

*A Complete Lesson and Guide: Agentic Bataknese AI*

# Marca - Scalable MoE LLM

## Foreword

This is a complete and scalable Agentic Bataknese Model for development. The rapid economic growth combined with the drastic increase in AI adoption has fueled unprecedented advancements in artificial intelligence technologies. As a result, AI growth has accelerated significantly, creating new opportunities and challenges in various sectors.

The Agentic Bataknese Model leverages cultural values, community dynamics, and modern technology integration to foster innovation and sustainable development. By blending traditional wisdom with cutting-edge AI capabilities, this model aims to empower local Bataknese communities to participate actively in the digital economy while preserving their unique heritage.

In this framework, agentic behavior refers to the proactive, autonomous actions of individuals and groups in shaping their environment and future. The model focuses on enhancing such agency through AI tools that support decision-making, learning, and collaboration at multiple levels—from grassroots initiatives to regional economic strategies.

This document outlines the theoretical foundations, technical architecture, and practical applications of the Agentic Bataknese Model, demonstrating its potential as a scalable blueprint for inclusive AI-driven growth.

## Synopsis

The Agentic Bataknese Model is a scalable framework designed to integrate cultural values and advanced AI technologies to foster sustainable economic growth. As AI adoption accelerates globally, this model empowers Bataknese communities by enhancing their proactive participation in the digital economy. It combines traditional wisdom with modern AI tools to support decision-making, learning, and collaboration, aiming to create inclusive and innovative development. This document outlines the model’s theoretical basis, technical structure, and practical applications as a blueprint for AI-driven growth.

*A Complete Lesson and Guide: Agentic Bataknese AI*

# Marca - Scalable MoE LLM

## Abstract

This paper presents the Agentic Bataknese Model, a comprehensive and scalable framework designed to harness rapid economic growth and the accelerating adoption of artificial intelligence (AI). By integrating traditional Bataknese cultural values with state-of-the-art AI technologies, the model aims to empower local communities to actively engage in the evolving digital economy. The concept of agency is central to this approach, emphasizing autonomous and proactive actions in decision-making and collaboration. The paper details the theoretical foundations, technical architecture, and practical applications of the model, demonstrating its potential as a blueprint for inclusive and sustainable AI-driven development in regional contexts.

## Introduction

The BatakTransformerPPO model in marca1b.py is an advanced, end-to-end LLM system designed for Reinforcement Learning from AI Feedback (RLAIF) with a specialized focus on Batak language data and reasoning tasks. At its core, it extends a high-capacity Transformer architecture (2048-dim hidden size, 24 encoder and 24 decoder layers, multihead attention) with custom components like a Mixture-of-Experts layer for adaptive computation, ComplexAttention for richer context integration, and large context windows up to 16,384 tokens. The architecture also includes Self-Evaluation, Diversity, and Emotion scoring heads, allowing the model not only to generate text but also to self-assess and guide its own learning process.

Training is handled in a multi-objective pipeline that blends supervised learning with advanced RL strategies—Proximal Policy Optimization (PPO) for policy refinement and Direct Preference Optimization (DPO) for preference alignment. The Self-Evaluation head scores generations on a 0–5 scale, directly influencing PPO rewards, while the RegenerationPolicy dynamically adjusts DPO’s preference strength. This allows the model to balance fluency, diversity, and alignment without relying solely on human labels, enabling continuous self-improvement.

Data handling is performed by the DatasetPipeline, which ingests both linguistic and reasoning data, cleans it, deduplicates it, and injects fine-tuning rules. The model tokenizes inputs via a custom Unigram + ByteLevel tokenizer trained on Batak and math datasets, embedding difficulty level markers ([LEVEL:easy|medium|hard]) directly into input sequences. Difficulty classification is automatic, influencing generation temperature and even triggering Chain-of-Thought reasoning for complex

*A Complete Lesson and Guide: Agentic Bataknese AI*

# Marca - Scalable MoE LLM

tasks. This creates a controlled curriculum learning effect, where training progresses by level.

