Representation-Aligned Auxiliary Supervision for Language Model Adaptation

Kyuyoung Kim[1] **Peiyao Sheng**[2] **Ashwin Hebbar**[3] **Peiyang Xu**[3]
Yunfei Xie[4] **Kevin Wang**[5] **Rui Xin**[6] **Chen Wei**[4] **Zhangyang Wang**[5]
Jinwoo Shin[1] **Pramod Viswanath**[2,3] **Sewoong Oh**[2,6]
[1]KAIST AI [2]Sentient Labs [3]Princeton University
[4]Rice University [5]University of Texas, Austin [6]University of Washington

Abstract

Language models exhibit strong reasoning capabilities, yet adapting them to structured domains remains challenging and can yield inconsistent outcomes. We identify *representation compatibility*, the alignment between a model's preferences and the representations used for training, as a key factor in effective adaptation. We study this in chess, which provides a controlled testbed with precise semantics, computable optimal actions, and multiple state representations, including a symbolic encoding (FEN) and a spatial format (ASCII). We find that models exhibit strong preferences that affect both learning and generalization. Building on this observation, we introduce *representation-aligned auxiliary supervision*, which aligns environment-derived tasks with compatible representations to improve adaptation. Across model scales and representations, auxiliary supervision consistently improves optimal-move prediction relative to target-only training under identical target data, with larger gains in cross-representation transfer. Auxiliary tasks that expose environment dynamics provide larger and more consistent gains than surface-level or static supervision, while remaining competitive with substantially increasing the amount of target-task data. Moreover, ASCII-trained models transfer substantially better to FEN than FEN-trained models do to ASCII, even surpassing the FEN target-only baseline on FEN evaluation. The gains also extend beyond optimal-move prediction to open-ended, factually grounded commentary generation. Overall, our results show that effective adaptation can be achieved by providing auxiliary knowledge in representations aligned with the model.

1 Introduction

Language models are increasingly used for prediction and decision-making in structured domains, where an input corresponds to an underlying state (s) and the desired output is an action (a). Adapting models to such domains typically relies on supervised fine-tuning (SFT) on task-specific data, optionally followed by reinforcement learning (RL) [1, 2]. While this directly supervises the desired state-to-action behavior, obtaining sufficient target-specific supervision can be costly. Structured environments, however, often provide additional signals that can be derived automatically, such as state properties, action constraints, and state transitions. Auxiliary-task learning leverages these signals to provide the model with exposure to the environment structure that may help it learn the target state-to-action mapping [3]. We ask when and how such auxiliary supervision can effectively improve adaptation of language models to structured domains.

A key question is how the underlying state should be represented when providing such auxiliary learning signals. In this work, we study *representation compatibility*, the extent to which a model's preferred ways of processing inputs align with the representations used during training, as a factor in

Preprint.

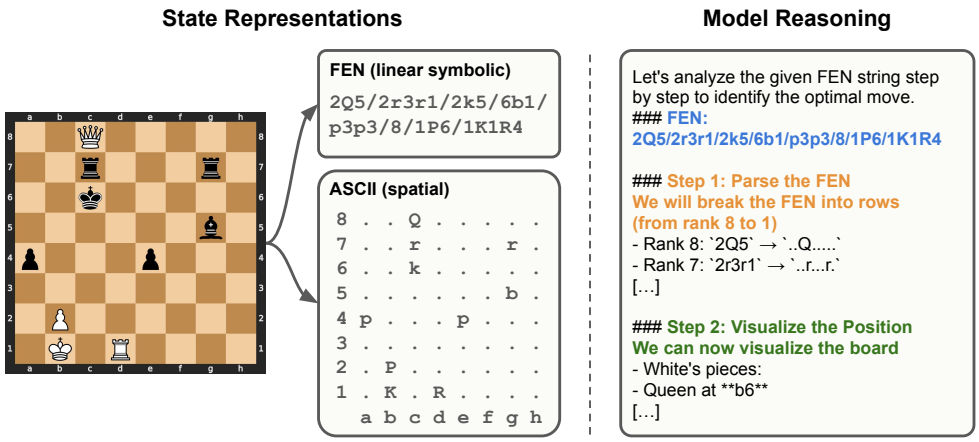

Figure 1: **Representation compatibility.** Given a chess position in FEN, Qwen3-8B reconstructs an ASCII board before reasoning about the task, illustrating how a model's processing of a structured state can often heavily depend on its representation. FEN is a compact linear encoding, while ASCII provides an explicit 8×8 spatial representation.

domain adaptation. This is motivated by an empirical observation: models often transform inputs into alternative forms prior to reasoning about the target task. For example, in chess, when given positions in Forsyth–Edwards Notation (FEN), a compact symbolic encoding, Qwen3-8B frequently converts them into ASCII board diagrams and performs reasoning in this spatial format (Figure 1). Consistent with this behavior, in the task of predicting next states after moves, the model achieves 47.4% accuracy in ASCII versus 3.9% in FEN, a twelve-fold gap despite their semantic equivalence (Figure 2). This asymmetry suggests that models do not process semantically equivalent representations uniformly, making representation compatibility an important consideration for adaptation.

Building on this observation, we ask whether expressing environment-derived auxiliary supervision in a representation compatible with the model can improve adaptation to the target state-to-action task. To this end, we introduce a simple training approach that jointly trains on the target objective and environment-derived auxiliary tasks. These tasks, including static state properties, state transitions, and action constraints, provide additional supervision about environment structure relevant to the target prediction task.

