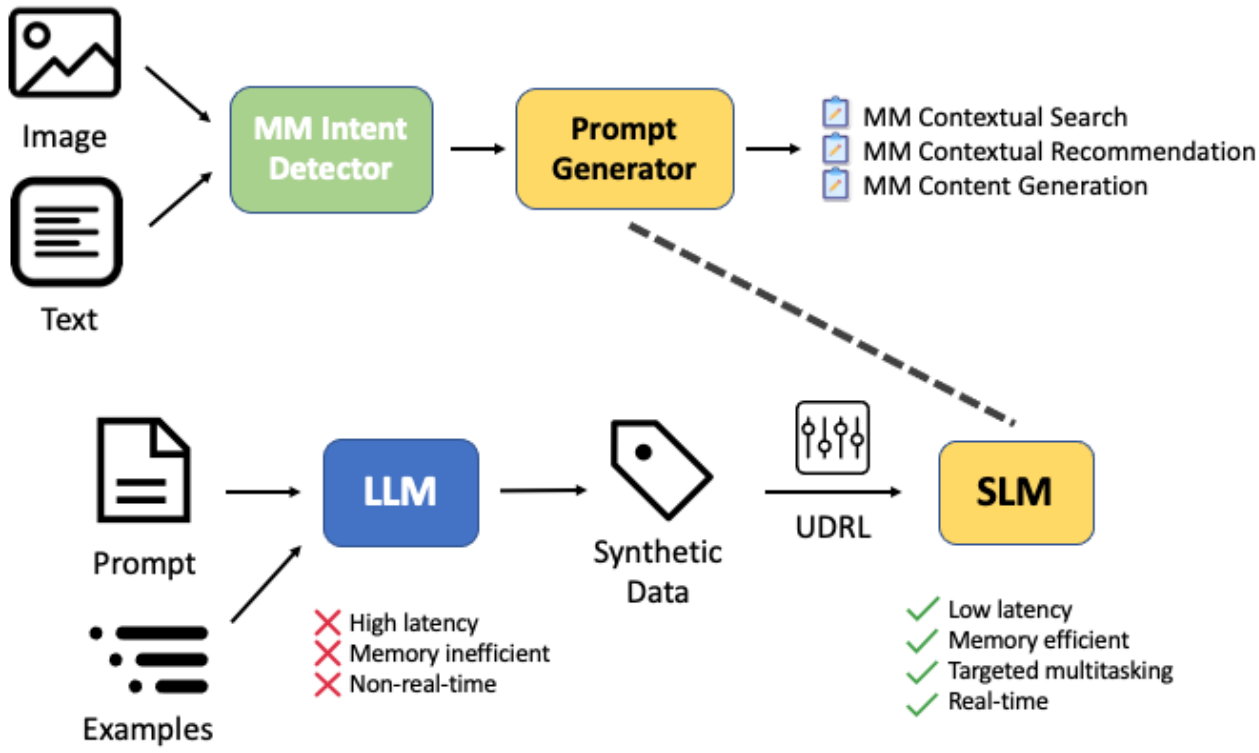

Fig. 1. The top workstream illustrates our multimodal system with intent detection and prompt generation for contextual search, recommendation, and content generation. The bottom workstream shows our pipeline for training a small language model (SLM) via synthetic data distillation and upside-down reinforcement learning, enabling an efficient real-time alternative to large language models.

A. Synthetic Dataset Distillation

To enable the SLM to capture complex task knowledge from a larger model, we first create a synthetic dataset using Llama-3. This involves generating high-quality, task-specific data that allows the SLM to learn from the representations of LLM effectively. We utilize vLLM [5] to parallelize our dataset generation process.

1) Dataset Curation Process:

- **Intent and Prompt Pair Generation:** We curated a dataset of 52 million intent-prompt pairs by prompting the LLM with various intent descriptions and collecting its responses. Each sample consists of an intent description paired with a generated prompt, tailored to a range of real-world tasks. For example, intents such as "birthday celebration," "holiday sale," or "product launch" were used to generate prompts that align with these contexts.
- **Structured Prompting for Targeted Tasks:** To ensure diversity and relevance in generated prompts, we used structured prompting with few-shot examples and task specifications. Each prompt to Llama-3 included high-level task descriptions and few-shot examples to encourage contextually appropriate and diverse outputs. Different prompting strategies are used to address the specific requirements of different tasks. By specifying factors like tone, style, and context, we created a dataset that is comprehensive and task-aligned.
- **Labeling and Intent Selection:** The intent descriptions were sourced from in-house creative knowledge graph and augmented with diverse asset metadata from internal images and templates. This process provided a broad set of intents that accurately represent real-world use cases. Each intent was paired with multiple generated prompts, thus ensuring a rich dataset of prompt variations for each concept.

This curated dataset of intent-prompt pairs serves as a distilled form of knowledge from Llama-3, allowing the SLM to learn diverse language patterns and contextual nuances without directly training on a massive language model. Examples of intents to generated prompts are shown in Table I.

B. Upside-Down Reinforcement Learning for SLM Optimization

To optimize the SLM's generation quality and control specific attributes like length and relevance, we utilize upside-down reinforcement learning (UDRL) [3], [4]. This approach allows the model to learn specific objectives based on desired outcomes rather than traditional reward structures.