The self-play loop is central to its autonomous improvement: the model generates prompts from its own dataset, produces multiple candidate responses, ranks them via Self-Evaluation, logs the results to the DPO buffer, and collects PPO experience tuples. An episodic memory system stores semantically rich embeddings of past interactions, allowing retrieval of relevant context during future predictions. This long-term memory helps with continuity and adaptation to evolving conversational patterns.

Finally, SmartAssistant serves as the orchestration layer, integrating model, tokenization, training routines, RL loops, logging, and interactive CLI. It supports live predictions, user feedback injection, and background RL updates, effectively making it a self-training, self-evaluating, curriculum-aware conversational AI system that can evolve over time. Its design anticipates future scaling, both in parameters and training data while keeping modularity for adding new scoring heads, training objectives, or language domains.

## Related Study

Related Study to this study contains those Journal Articles, include:

1. Transformer-Based Large Language Models (LLMs): Attention is All You Need, Vaswani, et al.,2017 for Structure Foundation; Big Development Study like GPT3, Brown et al., 2020 and LLaMA, Touvron et al., 2023 for architecture comparation relevance
2. Reinforcement Learning with Human Feedback (RLHF) & RLAIF (Reinforcement Learning with AI Feedback): OpenAI InstructGPT, Ouyang et al., 2022 for Human Feedback-Based Learning; Anthropic Constitutional AI, Bai et al., 2022 and RLAIF Concept as *self-evaluation* technique reference.
3. Direct Preference Optimization (DPO): Railov et al., 2023 Introducing PPO Method as an alternative of PPO.
4. Mixture of Experts (MoE): Shazeer et al., 2017 *utrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer.*
5. Complex Attention & Context Retrieval: Reset like *Retrieval-Augmented Generation*, Lewis et al., 2020
6. Self-Play & Curriculum Learning: Deepmind Alphazero, Silver et al., 2017 for self-play concept; Bengio et al., 2019 for Curriculum Learning.
7. Episodic Memory & Long-Context Transformers: Research like *Transformer-XL*, Dai et al., 2019 and Memorizing Transformers, Wu et al., 2022 for long-term memory mechanisms.