We study these questions in chess, where precise semantics, computable environment structure, and multiple representations enable controlled evaluation of auxiliary supervision and representation choice. Across our experiments, environment-derived auxiliary tasks consistently improve optimal-move prediction over target-only training. Moreover, these gains go beyond those obtained by simply increasing the amount of target-task data: although optimal-move accuracy improves with additional examples, environment-derived tasks provide notable gains without increasing the amount of target data. The largest gains come from tasks that expose environment dynamics, such as state transitions and action constraints, whereas simpler tasks based on state manipulation or static board properties provide smaller and less consistent improvements. We further find that the role of representation choice depends on the evaluation setting: its effects are most pronounced for state-understanding tasks and cross-representation transfer, while in-distribution optimal-move performance remains relatively robust across representations. Together, these results suggest that auxiliary supervision provides a complementary learning signal for structured state-to-action adaptation, and that the representation in which the underlying structure is learned is important for effective adaptation and transfer.

Contributions. In summary, we propose a simple approach for adapting language models to structured state-to-action tasks by jointly training on the target action-prediction objective and environment-derived auxiliary tasks, with the auxiliary supervision expressed in a representation compatible with the model. Our contributions are threefold:

- **Representation-aligned auxiliary supervision.** We introduce a training approach for incorporating environment-derived auxiliary tasks into language model adaptation, using representation compatibility to guide how the auxiliary supervision is expressed (Figure 2, Section 3).
- **Auxiliary supervision for structured adaptation.** We show that environment-derived auxiliary tasks improve target state-to-action prediction beyond target-only training and beyond simply

increasing the amount of target-task data (Table 1), with dynamics-oriented supervision providing particularly strong gains (Figure 3).

- **Representation compatibility and transfer.** Through matched-data controls and cross-representation experiments, we characterize how auxiliary supervision and its representation affect target performance, state understanding, and transfer (Figure 3, Figure 4, Table 3).

2 Background and problem setup

We consider fine-tuning a model $\mathcal{M}$ for a target task in which each input corresponds to an underlying *state* $s \in \mathcal{S}$ and the model outputs an *action* $a \in \mathcal{A}$. Fine-tuning can thus be viewed as learning a conditional distribution over actions given states, $p(a \mid s)$. For example, in chess, the state is the current board position and the action is the next move. A *representation* $r \in \mathcal{R}$ defines a mapping $\phi_r(s)$ from a state to an input sequence. A state may admit multiple semantically equivalent representations that differ substantially in how readily a model can process and reason over them.

Representation compatibility. Prior work has observed that reasoning performance can vary substantially across representations that encode the same underlying state. For example, in code generation, reasoning in Python rather than natural language has been shown to substantially improve performance [4, 5, 6, 7]. We use *representation compatibility* to describe the extent to which a representation supports effective processing of the underlying state by a model. Operationally, we assess compatibility by comparing downstream task performance across representations (e.g., Figure 2) and by examining reasoning behavior (e.g., Figure 1).

In chess reasoning, we observe a pronounced difference between ASCII, a spatially structured representation, and FEN, a linear symbolic encoding (Figure 2). This pattern is consistent with prior observations in other grid-based environments such as Sokoban, FrozenLake, and Sudoku, where spatial representations facilitate reasoning about structured environments [8]. A complementary qualitative indicator is whether a model transforms one representation into another during reasoning. Figure 1 illustrates this behavior: in an optimal-move prediction task, given a chess position in FEN, Qwen3-8B first converts it into an ASCII representation before reasoning about candidate moves. The mechanisms underlying these preferences, which may reflect factors such as pretraining data or architectural biases, are beyond the scope of this work. We instead focus on how representation compatibility can be diagnosed and leveraged to improve learning and transfer.

Environment-derived auxiliary tasks. In addition to the target objective of predicting the optimal action, we consider auxiliary tasks derived from the underlying environment. These tasks provide complementary supervision about aspects of the state, action space, and environment structure that are not directly specified by the target objective. They include tasks involving state properties, legal actions, and next states, covering static properties of the state, aspects of the action space, and transition dynamics. In particular, next-state prediction models the transition dynamics $p(s' \mid s, a)$, while legal-action prediction characterizes the set of valid actions $\mathcal{A}(s)$. Because the underlying state can be represented in multiple ways, these auxiliary tasks can likewise be instantiated using different representations. We investigate how this choice affects the effectiveness of supervision and its utility to the target task. Although the formulation is compatible with both supervised and reinforcement learning, we instantiate it using offline data and supervised fine-tuning in this work.

3 Representation-aligned auxiliary supervision

We propose a simple method for adapting language models to structured state-to-action tasks using environment-derived auxiliary supervision. The approach uses representation compatibility to determine how this auxiliary supervision is expressed, with the goal of improving target performance. It consists of three steps: (1) diagnosing representation compatibility, (2) constructing auxiliary tasks from environment structure, and (3) jointly training on the target and auxiliary objectives.

3.1 Step 1: Diagnosing representational preferences

Given a set of candidate representations $\mathcal{R}$, we estimate their compatibility with $\mathcal{M}$ by comparing performance across semantically equivalent inputs expressed in different representations. A natural

approach is to evaluate the base model directly on the target task under each representation. However, when the model lacks sufficient capability, performance may be uniformly low, providing little discriminative signal. In such cases, we instead use a more tractable *probe task* that preserves the relevant representational structure of the target task for diagnosing compatibility. For a probe task $\mathcal{T}_{\text{probe}}$ and representation r, we estimate compatibility from the model's performance on probe examples $\mathcal{D}_{\text{probe}}$:

$$C(\mathcal{M}, r; \mathcal{T}_{\text{probe}}) = f(\mathcal{M}, \phi_r, \mathcal{D}_{\text{probe}}),$$

where f denotes the task-specific evaluation metric. We use the highest-scoring representation r^* as the default representation for auxiliary supervision:

$$r^* = \arg\max_{r \in \mathcal{R}} C(\mathcal{M}, r; \mathcal{T}_{\text{probe}}).$$

In addition to task performance, qualitative analysis of reasoning traces provides a complementary signal of compatibility. For example, if a model consistently transforms inputs into an alternative form before reasoning, this provides behavioral evidence that the alternative representation may be more compatible with its reasoning process (e.g., Figure 1). Together, probe task performance and reasoning analysis provide practical signals for assessing representational preferences that guide the design of auxiliary supervision.