1) Reward-Based Prompt Generation: Upside-down reinforcement learning frames the prompt generation task as an optimization problem where the SLM aims to achieve target specifications for each generated output. This process is as follows:

TABLE I
EXAMPLES OF INTENTS AND GENERATED PROMPTS FOR IMAGE AND TEMPLATE GENERATION, INCLUDING FORMAT SPECIFICATIONS. HAVING TASK, LENGTH, AND INTENT TAGS HELPS GUIDE THE GENERATIONS.

Intents / Task	Generated Prompt	Training Data Format
Topic: birthday Scene object: balloon Task: prompt for text-to-image generation	Whimsical birthday celebration featuring giant balloons in fun shapes and sizes, tied to a birthday child's arm or wrist.	`<|19|> <|intent|> Topic: birthday, Scene object: balloon <|IP|> whimsical birthday celebration featuring giant balloons in fun shapes and sizes, tied to a birthday child's arm or wrist.`
Topic: birthday party Design type: invitation Task: prompt for text-to-template generation	Create a whimsical birthday party invitation template with balloons, confetti, and a playful theme.	`<|14|> <|intent|> Topic: birthday party, Design Type: invitation <|TP|> create a whimsical birthday party invitation template with balloons, confetti, and a playful theme.`

- **Controlled-Length Generation:** The SLM is trained to produce prompts within a desired length range (e.g., 10 to 35 words). Tokens indicating target lengths are incorporated into the input, guiding the model towards generating responses that match the specified length with a mean squared error consistently under two words. More evaluations are in Section III.
- **Modality-Agnostic Prompting:** We trained the SLM to handle both text-to-image (T2I) and text-to-template (T2T) prompts within the same framework. This was achieved by adding modality tokens to each training instance, allowing the model to distinguish between generation tasks and tailor its output accordingly.
- **Contextual Relevance and Specificity:** By assigning relevance scores to generated prompts based on a predefined metric (e.g., alignment with target intent or clarity), the model learns to prioritize contextually relevant and specific responses. During training, prompts that meet these objectives are positively reinforced, improving the SLM's ability to generate accurate and contextually appropriate prompts. While we do not implement this aspect in our current pipeline, it could be valuable for other applications.

C. Model Architecture and Key Capabilities

Our SLM is based on nanoGPT[1], a compact variant of GPT-2, with 104 million parameters, configured with 12 layers, 12 attention heads, and an embedding dimension of 768. The model architecture and training setup are designed to maximize efficiency and multitask performance.

1) Model Specifications:

- **Parameter Efficiency:** The SLM contains approximately 1/80th the parameters of Llama-3 8b and similar state-of-the-art LLMs, making it computationally efficient and suitable for deployment on standard hardware.
- **Inference Speed:** Our SLM achieves a processing speed of up to 338 tokens per second on a single A10G GPU (non-batched, non-quantized, or accelerated), making it suitable for real-time applications. This performance is especially advantageous in resource-constrained environments where inference latency is a critical factor.
- **Multitask Learning Capabilities:** The SLM is trained to handle both T2I and T2T prompts, making it a versatile tool for multimodal applications. Through the integration of task-specific tokens, the model can generate contextually accurate prompts tailored to the input task, whether it's for text-to-image generation or template-based design.

2) Training with UDRL: The training data are formatted as `<|# words of the prompt|> <|intent|> INTENT <|prompt for T2I (IP) or T2T (TP)|> PROMPT`. As outlined in Section II-B, we combine the word count and modality tokens to create a single training instance; refer to Table I for examples.

We utilize a vocabulary set with legal approval, ensuring it excludes any offensive, discriminatory, or biased language. We then train a BPE tokenizer with 25,600 tokens from scratch using our curated dataset. Subsequently, we train the nanoGPT model using next-token prediction, employing four A10G GPUs over 10 days, completing approximately 300,000 iterations with a batch size of 128 and a learning rate of 6e-4.

By combining this training approach with upside-down reinforcement learning, we ensure that our SLM can effectively manage length control and multitasking. During inference, we can easily control output length by specifying a length token and define the desired task using the modality token. Please refer to Section III for more evaluations of these abilities.

III. EXPERIMENTS AND EVALUATIONS

We focus on both qualitative and quantitative evaluations to judge the quality of our model.

A. Quantitative Evaluation

The quantitative evaluation is divided into two primary areas: relevance evaluation, which assesses how well the generated prompts align with the specified queries and modality requirements, and task adherence evaluation, which measures the SLM's accuracy in meeting specific prompt length requirements.

1) Relevance Evaluation: To evaluate relevance, we conducted experiments to measure how accurately the SLM-generated prompts aligned with the specified queries and task requirements. Relevance was assessed using an automatic method with LLM-as-a-judge, where GPT-4o-mini [11] served as the evaluator. Each generated prompt was rated on a scale from 1 to 10, where higher scores indicate stronger alignment

[1]https://github.com/karpathy/nanoGPT

TABLE II

RELEVANCE SCORES FOR PROMPT GENERATION TASKS, EVALUATED ON TEXT-TO-IMAGE AND TEXT-TO-TEMPLATE PROMPTS UNDER BOTH FEW-SHOT AND ZERO-SHOT PROMPTING. OUR SLM MODEL, WHICH OPERATES IN A ZERO-SHOT SETTING, DEMONSTRATES COMPETITIVE PERFORMANCE WITH MUCH LARGER MODELS. WITH THE UDRL TRAINING PARADIGM, SLM ACHIEVES EFFICIENCY SUITABLE FOR ENTERPRISE USE, OUTPERFORMING OR MATCHING MODELS WITH SIGNIFICANTLY HIGHER PARAMETER COUNTS.