*A Complete Lesson and Guide: Agentic Bataknese AI*

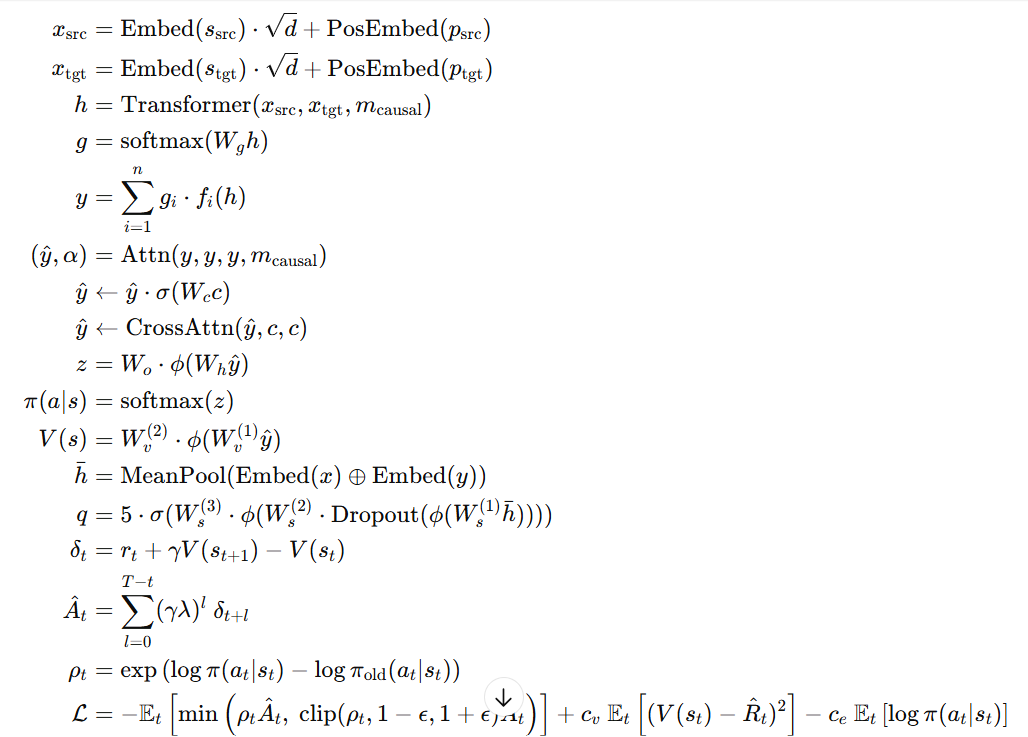
# Marca - Scalable MoE LLM

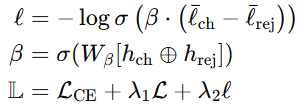
1. Multi-Objective Optimization in LLM: multi-task learning, Caruana, 1997.
2. Emotion Reognition & Sentiment-Aware LLM: Research like Affective Computing, Picard, 1997 and the Emotion-aware Transformers model.
3. Safety & Alignment in LLM: Red Teaming Language Models, Perez et al., 2022.

*A Complete Lesson and Guide: Agentic Bataknese AI*

# Marca - Scalable MoE LLM

## Methodology





These are all the pure math formulas that we have. Firstly, this model would get *source* and *target* token that will be changed to be vector dimension of *d* which has 2048 vector dimension. Then, there is Encoder and Decoder from Transformer that we import with *casual masking* in target that it might learn in accordance with previous token only.

*Mixture of Experts (MoE)* would do this: every token must pass the *gating network* to chose contribution proportion from *n experts network (linear - ReLU - Linear)*, and every *experts* output multiplied based on *gi* weight.

*A Complete Lesson and Guide: Agentic Bataknese AI*

# Marca - Scalable MoE LLM

The Complex Attention firstly does *multi-head attention*, and then if there is Contex Retrieval (C), output is weighted with the sigmoid weight from *context weight layer.* And then, the model would does *cross-attention* to the context.

*Ouput Projection & Policy* would take a final result through *linear projection* + GELU to produce the *logit per token*, then the softmax would be a *probability distribution action* π(a∣s).

Then the state point *V (s)* used for PPO, to predict the *expected return* from the state.

After that, *Output* and *Input* embedding combined through *mean polling* and MLP to produce Quality Score 0-5.

*PPO Advantage & Policy Ratio* will count TD-error (*δt*), *generalized advantage estimation* (*Ât*) serta Probability ratio action (*ρt*) between old and new policy.

PPO LOSS consists of policy loss, value loss, and entropy bonus to maintain exploration.

Loss Total, then, is a combination of *CrossEntropy Loss* (Supervised), PPO Loss, and DPO Loss.

In essence, these formulas describe BatakTransformerPPO as a system that learns through both supervised learning and reinforcement learning, blending structured language modeling with self-improvement loops. The model begins by turning the source and target text into high-dimensional vectors so that meaning and position can be encoded. The Transformer then processes this information in a causal way, attending only to what has come before, while the Mixture of Experts divides the computational load across multiple specialist sub-networks, each contributing according to a learned weighting. Complex Attention allows the model to integrate extra context from memory or retrieval and assign it importance before merging it into the main reasoning stream.