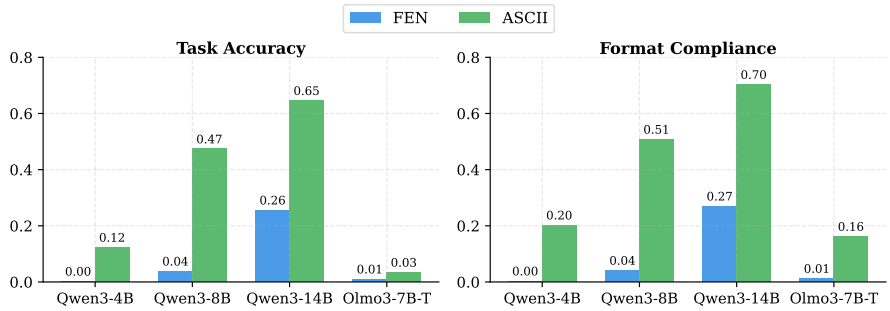

Figure 2: **Representation compatibility in next-state prediction.** Across model scales and families, ASCII inputs consistently yield higher board accuracy (left) and format compliance (right) than FEN prior to domain adaptation. The gap narrows with model size but remains substantial.

As an example, we diagnose compatibility in chess using next-state prediction as a probe, evaluating board accuracy and format compliance under FEN and ASCII (Figure 2). Across models, ASCII yields substantially higher accuracy than FEN, providing a practical signal for selecting the representation in which to express auxiliary supervision. Task accuracy and format compliance are formally defined in Section 4.1, and task performance after fine-tuning is reported in Table 2.

3.2 Step 2: Constructing environment-derived auxiliary tasks

Given a target task $\mathcal{T}$, we construct auxiliary tasks $\mathcal{T}_{\text{aux}}^{(i)}$ from information that can be computed from the underlying environment. These tasks provide additional supervision over the representation, underlying state, or environment dynamics without requiring additional target-task labels. We consider three types of auxiliary tasks:

- **Representation-level tasks:** operations over the input representation itself, such as extracting, concatenating, or manipulating components of the representation.
- **State-property tasks:** tasks that predict properties of the underlying state independently of a particular action, such as object counts, locations, or occupancy.
- **Dynamics tasks:** tasks that characterize how actions interact with the environment, including predicting state transitions and identifying valid actions.

We instantiate auxiliary supervision in chess as *surface*, *static*, and *dynamics* tasks, respectively. Surface tasks operate directly on the board representation, for example by extracting or manipulating ranks. Static tasks query properties of the board, such as piece counts or material balance. Dynamics tasks capture the consequences and constraints of actions, including next-state prediction and legal-move prediction. Each auxiliary task as well as the target task is expressed using the representation

Algorithm 1 Representation-Aligned Auxiliary Supervision

Require: Model $\mathcal{M}$, target task $\mathcal{T}$, auxiliary tasks $\{\mathcal{T}_{\text{aux}}^{(i)}\}$, candidate representations $\mathcal{R}$, probe task $\mathcal{T}_{\text{probe}}$, probe data $\mathcal{D}_{\text{probe}}$, training data $\mathcal{D}$, loss coefficients $\{\lambda_i\}$

// Step 1: Diagnose representation compatibility
for each $r \in \mathcal{R}$ **do**
  $C(\mathcal{M}, r; \mathcal{T}_{\text{probe}}) \leftarrow f(\mathcal{M}, \phi_r, \mathcal{D}_{\text{probe}})$
$r^* \leftarrow \arg\max_{r \in \mathcal{R}} C(\mathcal{M}, r; \mathcal{T}_{\text{probe}})$ ▷ select the most compatible representation
// Step 2: Instantiate tasks in r^*
for each $\mathcal{T}_i \in \mathcal{T} \cup \{\mathcal{T}_{\text{aux}}^{(i)}\}$ **do**
  $\mathcal{D}_{\mathcal{T}_i} \leftarrow \phi_{r^*}(\mathcal{D}_{\mathcal{T}_i})$ ▷ express tasks in r^*
// Step 3: Joint training
for each training step **do**
  Sample batch from $\mathcal{T} \cup \{\mathcal{T}_{\text{aux}}^{(i)}\}$ ▷ joint training on the target and auxiliary tasks
  $\mathcal{L} \leftarrow \mathcal{L}(\mathcal{T}) + \sum_i \lambda_i \cdot \mathcal{L}(\mathcal{T}_{\text{aux}}^{(i)})$
  Update $\mathcal{M}$ on $\mathcal{L}$
return $\mathcal{M}$

r^* selected in Step 1. This is motivated by the hypothesis that auxiliary supervision is most useful when the representation used to provide it is compatible with the model's processing preferences. As our experiments in Section 4 show, the representation choice is particularly important for state-understanding tasks and cross-representation transfer.