Model	**Few-Shot**		**Zero-Shot**		**# Tokens Per Second**		**# Params**
	text-to-image	**text-to-template**	**text-to-image**	**text-to-template**	**Rate**	**Memory Used**	
SLM (Ours)	N/A	N/A	8.31	8.45	**353**	**∼500MB**	**104M**
Llama-3 8B [1]	8.63	8.56	8.36	8.33	80	∼80GB (with vLLM)	8.0B
Ministral 8B [6]	8.55	8.40	7.87	7.18	78		8.0B
Gemma-3 4B [7]	8.66	8.41	8.51	**8.49**	100		4.3B
SmolLM3 3B [8]	8.70	**8.68**	8.19	8.35	121		3.1B
Llama-3.2 3B [1]	8.65	8.60	8.40	8.44	142		3.2B
Phi-4-mini [9]	8.59	8.45	8.27	8.15	118		3.8B
Qwen3 8B [10]	**8.80**	8.63	**8.52**	8.37	76		8.2B

with the target query. We provided several examples and criterion to the LLM judge prior to the evaluation. Some of these include

- Correctness - does the prompt contain grammatical or semantic errors.
- Clarity - is the prompt clear to understand, is the grammar structure sound.
- Completeness - does the prompt utilize all of the provided query context.
- Usefulness - is the prompt generated useful for the task provided.

For comparison, we evaluated our SLM with state-of-the-art LLMs ranging in size from 3B to 8B. Table II displays the relevance scores for both text-to-image (T2I) and text-to-template (T2T) prompt generation tasks. We investigated the performance of these LLMs under both few-shot and zero-shot prompting. For our trained SLM, there is no need for demonstration, which places it in the zero-shot setting. We observe the following.

1) **Effective Distillation:** In the few-shot setting, our SLM achieves relevance scores nearly on par with the teacher model, Llama-3 8B, with only a minor performance gap of approximately 3%. In the zero-shot setting, the performance of our SLM matches Llama-3 8B, and it even outperforms Llama-3 8B in T2T prompt generation. This illustrates the success of our distillation process, effectively transferring high-performance prompt generation capabilities to a more efficient model, without the need for task-specific demonstrations.
2) **Competitive Performance with Minimal Parameters:** Despite being over an order of magnitude smaller, our SLM competes closely with models between 3B–8B parameters. On T2I zero-shot, SLM scores 8.31, only slightly behind the strongest model (Qwen-3 8B, 8.52), while in T2T zero-shot it ranks second overall (8.45). Importantly, these results are achieved with just 104M parameters—two orders of magnitude fewer than most baselines—demonstrating a remarkable balance of compactness and accuracy.
3) **Efficiency and Deployment Readiness:** SLM is highly efficient in both speed and memory. Running on a single A100 80GB, it requires only ∼500MB of memory while generating at 353 tokens/second. In contrast, all baseline LLMs evaluated under the vLLM runtime require ∼80GB of GPU memory and achieve significantly slower generation rates (76–142 tokens/second). This means SLM is nearly 3–5× faster while being vastly more memory-efficient. Even compared to the smallest baseline (Llama-3.2 3B), SLM delivers superior or comparable prompt relevance with a model size that is over 30× smaller.

TABLE III

TASK ADHERENCE RESULTS FOR SPECIFIED PROMPT LENGTHS. MSE MEASURES THE DEVIATION FROM THE TARGET LENGTH, AND THE FINAL ROW SHOWS THE PERCENTAGE OF PROMPTS WITHIN ±2 WORDS OF THE TARGET LENGTH.

Target Length (words)	10	20	30	35
Mean Length (words)	10.3	19.2	31.1	33.8
Mean Square Error	0.365	0.881	1.131	1.179
% Prompts Within ±2 Words	98%	96%	93%	95%

In summary, our SLM delivers contextually accurate prompt generation while being substantially more efficient and lightweight than existing LLMs. Its strong performance across both T2I and T2T tasks, combined with extreme efficiency in speed and memory, makes it a compelling solution for real-world deployment, especially in resource-constrained environments for multimodal discovery applications.

2) Task Adherence Evaluation (Length): In addition to relevance, we evaluated the SLM's ability to meet target length requirements for generated prompts. The model was trained to produce prompts within specified lengths, typically in the range of 10 to 35 words. Length adherence was measured by calculating the mean squared error (MSE) between the generated prompt length and the target length, and by recording the percentage of prompts that fell within an acceptable range (±2 words from the target length). The evaluation results are in Table III.