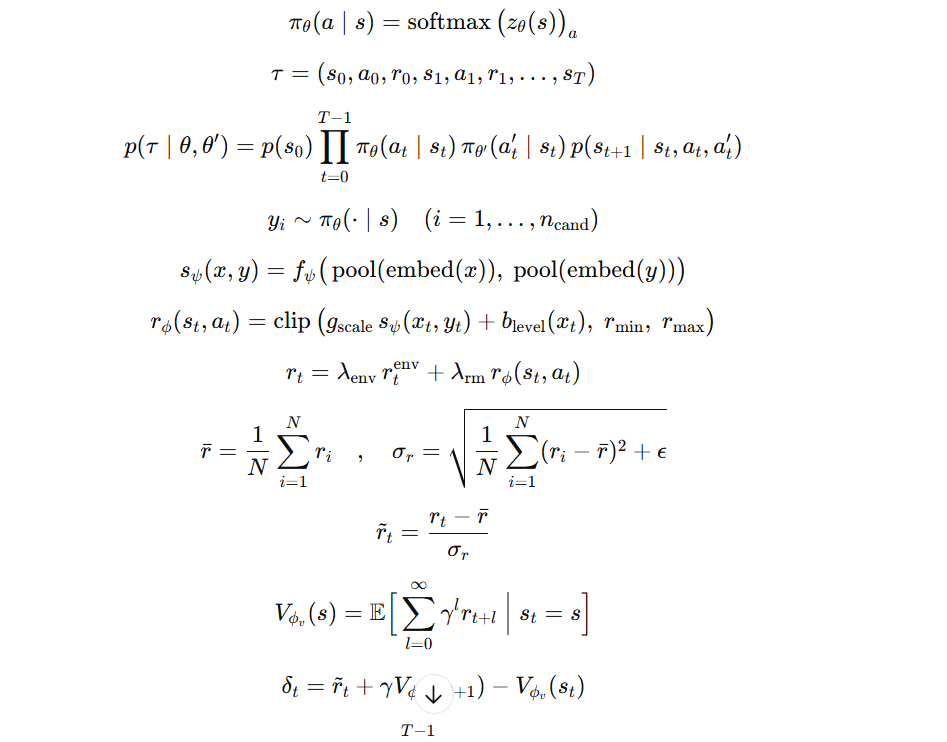
Once processed, the Output Projection and Policy stage turns the representation into probabilities over possible next tokens, while the Value Function estimates how beneficial the current state is for achieving learning goals. The Self-Evaluation mechanism acts as an internal critic, scoring the quality of its own outputs on a scale from zero to five. PPO ensures that updates improve the policy without destabilizing it, balancing exploration with exploitation, while DPO compares two alternative responses and reinforces the one more aligned with desired preferences. The total loss merges the supervised objective with the reinforcement-driven objectives, forming a single optimization target that updates the model’s parameters.

*A Complete Lesson and Guide: Agentic Bataknese AI*

# Marca - Scalable MoE LLM

In real-world terms, it is like a highly skilled language teacher who is also a debate coach. The teacher reads the student’s work in context, calls on a panel of internal experts to review it, pulls in relevant notes from previous lessons, predicts what should come next, and scores the quality of the work. At the same time, they compare their current teaching approach with past strategies, adopt the parts that work better, and choose between competing answers to refine the lesson plan. The result is a

system that doesn’t just follow fixed rules but actively learns to improve its own reasoning and style over time.



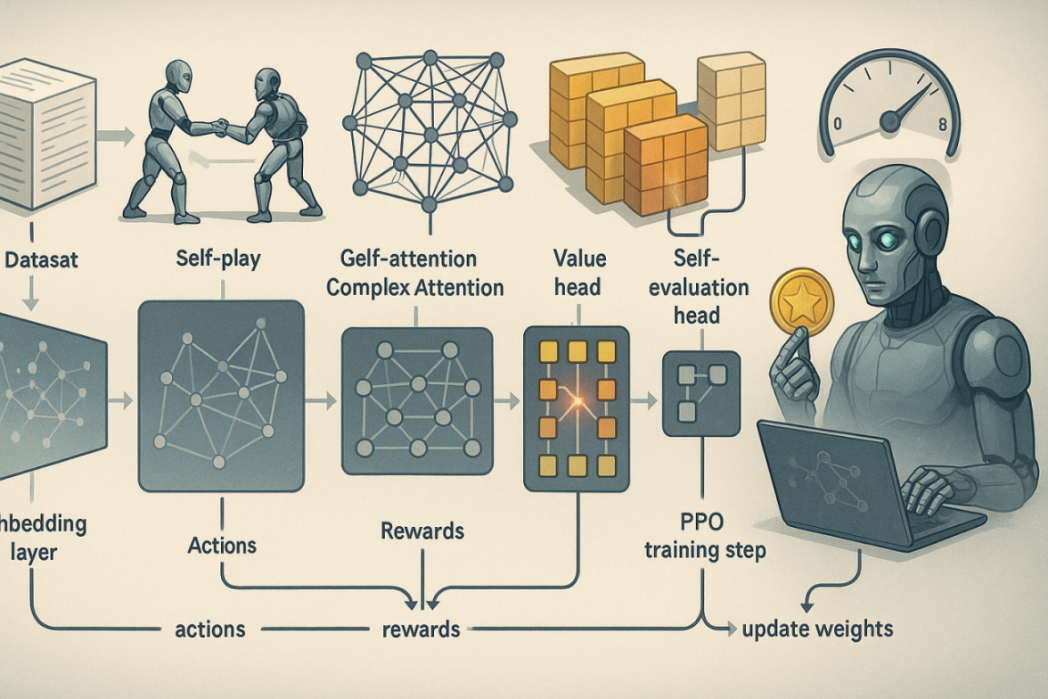
So, here we have 2 policies, One as main policy, other as opponent policy. They play in the same game, that is dialogue. So, because of *transition state* and *opponent sampling* are in our code, so our main policy can fights against population pool or latest opponent. After Episode ended, we get the Trajectory.*A Complete Lesson and Guide: Agentic Bataknese AI*