3.3 Step 3: Joint training

We jointly train the model on the target task $\mathcal{T}$ and auxiliary tasks $\mathcal{T}_{\text{aux}}^{(i)}$, with inputs expressed in r^*. Let $\mathcal{L}(\mathcal{T})$ and $\mathcal{L}(\mathcal{T}_{\text{aux}}^{(i)})$ denote their respective losses. The training objective is then

$$\mathcal{L} = \mathcal{L}(\mathcal{T}) + \sum_i \lambda_i \cdot \mathcal{L}(\mathcal{T}_{\text{aux}}^{(i)}),$$

where the weights $\lambda_i \geq 0$ control the contribution of each auxiliary task. The target task remains the primary objective, while the auxiliary losses provide additional learning signals derived from the environment. We use joint training so that these signals shape the same model used for the target task rather than requiring auxiliary capabilities to be learned independently and composed at inference time. In our experiments, we compare this joint-training strategy with such sub-skill mixing, where component skills are learned independently and composed at inference time; the latter does not recover the same gains (Table 7).

4 Experiments

4.1 Experimental setup

We study domain adaptation in chess, where board states have precise semantics, environment-derived supervision can be computed automatically, and the same state admits multiple representations. This setting allows us to control both the type of auxiliary supervision and its representation while evaluating performance on state-to-action prediction.

Representations. We consider two representations of chess positions: Forsyth–Edwards Notation (FEN) and ASCII board diagrams. FEN encodes the board as a linear symbolic sequence, where each rank is described by characters for piece types and digits indicating consecutive empty squares. In contrast, ASCII represents the board as an 8×8 spatial grid with piece symbols placed at their corresponding locations and empty squares denoted by dots. The two representations are semantically equivalent and admit exact transformations.

Table 1: **Environment-derived supervision improves optimal-move prediction.** OM trains on optimal-move prediction only; OM + dynamics additionally includes next-state and legal-move supervision. Training representations are shown in parentheses; FEN and ASCII columns report accuracy under each representation. Shading indicates in-distribution performance.

Training	Qwen3-8B		Qwen3-14B		Olmo3-7B-T	
	FEN	ASCII	FEN	ASCII	FEN	ASCII
Baseline	0.0	0.001	0.020	0.041	0.0	0.0
OM only (FEN)	0.181	0.146	0.199	0.161	0.172	0.152
OM + dynamics (FEN)	**0.192**	0.153	**0.223**	0.166	**0.185**	0.077
OM only (ASCII)	0.168	0.181	0.194	0.206	0.157	0.176
OM + dynamics (ASCII)	0.173	**0.201**	0.202	**0.226**	0.185	**0.199**

Tasks. Our primary target task is optimal-move prediction: given a board state, the model predicts the best move as evaluated by a chess engine. This task requires reasoning about the current board state, evaluating candidate moves, and identifying the strongest one. We consider three types of environment-derived auxiliary supervision. *Surface* tasks operate on the representation itself through operations such as extracting, concatenating, or reversing portions of the board encoding. *Static* tasks predict properties of the board state, such as material counts, piece locations, or square occupancy. *Dynamics* supervision captures relationships between states and actions through next-state prediction and legal-move prediction. We also consider *filler* examples, which provide additional optimal-move training data and serve as a control for training-data volume. All tasks use chain-of-thought prompting, with reasoning enclosed in `<think>` tags and final answers in `<answer>` tags.

Data. We construct datasets for optimal-move prediction and environment-derived auxiliary tasks using the Lichess Evaluation Database, which contains diverse chess positions annotated with Stockfish analysis. The primary optimal-move dataset contains 100k board-move pairs, augmented with reasoning traces derived from engine evaluations. For the auxiliary tasks, we construct 200k examples for each supervision condition, with task-specific details described in Appendix A.1. Evaluation is performed on a held-out set of 5k positions disjoint from the training data.

Training and evaluation. Our main experiments use Qwen3-8B and Qwen3-14B [9] as the base models. We additionally evaluate Olmo-3-7B-Think [10] to assess generalization across model families. All models are trained via supervised fine-tuning (SFT) using the objective described in Section 3. Further training details are provided in Appendix A.2.

We report exact-match (EM) accuracy, the fraction of predictions that exactly match the ground truth. For next-state prediction, we additionally report board accuracy (BA), which scores predictions rank-by-rank and assigns $k/8$ for k correctly predicted ranks. For legal-move prediction, we use binary correctness: 1 if the predicted move is legal and 0 otherwise. Format compliance measures the fraction of responses that conform to the expected output format and is reported where relevant as a measure of instruction-following capability.

4.2 Environment-derived supervision for target adaptation

We first ask whether environment-derived auxiliary supervision improves performance on the target task. Table 1 compares optimal-move prediction with and without dynamics supervision, reporting mean accuracy across training seeds. For the evaluated models, adding this auxiliary signal consistently improves optimal-move accuracy. For example, with ASCII training, the additional supervision improves Qwen3-14B accuracy from 0.206 to 0.226 (+9.7%), while with FEN training it improves accuracy from 0.199 to 0.223 (+12.1%). These improvements are obtained with the same 100k optimal-move training examples, indicating that environment-derived supervision provides a useful learning signal beyond the target objective.