The SLM achieves precise control over prompt length,

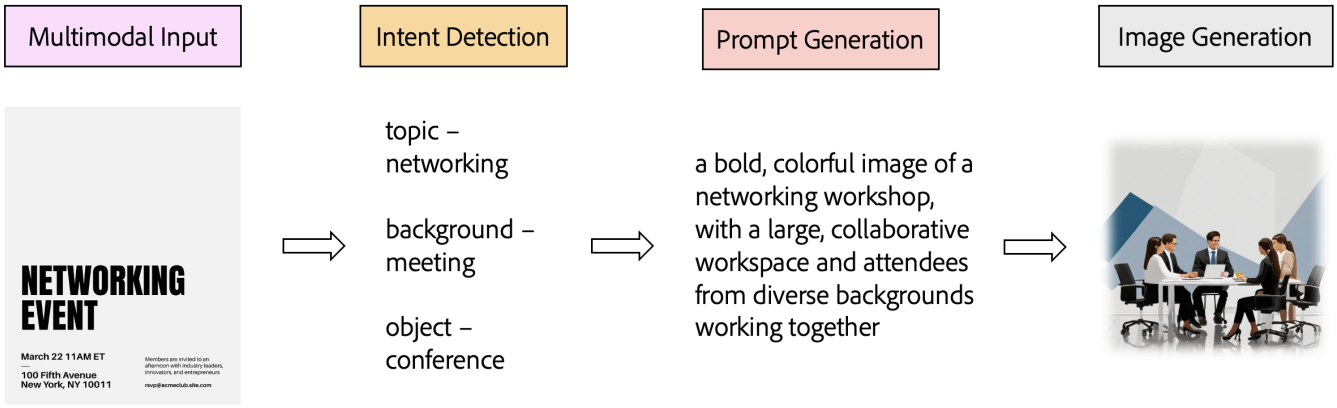

Fig. 2. A sample pipeline for multimodal contextual recommendation incorporating our SLM model for prompt generation.

with MSE consistently around 1. These results indicate that the model can reliably generate prompts of desired lengths, which is essential for applications requiring specific or concise outputs. This ability provides significant flexibility across diverse multimodal use cases.

B. Qualitative Evaluation

To complement the quantitative results, we conducted qualitative evaluations with human reviewers, tailored to real-world multimodal use cases.

1) Prompt Quality: We chose 18 human reviewers with experience in writing prompts for image and template generative models. The reviewers assessed the SLM-generated prompts on two criteria:

- Relevance - how closely does the prompt align with the query.
- Correctness - is the prompt clear and easy to understand, with correct grammar and sound structure.

Each criterion was rated on a scale from 1 to 3 (corresponding to not, somewhat, very, e.g., not relevant, somewhat relevant and very relevant respectively), and reviewers provided additional qualitative feedback on the prompts.

In total, we collected 180 human evaluations. The human reviewers scored the prompts with an average relevance rating of **87%**, highlighting the model's ability to generate prompts that align well with the provided queries and the task specification. For accuracy and correctness, the SLM received a score of **96%** somewhat or very correct, indicating that the prompts were contextually accurate, grammatically correct and generally maintained coherence to the topic. We emphasize again that the prompt, which incorporates core user intents, can be leveraged to support a wide range of multimodal tasks, including contextual search, prompt suggestion, generative autocomplete, and creative content generation.

2) Multimodal Contextual Recommendation: We evaluated the end-to-end experience for multimodal contextual recommendation, as illustrated in Fig. 2. Suppose a user is editing a canvas containing some pre-specified text; in this case, we can incorporate those multimodal inputs for intent understanding and prompt generation. The generated prompt can then be used to produce images that the user can integrate into the canvas.

For this evaluation, we recruited 17 evaluators with prior experience in editing templates. We collected 210 evaluations, asking the evaluators to rank the work-in-progress canvas alongside the generated elements. The results indicate that **86%** of evaluators were willing to incorporate the generated images into the specified canvas, demonstrating strong user acceptance.

Overall, the qualitative feedback aligns with our quantitative findings, showing that the SLM not only achieves high relevance and adherence to content length but also produces contextually rich prompts. This analysis highlights the practical applicability of SLMs for multimodal tasks, especially in scenarios where low latency and computational efficiency are essential.

IV. Conclusion

This work demonstrates the potential of small language models (SLMs) as efficient and versatile multitask learners. By integrating upside-down reinforcement learning with supervised training, we show that an SLM can achieve performance competitive with much larger models, such as Llama-3, while reducing model size by up to 80-fold for targeted tasks. Our evaluations indicate that SLMs can preserve high contextual relevance, precision, task adherence, and precise length control, even in resource-constrained environments.

The substantial gains in efficiency and low-latency inference highlight the practicality of SLMs for real-world deployment, particularly in enterprise scenarios where computational resources and response speed are critical. These findings underscore the promise of lightweight models as scalable engines for multimodal search, recommendation, and creative content generation.