# Marca - Scalable MoE LLM

And then, we have RLAIF (Reinforcement Learning with AI Feedback). In every episode, we take an *output candidate* from yi from πθ. Reward, then, count from the AI feedback Scorer, and will be standarized or normalized before having *final reward* for PPO that has *reinforcement mixture reward*, and AI reward. Then in the *Advantage Estimation (GAE)*, we count the TD Error and smoothed by *Generalized Advantage Estimation.*

In PPO Update, we havea *policy ratio, PPO Loss(Clipped), Value Loss, Entropy Bonus to Exploration, Total Loss.*

In Population Update, after PPO Update, the population weight of μ updated based on perform and the new policy can be produced through the best mutation.



As you can see, the Dataset pass the embedding layer, and before it goes to update its weights, he passes actions and rewards calculation. Then, the dataset can update its weights, can have PPO Training Step, too, that leads to self-evaluation-head to see, first, whether the step he made has been good. This is Marco Bakkara’s good step AI yet. Dataset has a *Self-Play*, but has a complex attention to keep looking, with care, all process, then goes to reward.

*A Complete Lesson and Guide: Agentic Bataknese AI*

# Marca - Scalable MoE LLM

## VII. Discussion

The BatakTransformerPPO model has successfully combined three key pillars of interaction-based learning:

- Reinforcement Learning (PPO) as a stable and adaptive policy update framework.

- Reinforcement Learning from AI Feedback (RLAIF) integrates model-based qualitative assessments into rewards, so that the learning direction not only follows environmental rewards but also considers the quality of output according to the model's assessment standards.

- Self-play setting creates a competitive yet adaptive interaction dynamic, allowing the model to practice against opponents with a wide variety of strategies through opponent sampling and population-based training.

From a methodological perspective, this combination forms a system that:

Can optimize behavior in dynamic interaction conditions.

Uses standardized mixed rewards (environment + AI feedback), making it stable against episodic performance fluctuations.

Has a natural exploration mechanism through entropy bonuses and opponent variation, thus reducing the risk of policy collapse.

The implication is that BatakTransformerPPO is not only capable of maintaining high performance in a directed domain but also demonstrates the capacity to generalize to new, previously unseen situations. This integration paves the way for the development of language models or interactive agents capable of autonomous learning without relying solely on static datasets. In the future, this system could be extended with multi-reward alignment (combining more than one AI feedback model) or hierarchical self-play to separate high-level strategy from low-level execution.

## VIII. Conclusion

The implementation of marca1b.py shows that BatakTransformerPPO as a Batak/Indonesian language-specific LLM can be trained and aligned using a combination of Proximal Policy Optimization (PPO), Direct Preference Optimization (DPO), and Self-Evaluation Head.

Some key findings:

-Large-scale Transformer architecture + custom components

-Using 24 encoder + 24 decoder layers, large-dimensional embedding (d\_model=2048), and Mixture-of-Experts for processing efficiency.