The gains in Table 1 could in principle arise simply from exposing the model to more training data, rather than from the information provided by the auxiliary tasks. We therefore compare auxiliary supervision with increasing the amount of target-task data. As shown in Figure 3a, optimal-move performance in ASCII increases from 0.206 with 100k target examples to 0.214 with 200k and

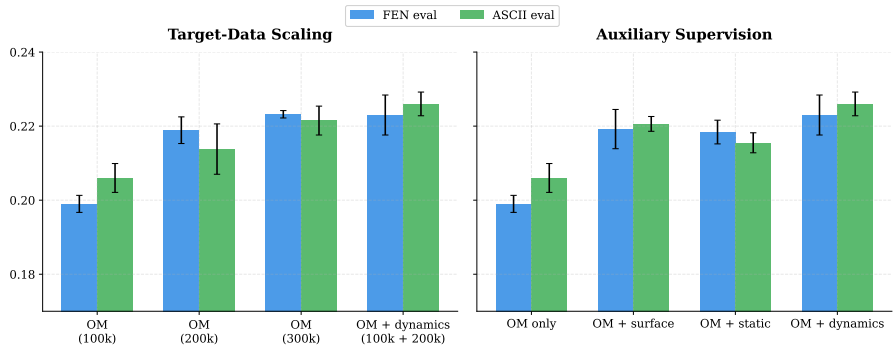

Figure 3: **Environment-derived supervision improves beyond target-data scaling and matched-data controls.** (a) Optimal-move accuracy as the amount of target-task training data increases, compared with 100k target-task examples plus 200k environment-derived dynamics examples. (b) Optimal-move accuracy under matched-data auxiliary supervision conditions. All results use Qwen3-14B, with models trained and evaluated in either FEN or ASCII.

0.222 with 300k. In contrast, combining the original 100k target examples with 200k additional dynamics examples yields an accuracy of 0.226, comparable to that achieved by training on 300k target examples. A similar pattern is observed for FEN. Thus, environment-derived supervision provides an additional learning signal beyond simply increasing the amount of target-task data.

We next examine which properties of the auxiliary supervision contribute to these gains. Figure 3b compares supervision conditions under a matched budget of 200k additional examples. Surface tasks and static state-property tasks produce smaller and less consistent improvements, whereas dynamics supervision yields the largest gain. This pattern suggests that the most useful auxiliary supervision is not merely additional multitask training or exposure to static properties of the state, but supervision that captures relationships between states and actions through transitions and action constraints.

Together, these results show that environment-derived auxiliary supervision can improve adaptation to structured state-to-action tasks beyond target-only training and beyond simply increasing target-task data. Moreover, the matched-data comparisons indicate that the content of the auxiliary supervision matters: supervision that exposes environment dynamics provides particularly strong gains.

4.3 Representation compatibility and transfer

We next examine how the representation used for auxiliary supervision affects what the model learns and how that knowledge transfers. Our probe experiments demonstrate a notable gap for next-state prediction, with models performing substantially better in ASCII than in FEN (Figure 2). For optimal-move prediction, however, the effect is comparatively modest. As shown in Table 1, models trained and evaluated in their respective representations achieve broadly similar performance, with the differences between FEN and ASCII smaller than those observed in state understanding. Thus, representation compatibility does not necessarily translate into large in-distribution gains on the target task, particularly when it can be learned reasonably well from either representation.

The effect becomes more pronounced when evaluating transfer across representations. Models trained with dynamics supervision in ASCII consistently transfer more effectively to FEN than models trained in FEN transfer to ASCII. For example, with Qwen3-14B, OM + dynamics (ASCII) achieves 0.202 when evaluated in FEN, compared with 0.166 for OM + dynamics (FEN) evaluated in ASCII. The former also exceeds the 0.199 achieved by OM-only training in FEN. A similar pattern is observed for Olmo3-7B: OM + dynamics (ASCII) achieves 0.185 when evaluated in FEN, matching the in-distribution performance of FEN, whereas OM + dynamics (FEN) evaluated in ASCII drops substantially to 0.077. This asymmetry suggests that the representation used for auxiliary training affects how effectively the resulting knowledge transfers across semantically equivalent forms.

The contrast between representations is particularly clear for state-understanding tasks. After fine-tuning Qwen3-8B, ASCII→ASCII achieves 93.4% exact match compared with 61.3% for FEN→FEN (Table 2), consistent with the gap observed in the base-model probe. Together with the results above,

Table 2: **Representation on performance and reasoning efficiency.** Each row corresponds to Qwen3-8B fine-tuned on a representation and reasoning length combination. **Avg. Len.** denotes average response length. **Fmt**: format compliance; **EM**: exact match; **BA**: board accuracy.

Training Data	Avg. Len.	Fmt	EM	BA
FEN→FEN (short)	267	1.000	0.539	0.916
FEN→FEN (long)	622	1.000	**0.613**	**0.930**
ASCII→ASCII (short)	258	1.000	**0.934**	**0.988**
ASCII→ASCII (long)	681	0.999	0.713	0.942
FEN→ASCII→FEN (short)	276	1.000	0.824	0.971
FEN→ASCII→FEN (long)	664	1.000	**0.912**	**0.981**

this suggests that representation choice is particularly consequential for learning and transferring structured state knowledge, even when its effect on the target task is comparatively modest.

4.4 Additional analyses

Representation compatibility and reasoning length. In next-state prediction, representation compatibility also changes how additional reasoning affects learning: longer reasoning improves performance under FEN but degrades performance under ASCII (Table 2). Moreover, using ASCII as an intermediate representation in the FEN→ASCII→FEN setting substantially improves performance over direct FEN→FEN, despite identical input and output formats. These results suggest that representation compatibility affects both the amount of reasoning that benefits learning and the usefulness of intermediate representations for structured state transformations.