REFERENCES

[1] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan, A. Goyal, A. Hartshorn, A. Yang, A. Mitra, A. Sravankumar, A. Korenev, A. Hinsvark, A. Rao, A. Zhang, A. Rodriguez, A. Gregerson, A. Spataru, B. Roziere, B. Biron, B. Tang, B. Chern, C. Caucheteux, C. Nayak, C. Bi, C. Marra, C. McConnell, C. Keller, C. Touret, C. Wu, C. Wong, C. C. Ferrer, C. Nikolaidis, D. Allonsius, D. Song, D. Pintz, D. Livshits, D. Esiobu, D. Choudhary, D. Mahajan, D. Garcia-Olano, D. Perino, D. Hupkes, E. Lakomkin, E. AlBadawy, E. Lobanova, E. Dinan, E. M. Smith, F. Radenovic, F. Zhang, G. Synnaeve, G. Lee, G. L. Anderson, G. Nail, G. Mialon, G. Pang, G. Cucurell, H. Nguyen, H. Korevaar, H. Xu, H. Touvron, I. Zarov, I. A. Ibarra, I. Kloumann, I. Misra, I. Evtimov, J. Copet, J. Lee, J. Geffert, J. Vranes, J. Park, J. Mahadeokar, J. Shah, J. van der Linde, J. Billock, J. Hong, J. Lee, J. Fu, J. Chi, J. Huang, J. Liu, J. Wang, J. Yu, J. Bitton, J. Spisak, J. Park, J. Rocca, J. Johnstun, J. Saxe, J. Jia, K. V. Alwala, K. Upasani, K. Plawiak, K. Li, K. Heafield, K. Stone, K. El-Arini, K. Iyer, K. Malik, K. Chiu, K. Bhalla, L. Rantala-Yeary, L. van der Maaten, L. Chen, L. Tan, L. Jenkins, L. Martin, L. Madaan, L. Malo, L. Blecher, L. Landzaat, L. de Oliveira, M. Muzzi, M. Pasupuleti, M. Singh, M. Paluri, M. Kardas, M. Oldham, M. Rita, M. Pavlova, M. Kambadur, M. Lewis, M. Si, M. K. Singh, M. Hassan, N. Goyal, N. Torabi, N. Bashlykov, N. Bogoychev, N. Chatterji, O. Duchenne, O. Çelebi, P. Alrassy, P. Zhang, P. Li, P. Vasic, P. Weng, P. Bhargava, P. Dubal, P. Krishnan, P. S. Koura, P. Xu, Q. He, Q. Dong, R. Srinivasan, R. Ganapathy, R. Calderer, R. S. Cabral, R. Stojnic, R. Raileanu, R. Girdhar, R. Patel, R. Sauvestre, R. Polidoro, R. Sumbaly, R. Taylor, R. Silva, R. Hou, R. Wang, S. Hosseini, S. Chennabasappa, S. Singh, S. Bell, S. S. Kim, S. Edunov, S. Nie, S. Narang, S. Raparthy, S. Shen, S. Wan, S. Bhosale, S. Zhang, S. Vandenhende, S. Batra, S. Whitman, S. Sootla, S. Collot, S. Gururangan, S. Borodinsky, T. Herman, T. Fowler, T. Sheasha, T. Georgiou, T. Scialom, T. Speckbacher, T. Mihaylov, T. Xiao, U. Karn, V. Goswami, V. Gupta, V. Ramanathan, V. Kerkez, V. Gonguet, V. Do, V. Vogeti, V. Petrovic, W. Chu, W. Xiong, W. Fu, W. Meers, X. Martinet, X. Wang, X. E. Tan, X. Xie, X. Jia, X. Wang, Y. Goldschlag, Y. Gaur, Y. Babaei, Y. Wen, Y. Song, Y. Zhang, Y. Li, Y. Mao, Z. D. Coudert, Z. Yan, Z. Chen, Z. Papakipos, A. Singh, A. Grattafiori, A. Jain, A. Kelsey, A. Shajnfeld, A. Gangidi, A. Victoria, A. Goldstand, A. Menon, A. Sharma, A. Boesenberg, A. Vaughan, A. Baevski, A. Feinstein, A. Kallet, A. Sangani, A. Yunus, A. Lupu, A. Alvarado, A. Caples, A. Gu, A. Ho, A. Poulton, A. Ryan, A. Ramchandani, A. Franco, A. Saraf, A. Chowdhury, A. Gabriel, A. Bharambe, A. Eisenman, A. Yazdan, B. James, B. Maurer, B. Leonhardi, B. Huang, B. Loyd, B. D. Paola, B. Paranjape, B. Liu, B. Wu, B. Ni, B. Hancock, B. Wasti, B. Spence, B. Stojkovic, B. Gamido, B. Montalvo, C. Parker, C. Burton, C. Mejia, C. Wang, C. Kim, C. Zhou, C. Hu, C.