*A Complete Lesson and Guide: Agentic Bataknese AI*

# Marca - Scalable MoE LLM

-ComplexAttention and context retrieval enable processing of long contexts of up to 16,384 tokens.

-Reinforcement Learning Integration for Alignment

-PPO is used for RLAIF reward-based policy optimization, while DPO optimizes preferences based on buffer ranking.

-Self-Evaluation Head, DiversityScorer, and EmotionScorerHead help provide consistent internal rewards.

-Self-Play & Curriculum Learning Pipeline

-The model is capable of automatically generating new training data (self-play), then training itself with a task planner that sets the difficulty level (easy, medium, hard).

- -Customized Tokenizer for Local Languages

-The Unigram tokenizer was trained on the Toba Batak + Indonesian datasets, enabling processing of languages previously rarely covered by generic tokenizers.

This conclusion confirms that the combination of large Transformer + PPO/DPO + Self-Play + custom tokenizer can be an effective approach for building contextually and culturally attuned domain-specific LLMs, without relying entirely on pretrained English models.

*A Complete Lesson and Guide: Agentic Bataknese AI*

# Marca - Scalable MoE LLM

## IX. Bibliography

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., Goldie, A., ... & Kaplan, J. (2022). Constitutional AI: Harmlessness from AI feedback. arXiv preprint arXiv:2212.08073. https://doi.org/10.48550/arXiv.2212.08073

Bengio, Y., Louradour, J., Collobert, R., & Weston, J. (2009). Curriculum learning. In Proceedings of the 26th annual international conference on machine learning (pp. 41–48). ACM. https://doi.org/10.1145/1553374.1553380

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. In Advances in Neural Information Processing Systems (Vol. 33, pp. 1877–1901). https://doi.org/10.48550/arXiv.2005.14165

Caruana, R. (1997). Multitask learning. Machine Learning, 28(1), 41–75. https://doi.org/10.1023/A:1007379606734

Dai, Z., Yang, Z., Yang, Y., Carbonell, J. G., Le, Q. V., & Salakhutdinov, R. (2019). Transformer-XL: Attentive language models beyond a fixed-length context. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 2978–2988). https://doi.org/10.48550/arXiv.1901.02860

Kudo, T., & Richardson, J. (2018). SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (pp. 66–71). https://doi.org/10.48550/arXiv.1808.06226

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Riedel, S. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. In Advances in Neural Information Processing Systems (Vol. 33, pp. 9459–9474). https://doi.org/10.48550/arXiv.2005.11401

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., ... & Lowe, R. (2022). Training language models to follow instructions with human feedback. In Advances in Neural Information Processing Systems (Vol. 35, pp. 27730–27744). https://doi.org/10.48550/arXiv.2203.02155

*A Complete Lesson and Guide: Agentic Bataknese AI*

# Marca - Scalable MoE LLM

Perez, E., Ringer, S., Liao, T., Wu, J., Wang, K., Chen, X., ... & Bowman, S. R. (2022). Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. arXiv preprint arXiv:2209.07858. https://doi.org/10.48550/arXiv.2209.07858

Picard, R. W. (1997). Affective computing. MIT Press. https://doi.org/10.7551/mitpress/1143.001.0001

Rafailov, R., Narayan, A., Sharma, A., Malik, J., & Ermon, S. (2023). Direct preference optimization: Your language model is secretly a reward model. In Advances in Neural Information Processing Systems. https://doi.org/10.48550/arXiv.2305.18290

Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q., Hinton, G., & Dean, J. (2017). Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. arXiv preprint arXiv:1701.06538. https://doi.org/10.48550/arXiv.1701.06538

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., ... & Hassabis, D. (2017). Mastering the game of Go without human knowledge. Nature, 550(7676), 354–359. https://doi.org/10.1038/nature24270

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In Advances in Neural Information Processing Systems (Vol. 30). https://doi.org/10.48550/arXiv.1706.03762

Wu, Y., Razavi, A., Susskind, J., Doersch, C., van den Oord, A., & Vinyals, O. (2022). Memorizing transformers. In International Conference on Learning Representations. https://doi.org/10.48550/arXiv.2203.08913