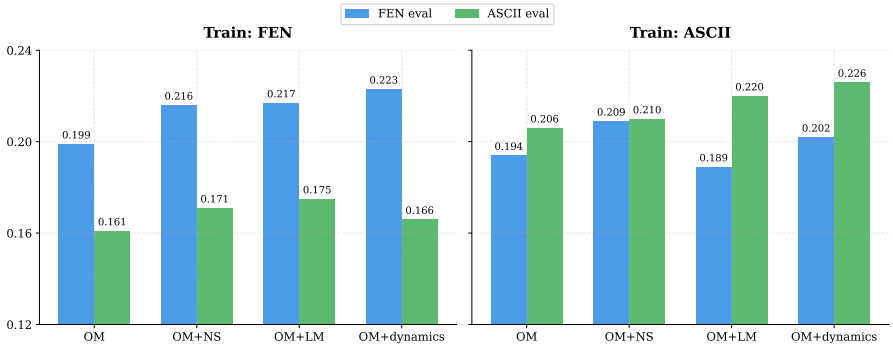

Figure 4: **Contribution of environment-derived auxiliary tasks.** Optimal-move accuracy for models trained with different combinations of tasks. Each task provides consistent gains over target-only training, but their combination (OM + dynamics) yields the most reliable improvements.

Auxiliary task composition. We analyze the contribution of individual auxiliary tasks by augmenting optimal-move prediction (OM) with next-state prediction (NS), legal-move prediction (LM), and their combination (dynamics). Figure 4 reports results for the 14B model. Across both training representations, NS and LM each provide modest but consistent gains over target-only training. However, combining the two tasks yields larger and more consistent improvements, particularly under ASCII training, which achieves the highest in-distribution performance. These results suggest that NS and LM provide complementary supervision about environment dynamics and action constraints, and that jointly exposing the model to both aspects is more effective than either task alone.

Auxiliary supervision on open-ended generation. We additionally evaluate whether the benefits of environment-derived auxiliary supervision extend beyond move prediction to open-ended generation. Given a board state and a move, the model generates an instructive commentary, which we evaluate for factual correctness using a tool-augmented judge. Table 3 reports results under different auxiliary warm-up conditions and commentary representations. Each row specifies the model initialization: the base model, no auxiliary warm-up, or a model first trained on optimal-move prediction and dynamics supervision in FEN or ASCII. The columns then indicate the representation used for commentary fine-tuning and evaluation. See Appendix C for more details.

Table 3: **Auxiliary supervision on commentary generation.** Claim accuracy and error rate for Qwen3-8B across auxiliary warm-up conditions. Columns indicate the representation used for commentary training and evaluation; rows indicate the preceding training or auxiliary warm-up condition. For the OM + dynamics rows, the auxiliary model is trained in the representation shown in parentheses. Shading indicates matched auxiliary and commentary representations.

Initialization	**FEN**		**ASCII**	
	Claim Acc. ↑	Error Rate ↓	Claim Acc. ↑	Error Rate ↓
Baseline (Qwen3-8B)	20.1	61.5	17.1	56.6
No auxiliary supervision	33.7	39.0	44.0	32.2
OM + dynamics (FEN)	50.7 (+17.0)	23.2 (-15.8)	41.7 (-2.3)	35.7 (+3.5)
OM + dynamics (ASCII)	44.0 (+10.3)	36.3 (-2.7)	51.0 (+7.0)	27.9 (-4.3)

Without auxiliary warm-up, training and evaluating commentary generation in ASCII outperforms FEN by 10.3% in claim accuracy, consistent with the model's preference for spatially structured inputs. Auxiliary supervision further improves commentary generation when the warm-up and commentary representations are aligned. Starting from the OM + dynamics (ASCII) checkpoint and training commentary generation in ASCII yields 51.0% claim accuracy and reduces the error rate to 27.9%. The corresponding FEN-aligned condition achieves 50.7% claim accuracy and a 23.2% error rate. We also observe stronger cross-representation transfer from ASCII-aligned auxiliary supervision: compared with no warm-up, ASCII warm-up improves claim accuracy by 10.3% when commentary is trained and evaluated in FEN, whereas FEN warm-up does not improve performance in ASCII. Overall, these results suggest that the benefits of environment-derived supervision extend to open-ended, factually grounded generation, while the representation also affects transfer.

5 Related work

Chess is a common testbed for studying reasoning in language models due to its precise semantics and computable ground truth. Prior work has studied state tracking [11], policy-language training [12], engine distillation [13], strategy-aware fine-tuning [14], and commentary generation [15], as well as strategic reasoning with RL [16, 17] and planning with learned world models [18]. We instead use chess as a controlled setting for studying how representation choice affects domain adaptation.

Model performance is known to depend strongly on input representation. For example, expressing reasoning as code rather than natural language improves accuracy across a range of tasks [4, 5, 6, 7, 19], and structured formats can elicit stronger decision-making [20]. These findings have primarily been exploited through inference-time prompting or representation design. We instead treat representation compatibility as a training-time consideration, using it to select how environment-derived auxiliary supervision is expressed.

Our work also relates to auxiliary-task learning. In RL, environment-derived supervision has been widely used to accelerate learning by developing richer representations [3, 21, 22]. We study a complementary question in language model adaptation: not only which environment-derived signals to supervise, but also which representation should be used to express them.

6 Conclusion

We study domain adaptation in language models through the lens of representation compatibility, showing that the effectiveness of environment-derived auxiliary supervision depends on both the auxiliary task and the representation in which it is expressed. We propose representation-aligned auxiliary supervision, which jointly trains the target task with environment-derived auxiliary tasks expressed in a representation compatible with the model. Across controlled experiments in chess, auxiliary supervision improves target performance beyond target-only training, while representation alignment leads to stronger cross-representation transfer and substantially improves state-understanding tasks. We further find that dynamics-oriented supervision provides the strongest gains, while post hoc composition of independently learned sub-skills provides limited transfer. Overall, our results suggest that choosing how environment structure is represented can be an important part of designing auxiliary supervision for structured domain adaptation.