-H. Chu, C. Cai, C. Tindal, C. Feichtenhofer, D. Civin, D. Beaty, D. Kreymer, D. Li, D. Wyatt, D. Adkins, D. Xu, D. Testuggine, D. David, D. Parikh, D. Liskovich, D. Foss, D. Wang, D. Le, D. Holland, E. Dowling, E. Jamil, E. Montgomery, E. Presani, E. Hahn, E. Wood, E. Brinkman, E. Arcaute, E. Dunbar, E. Smothers, F. Sun, F. Kreuk, F. Tian, F. Ozgenel, F. Caggioni, F. Guzmán, F. Kanayet, F. Seide, G. M. Florez, G. Schwarz, G. Badeer, G. Swee, G. Halpern, G. Thattai, G. Herman, G. Sizov, Guangyi, Zhang, G. Lakshminarayanan, H. Shojanazeri, H. Zou, H. Wang, H. Zha, H. Habeeb, H. Rudolph, H. Suk, H. Aspegren, H. Goldman, I. Damlaj, I. Molybog, I. Tufanov, I.-E. Veliche, I. Gat, J. Weissman, J. Geboski, J. Kohli, J. Asher, J.-B. Gaya, J. Marcus, J. Tang, J. Chan, J. Zhen, J. Reizenstein, J. Teboul, J. Zhong, J. Jin, J. Yang, J. Cummings, J. Carvill, J. Shepard, J. McPhie, J. Torres, J. Ginsburg, J. Wang, K. Wu, K. H. U, K. Saxena, K. Prasad, K. Khandelwal, K. Zand, K. Matosich, K. Veeraraghavan, K. Michelena, K. Li, K. Huang, K. Chawla, K. Lakhotia, K. Huang, L. Chen, L. Garg, L. A, L. Silva, L. Bell, L. Zhang, L. Guo, L. Yu, L. Moshkovich, L. Wehrstedt, M. Khabsa, M. Avalani, M. Bhatt, M. Tsimpoukelli, M. Mankus, M. Hasson, M. Lennie, M. Reso, M. Groshev, M. Naumov, M. Lathi, M. Keneally, M. L. Seltzer, M. Valko, M. Restrepo, M. Patel, M. Vyatskov, M. Samvelyan, M. Clark, M. Macey, M. Wang, M. J. Hermoso, M. Metanat, M. Rastegari, M. Bansal, N. Santhanam, N. Parks, N. White, N. Bawa, N. Singhal, N. Egebo, N. Usunier, N. P. Laptev, N. Dong, N. Zhang, N. Cheng, O. Chernoguz, O. Hart, O. Salpekar, O. Kalinli, P. Kent, P. Parekh, P. Saab, P. Balaji, P. Rittner, P. Bontrager, P. Roux, P. Dollar, P. Zvyagina, P. Ratanchandani, P. Yuvraj, Q. Liang, R. Alao, R. Rodriguez, R. Ayub, R. Murthy, R. Nayani, R. Mitra, R. Li, R. Hogan, R. Battey, R. Wang, R. Maheswari, R. Howes, R. Rinott, S. J. Bondu, S. Datta, S. Chugh, S. Hunt, S. Dhillon, S. Sidorov, S. Pan, S. Verma, S. Yamamoto, S. Ramaswamy, S. Lindsay, S. Lindsay, S. Feng, S. Lin, S. C. Zha, S. Shankar, S. Zhang, S. Zhang, S. Wang, S. Agarwal, S. Sajuyigbe, S. Chintala, S. Max, S. Chen, S. Kehoe, S. Satterfield, S. Govindaprasad, S. Gupta, S. Cho, S. Virk, S. Subramanian, S. Choudhury, S. Goldman, T. Remez, T. Glaser, T. Best, T. Kohler, T. Robinson, T. Li, T. Zhang, T. Matthews, T. Chou, T. Shaked, V. Vontimitta, V. Ajayi, V. Montanez, V. Mohan, V. S. Kumar, V. Mangla, V. Albiero, V. Ionescu, V. Poenaru, V. T. Mihailescu, V. Ivanov, W. Li, W. Wang, W. Jiang, W. Bouaziz, W. Constable, X. Tang, X. Wang, X. Wu, X. Wang, X. Xia, X. Wu, X. Gao, Y. Chen, Y. Hu, Y. Jia, Y. Qi, Y. Li, Y. Zhang, Y. Zhang, Y. Adi, Y. Nam, Yu, Wang, Y. Hao, Y. Qian, Y. He, Z. Rait, Z. DeVito, Z. Rosnbrick, Z. Wen, Z. Yang, and Z. Zhao, "The llama 3 herd of models," 2024. [Online]. Available: https://arxiv.org/abs/2407.21783

