Maia-2: A Unified Model for Human-AI Alignment in Chess

Zhenwei Tang*, **Difan Jiao***, **Reid McIlroy-Young^**, **Jon Kleinberg~**, **Siddhartha Sen'**, **Ashton Anderson***

*University of Toronto, ^Harvard University, ~Cornell University, 'Microsoft Research NYC

Keywords: Human-AI Alignment, Action Prediction, Chess, Skill-aware Attention, Coherence | **Contact:** {josephtang, ashton}@cs.toronto.edu

Paper

Code

Introduction

Research Goal: Coherent human decision-making modeling at various skill levels

- For enhanced human-AI alignment
- For more relatable AI partners (humanness)
- For algorithmically-informed teaching

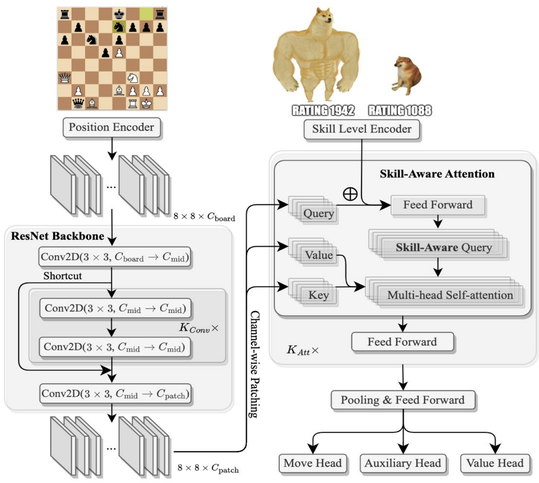
Why Chess?

- A historical testbed for AI research
- Precise measurements of skill via rating systems
- Data Availability: Online chess

Previous Work: Maia

- **Independent** models for different skill levels
- Lacks coherence in adapting to human improvement
- Limited effectiveness as AI partners and teaching tools

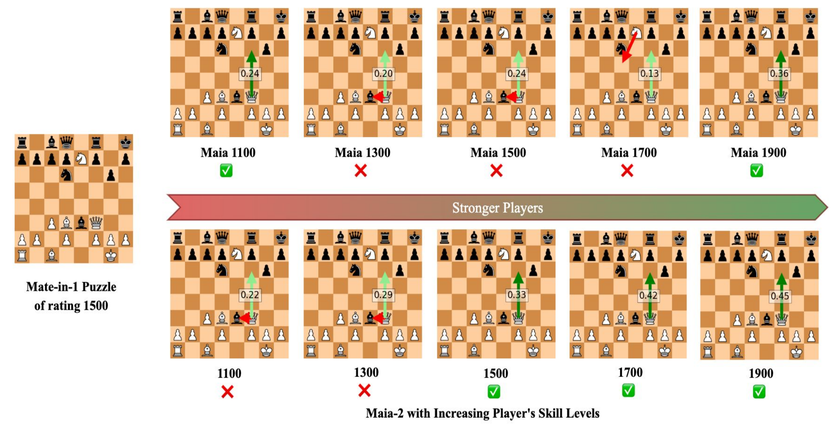
Methodology



An Unified One-for-all Model

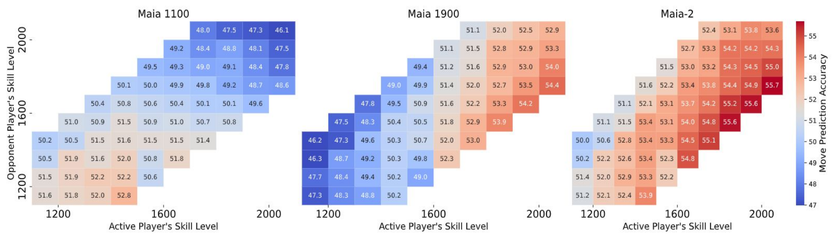
- ResNet Position Encoding
- Categorical Skill Level Encoding
- Bridging Skill Levels and Positions via:
  - **Skill-aware Attention**
  - **Channel-wise Patching**

Demo



Maia-2 and Maia solving a Mate-in-1 chess puzzle of rating 1500.