References

[1] D. Guo, D. Yang, H. Zhang, J. Song, P. Wang, Q. Zhu, R. Xu, R. Zhang, S. Ma, X. Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[2] Z. Shao, P. Wang, Q. Zhu, R. Xu, J. Song, X. Bi, H. Zhang, M. Zhang, Y. Li, Y. Wu, and D. Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

[3] M. Jaderberg, V. Mnih, W. M. Czarnecki, T. Schaul, J. Z. Leibo, D. Silver, and K. Kavukcuoglu. Reinforcement learning with unsupervised auxiliary tasks. In *International Conference on Learning Representations*, 2017.

[4] C. Li, J. Liang, A. Zeng, X. Chen, K. Hausman, D. Sadigh, S. Levine, L. Fei-Fei, F. Xia, and brian ichter. Chain of code: Reasoning with a language model-augmented code emulator. In *International Conference on Machine Learning*, 2024.

[5] H. Puerto, M. Tutek, S. Aditya, X. Zhu, and I. Gurevych. Code prompting elicits conditional reasoning abilities in Text+Code LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, November 2024.

[6] J. Wen, J. Guan, H. Wang, W. Wu, and M. Huang. Codeplan: Unlocking reasoning potential in large language models by scaling code-form planning. In *International Conference on Learning Representations*, 2024.

[7] N. Weir, M. Khalifa, L. Qiu, O. Weller, and P. Clark. Learning to reason via program generation, emulation, and search. *Advances in Neural Information Processing Systems*, 37, 2024.

[8] S. Chen, T. Zhu, Z. Wang, J. Zhang, K. Wang, S. Gao, T. Xiao, Y. W. Teh, J. He, and M. Li. Internalizing world models via self-play finetuning for agentic rl. *arXiv preprint arXiv:2510.15047*, 2025.

[9] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

[10] A. Ettinger, A. Bertsch, B. Kuehl, D. Graham, D. Heineman, D. Groeneveld, F. Brahman, F. Timbers, H. Ivison, J. Morrison, J. Poznanski, K. Lo, L. Soldaini, M. Jordan, M. Chen, M. Noukhovitch, N. Lambert, P. Walsh, P. Dasigi, R. Berry, S. Malik, S. Shah, S. Geng, S. Arora, S. Gupta, T. Anderson, T. Xiao, T. Murray, T. Romero, V. Graf, A. Asai, A. Bhagia, A. Wettig, A. Liu, A. Rangapur, C. Anastasiades, C. Huang, D. Schwenk, H. Trivedi, I. Magnusson, J. Lochner, J. Liu, L. J. V. Miranda, M. Sap, M. Morgan, M. Schmitz, M. Guerquin, M. Wilson, R. Huff, R. Le Bras, R. Xin, R. Shao, S. Skjonsberg, S. Z. Shen, S. S. Li, T. Wilde, V. Pyatkin, W. Merrill, Y. Chang, Y. Gu, Z. Zeng, A. Sabharwal, L. Zettlemoyer, P. W. Koh, A. Farhadi, N. A. Smith, and H. Hajishirzi. Olmo 3. *arXiv preprint arXiv:2512.13961*, 2025.

[11] S. Toshniwal, S. Wiseman, K. Livescu, and K. Gimpel. Chess as a testbed for language model state tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 2022.

[12] X. Feng, Y. Luo, Z. Wang, H. Tang, M. Yang, K. Shao, D. Mguni, Y. Du, and J. Wang. Chessgpt: Bridging policy learning and language modeling. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

[13] A. Ruoss, G. Delétang, S. Medapati, J. Grau-Moya, L. K. Wenliang, E. Catt, J. Reid, C. A. Lewis, J. Veness, and T. Genewein. Amortized planning with large-scale transformers: A case study on chess. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

[14] S. Wang, L. Ji, R. Wang, W. Zhao, H. Liu, Y. Hou, and Y. N. Wu. Explore the reasoning capability of llms in the chess testbed. *arXiv preprint arXiv:2411.06655*, 2024.

[15] J. Kim, J. Goh, I. Hwang, J. Cho, and J. Ok. Bridging the gap between expert and language models: Concept-guided chess commentary generation and evaluation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. Association for Computational Linguistics, 2025.

[16] D. Hwang, H. Lee, J. Choo, D. Park, and J. Park. Can large language models develop strategic reasoning? post-training insights from learning chess. In *The First Workshop on Test-time Scaling and Reasoning Models at the Conference on Language Modeling*, 2025.

[17] K. Kim, K. Wang, Y. Xie, P. Xu, P. Sheng, C. Wei, Z. Wang, J. Shin, P. Viswanath, and S. Oh. Correct answers from sound reasoning: Verifiable process supervision for language models. In *Conference on Language Modeling*, 2026.

[18] J. X. Zhang, Y. Jiang, S. Shang, and S. S. Du. Mastering board games by external and internal planning with language models. *arXiv preprint arXiv:2412.12119*, 2024.

[19] L. Gao, A. Madaan, S. Zhou, U. Alon, P. Liu, Y. Yang, J. Callan, and G. Neubig. Pal: Program-aided language models. In *International Conference on Machine Learning*, 2023.

[20] Y. Hu, X. Wang, W. Yao, Y. Lu, D. Zhang, H. Foroosh, D. Yu, and F. Liu. DeFine: Decision-making with analogical reasoning over factor profiles. In *Findings of the Association for Computational Linguistics*. Association for Computational Linguistics, July 2025.