[2] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," 2019.

[3] J. Schmidhuber, "Reinforcement learning upside down: Don't predict rewards – just map them to actions," 2020. [Online]. Available: https://arxiv.org/abs/1912.02875

[4] R. K. Srivastava, P. Shyam, F. Mutz, W. Jaśkowski, and J. Schmidhuber, "Training agents using upside-down reinforcement learning," 2021. [Online]. Available: https://arxiv.org/abs/1912.02877

[5] W. Kwon, Z. Li, S. Zhuang, Y. Sheng, L. Zheng, C. H. Yu, J. E. Gonzalez, H. Zhang, and I. Stoica, "Efficient memory management for large language model serving with pagedattention," in *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.

[6] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, and W. E. Sayed, "Mistral 7b," 2023. [Online]. Available: https://arxiv.org/abs/2310.06825

[7] G. Team, A. Kamath, J. Ferret, S. Pathak, N. Vieillard, R. Merhej, S. Perrin, T. Matejovicova, A. Ramé, M. Rivière, L. Rouillard, T. Mesnard, G. Cideron, J. bastien Grill, S. Ramos, E. Yvinec, M. Casbon, E. Pot, I. Penchev, G. Liu, F. Visin, K. Kenealy, L. Beyer, X. Zhai, A. Tsitsulin, R. Busa-Fekete, A. Feng, N. Sachdeva, B. Coleman, Y. Gao, B. Mustafa, I. Barr, E. Parisotto, D. Tian, M. Eyal, C. Cherry, J.-T. Peter, D. Sinopalnikov, S. Bhupatiraju, R. Agarwal, M. Kazemi, D. Malkin, R. Kumar, D. Vilar, I. Brusilovsky, J. Luo, A. Steiner, A. Friesen, A. Sharma, A. Sharma, A. M. Gilady, A. Goedeckemeyer, A. Saade, A. Feng, A. Kolesnikov, A. Bendebury, A. Abdagic, A. Vadi, A. György, A. S. Pinto, A. Das, A. Bapna, A. Miech, A. Yang, A. Paterson, A. Shenoy, A. Chakrabarti, B. Piot, B. Wu, B. Shahriari, B. Petrini, C. Chen, C. L. Lan, C. A. Choquette-Choo, C. Carey, C. Brick, D. Deutsch, D. Eisenbud, D. Cattle, D. Cheng, D. Paparas, D. S. Sreepathihalli, D. Reid, D. Tran, D. Zelle, E. Noland, E. Huizenga, E. Kharitonov, F. Liu, G. Amirkhanyan, G. Cameron, H. Hashemi, H. Klimczak-Plucińska, H. Singh, H. Mehta, H. T. Lehri, H. Hazimeh, I. Ballantyne, I. Szpektor, I. Nardini, J. Pouget-Abadie, J. Chan, J. Stanton, J. Wieting, J. Lai, J. Orbay, J. Fernandez, J. Newlan, J. yeong Ji, J. Singh, K. Black, K. Yu, K. Hui, K. Vodrahalli, K. Greff, L. Qiu, M. Valentine, M. Coelho, M. Ritter, M. Hoffman, M. Watson, M. Chaturvedi, M. Moynihan, M. Ma, N. Babar, N. Noy, N. Byrd, N. Roy, N. Momchev, N. Chauhan, N. Sachdeva, O. Bunyan, P. Botarda, P. Caron, P. K. Rubenstein, P. Culliton, P. Schmid, P. G. Sessa, P. Xu, P. Stanczyk, P. Tafti, R. Shivanna, R. Wu, R. Pan, R. Rokni, R. Willoughby, R. Vallu, R. Mullins, S. Jerome, S. Smoot, S. Girgin, S. Iqbal, S. Reddy, S. Sheth, S. Põder, S. Bhatnagar, S. R. Panyam, S. Eiger, S. Zhang, T. Liu, T. Yacovone, T. Liechty, U. Kalra, U. Evci, V. Misra, V. Roseberry, V. Feinberg, V. Kolesnikov, W. Han, W. Kwon, X. Chen, Y. Chow, Y. Zhu, Z. Wei, Z. Egyed, V. Cotruta, M. Giang, P. Kirk, A. Rao, K. Black, N. Babar, J. Lo, E. Moreira, L. G. Martins, O. Sanseviero, L. Gonzalez, Z. Gleicher, T. Warkentin, V. Mirrokni, E. Senter, E. Collins, J. Barral, Z. Ghahramani, R. Hadsell, Y. Matias, D. Sculley, S. Petrov, N. Fiedel, N. Shazeer, O. Vinyals, J. Dean, D. Hassabis, K. Kavukcuoglu, C. Farabet, E. Buchatskaya, J.-B. Alayrac, R. Anil, Dmitry, Lepikhin, S. Borgeaud, O. Bachem, A. Joulin, A. Andreev, C. Hardin, R. Dadashi, and L. Hussenot, "Gemma 3 technical report,"

2025. [Online]. Available: https://arxiv.org/abs/2503.19786

[8] E. Bakouch, L. Ben Allal, A. Lozhkov, N. Tazi, L. Tunstall, C. M. Patiño, E. Beeching, A. Roucher, A. J. Reedi, Q. Gallouédec, K. Rasul, N. Habib, C. Fourrier, H. Kydlicek, G. Penedo, H. Larcher, M. Morlon, V. Srivastav, J. Lochner, X.-S. Nguyen, C. Raffel, L. von Werra, and T. Wolf, "SmolLM3: smol, multilingual, long-context reasoner," https://huggingface.co/blog/smollm3, 2025.

[9] M. Abdin, J. Aneja, H. Behl, S. Bubeck, R. Eldan, S. Gunasekar, M. Harrison, R. J. Hewett, M. Javaheripi, P. Kauffmann, J. R. Lee, Y. T. Lee, Y. Li, W. Liu, C. C. T. Mendes, A. Nguyen, E. Price, G. de Rosa, O. Saarikivi, A. Salim, S. Shah, X. Wang, R. Ward, Y. Wu, D. Yu, C. Zhang, and Y. Zhang, "Phi-4 technical report," 2024. [Online]. Available: https://arxiv.org/abs/2412.08905

[10] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv, C. Zheng, D. Liu, F. Zhou, F. Huang, F. Hu, H. Ge, H. Wei, H. Lin, J. Tang, J. Yang, J. Tu, J. Zhang, J. Yang, J. Yang, J. Zhou, J. Zhou, J. Lin, K. Dang, K. Bao, K. Yang, L. Yu, L. Deng, M. Li, M. Xue, M. Li, P. Zhang, P. Wang, Q. Zhu, R. Men, R. Gao, S. Liu, S. Luo, T. Li, T. Tang, W. Yin, X. Ren, X. Wang, X. Zhang, X. Ren, Y. Fan, Y. Su, Y. Zhang, Y. Zhang, Y. Wan, Y. Liu, Z. Wang, Z. Cui, Z. Zhang, Z. Zhou, and Z. Qiu, "Qwen3 technical report," 2025. [Online]. Available: https://arxiv.org/abs/2505.09388