Results: Accuracy



Move prediction accuracy across diverse skill levels. Colors represent performance, with warmer tones indicating higher accuracy.

Experiments

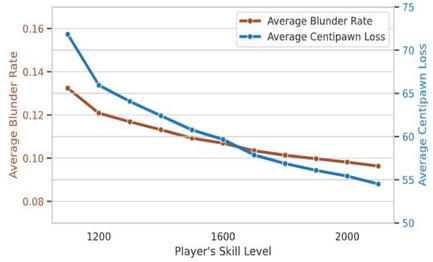
Datasets

- Training: 169M games (9.1B positions) played between May 2018 and Nov 2023
- Evaluation:
  - Maia benchmark dataset
  - Cross-Skill dataset: Lichess games played in Dec 2023
  - Grounded dataset: 450k positions from Dec 2023 games with Stockfish evaluations

Model Training

- Removing lookbacks according to Markov assumption for efficiency and flexibility
- Infusing auxiliary information
  - Objective, e.g., check delivery
  - Behavioral, e.g., moves' originating and destination squares
- Data Balancing w.r.t self and opponent skill-levels

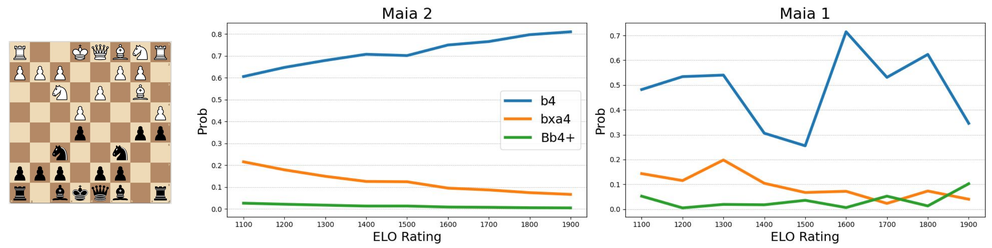
Results: Coherency



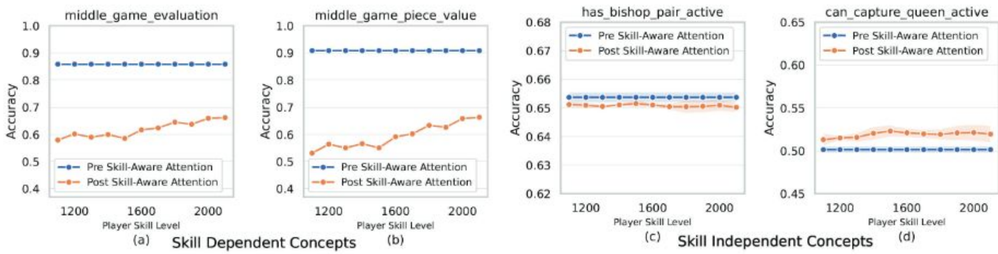
Quality of predicted moves.

	%Monotonic			%Transitional		
	Skilled	Advanced	Master	Skilled	Advanced	Master
Maia-1	1.61	1.42	1.14	13.34	18.14	20.48
Maia-2	27.61	28.51	26.38	22.59	23.39	21.72

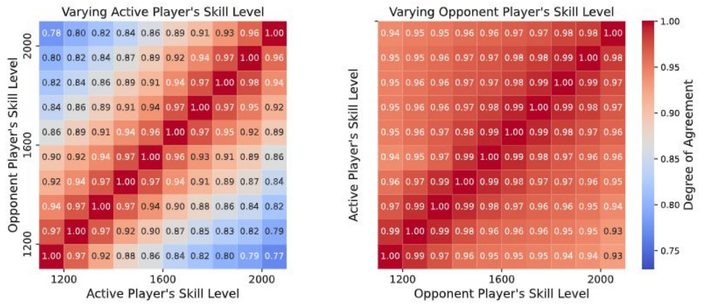
Percentage of monotonic and transitional positions.



An example of the more coherent predictions made by Maia-2.



Maia-2's chess concept recognition as a function of skill level, as measured by linear activation probes.



Move prediction agreement as (left) active player and (right) opponent player skill are varied.

Play Now

NEURAL INFORMATION PROCESSING SYSTEMS