[11] OpenAI, J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, R. Avila, I. Babuschkin, S. Balaji, V. Balcom, P. Baltescu, H. Bao, M. Bavarian, J. Belgum, I. Bello, J. Berdine, G. Bernadett-Shapiro, C. Berner, L. Bogdonoff, O. Boiko, M. Boyd, A.-L. Brakman, G. Brockman, T. Brooks, M. Brundage, K. Button, T. Cai, R. Campbell, A. Cann, B. Carey, C. Carlson, R. Carmichael, B. Chan, C. Chang, F. Chantzis, D. Chen, S. Chen, R. Chen, J. Chen, M. Chen, B. Chess, C. Cho, C. Chu, H. W. Chung, D. Cummings, J. Currier, Y. Dai, C. Decareaux, T. Degry, N. Deutsch, D. Deville, A. Dhar, D. Dohan, S. Dowling, S. Dunning, A. Ecoffet, A. Eleti, T. Eloundou, D. Farhi, L. Fedus, N. Felix, S. P. Fishman, J. Forte, I. Fulford, L. Gao, E. Georges, C. Gibson, V. Goel, T. Gogineni, G. Goh, R. Gontijo-Lopes, J. Gordon, M. Grafstein, S. Gray, R. Greene, J. Gross, S. S. Gu, Y. Guo, C. Hallacy, J. Han, J. Harris, Y. He, M. Heaton, J. Heidecke, C. Hesse, A. Hickey, W. Hickey, P. Hoeschele, B. Houghton, K. Hsu, S. Hu, X. Hu, J. Huizinga, S. Jain, S. Jain, J. Jang, A. Jiang, R. Jiang, H. Jin, D. Jin, S. Jomoto, B. Jonn, H. Jun, T. Kaftan, Łukasz Kaiser, A. Kamali, I. Kanitscheider, N. S. Keskar, T. Khan, L. Kilpatrick, J. W. Kim, C. Kim, Y. Kim, J. H. Kirchner, J. Kiros, M. Knight, D. Kokotajlo, Łukasz Kondraciuk, A. Kondrich, A. Konstantinidis, K. Kosic, G. Krueger, V. Kuo, M. Lampe, I. Lan, T. Lee, J. Leike, J. Leung, D. Levy, C. M. Li, R. Lim, M. Lin, S. Lin, M. Litwin, T. Lopez, R. Lowe, P. Lue, A. Makanju, K. Malfacini, S. Manning, T. Markov, Y. Markovski, B. Martin, K. Mayer, A. Mayne, B. McGrew, S. M. McKinney, C. McLeavey, P. McMillan, J. McNeil, D. Medina, A. Mehta, J. Menick, L. Metz, A. Mishchenko, P. Mishkin, V. Monaco, E. Morikawa, D. Mossing, T. Mu, M. Murati, O. Murk, D. Mély, A. Nair, R. Nakano, R. Nayak, A. Neelakantan, R. Ngo, H. Noh, L. Ouyang, C. O'Keefe, J. Pachocki, A. Paino, J. Palermo, A. Pantuliano, G. Parascandolo, J. Parish, E. Parparita, A. Passos, M. Pavlov, A. Peng, A. Perelman, F. de Avila Belbute Peres, M. Petrov, H. P. de Oliveira Pinto, Michael, Pokorny, M. Pokrass, V. H. Pong, T. Powell, A. Power, B. Power, E. Proehl, R. Puri, A. Radford, J. Rae, A. Ramesh, C. Raymond, F. Real, K. Rimbach, C. Ross, B. Rotsted, H. Roussez, N. Ryder, M. Saltarelli, T. Sanders, S. Santurkar, G. Sastry, H. Schmidt, D. Schnurr, J. Schulman, D. Selsam, K. Sheppard, T. Sherbakov, J. Shieh, S. Shoker, P. Shyam, S. Sidor, E. Sigler, M. Simens, J. Sitkin, K. Slama, I. Sohl, B. Sokolowsky, Y. Song, N. Staudacher, F. P. Such, N. Summers, I. Sutskever, J. Tang, N. Tezak, M. B. Thompson, P. Tillet, A. Tootoonchian, E. Tseng, P. Tuggle, N. Turley, J. Tworek, J. F. C. Uribe, A. Vallone, A. Vijayvergiya, C. Voss, C. Wainwright, J. J. Wang, A. Wang, B. Wang, J. Ward, J. Wei, C. Weinmann, A. Welihinda, P. Welinder, J. Weng, L. Weng, M. Wiethoff, D. Willner, C. Winter, S. Wolrich, H. Wong, L. Workman, S. Wu, J. Wu, M. Wu, K. Xiao, T. Xu, S. Yoo, K. Yu, Q. Yuan, W. Zaremba, R. Zellers, C. Zhang, M. Zhang, S. Zhao, T. Zheng, J. Zhuang, W. Zhuk, and B. Zoph, "Gpt-4 technical report," 2024. [Online]. Available: https://arxiv.org/abs/2303.08774