[21] D. Ha and J. Schmidhuber. Recurrent world models facilitate policy evolution. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

[22] D. Hafner, J. Pasukonis, J. Ba, and T. Lillicrap. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023.

[23] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

[24] T. Dao. Flashattention-2: Faster attention with better parallelism and work partitioning. In *International Conference on Learning Representations*, 2023.

A Experimental details

A.1 Data construction

We sample data from the Lichess Evaluations Database[1], a large collection of chess positions analyzed by the open-source chess engine Stockfish. Each entry contains a position in FEN, candidate moves with centipawn scores (measuring positional advantage), mate evaluations when applicable, and a principal variation (PV) corresponding to the engine's predicted best continuation. For each task and representation, we construct examples consisting of a ground-truth answer and a synthetic reasoning trace tailored to the task. For optimal-move prediction, the traces reason over Stockfish's top three candidate moves. For auxiliary tasks, the traces focus on the aspects of the board representation relevant to producing the correct answer.

Surface tasks operate purely on the board representation. Given a board, we sample one of five templates: extracting a specified rank verbatim; counting occurrences of a specified character within a rank; concatenating two specified ranks; identifying the lexicographically greatest rank when ranks are treated as strings; or reversing the character order of a specified rank. These tasks can be solved by manipulating the board string as text and require no chess-specific knowledge, such as piece identity, move legality, or board geometry.

Static tasks query factual properties of the position without involving transition dynamics. We sample one of five templates: counting a given piece type for a given side; computing the material balance using standard $1, 3, 3, 5, 9$ piece values while ignoring kings; listing all pieces for one side; counting the total number of pieces including kings; or determining which side has more heavy pieces (queens and rooks). These tasks require interpreting the board state but do not require reasoning about how the position may evolve.

Legal-move prediction requires reasoning about the rules that determine valid chess moves. Although chess positions follow well-defined rules, some positions involve additional constraints, including moves that expose the king to check, pinned pieces, check and double-check situations, pawn promotion, and en passant. We therefore oversample positions exhibiting these conditions to ensure adequate coverage of such edge cases during training and evaluation.

A.2 Training details

Table 4: **Supervised fine-tuning hyperparameters.**

	Hyperparameter	Value
Training	Epochs	3
	Batch size	256
	Gradient clipping	1.0
	Precision	bf16
Optimization	Learning rate	1e-5
	LR scheduler	cosine
	Warmup ratio	0.1
	Weight decay	0.01
	Loss coefficients λ_i for all i	1.0

Experiments are conducted primarily on NVIDIA H100 GPUs. Training time varies across different data mixtures, but most experiments complete within one day on 4 H100 GPUs. All models are trained with the AdamW optimizer [23] using $\beta_1 = 0.9$ and $\beta_2 = 0.95$. FlashAttention 2 [24] is enabled for all runs. Table 4 summarizes the hyperparameters used in training.

[1] https://database.lichess.org

Table 5: **Comparison of move quality in win probability loss (Qwen3-8B).** We report mean and median win probability loss (WPL), as a softer measure of move quality based on changes in engine-estimated win probability. Shading indicates in-distribution performance.

Training	Mean WPL ↓		Median WPL ↓	
	FEN	ASCII	FEN	ASCII
OM only (FEN)	21.8 (±27.2)	24.8 (±27.2)	7.8	12.9
OM + dynamics (FEN)	**21.0 (±26.7)**	23.8 (±26.8)	**7.1**	11.1
OM only (ASCII)	22.6 (±26.2)	22.0 (±27.2)	9.4	8.0
OM + dynamics (ASCII)	22.5 (±26.5)	**20.9 (±27.0)**	10.0	**6.4**

B Additional experimental results

B.1 Win probability loss evaluation

Table 1 reports optimal-move accuracy, which treats all non-optimal moves equally. However, chess positions often admit multiple moves with similar quality, making exact-match accuracy a coarse measure of move strength. To complement this evaluation, we measure *win probability loss* (WPL), defined as the absolute difference in win probability between the optimal Stockfish move and the model-selected move:

$$\text{WPL} = \left|\text{WinProb}(\text{Stockfish move}) - \text{WinProb}(\text{Model move})\right|.$$

We convert engine centipawn (cp) scores to win probabilities using the standard Lichess formula[2]:

$$\text{WinProb} = 50 + 50\left(\frac{2}{1 + e^{-0.00368208 \cdot \text{cp}}} - 1\right).$$

Table 5 shows that auxiliary dynamics supervision reduces both mean and median WPL in-distribution for both FEN and ASCII training. The improvements are also reflected in cross-representation evaluation: dynamics supervision in ASCII achieves lower WPL when evaluated in FEN than dynamics supervision in FEN achieves when evaluated in ASCII, consistent with the stronger cross-representation transfer observed in Table 1. Although the reductions in mean WPL are modest, the lower median WPL indicates that auxiliary supervision also improves the general quality of the model's selected move beyond exact optimal-move accuracy.

B.2 Mixture of representations

We examine whether training on multiple representations can improve performance across evaluation settings, and whether exposure to multiple representations alone induces cross-representation reasoning. To evaluate this, we train Qwen3-8B on three variants of next-state prediction: FEN→FEN, ASCII→ASCII, and FEN→ASCII→FEN (F2A2F), using the best-performing reasoning length for each setting identified in earlier experiments. We additionally consider explicit representation-mapping tasks to provide direct supervision for conversions between FEN and ASCII. The resulting training mixtures are:

- **FEN + ASCII (100k):** 50k examples of each single-representation task.
- **FEN + ASCII + F2A2F (150k):** additionally includes 50k mixed-representation examples.
- **FEN + ASCII + F2A2F + Conv (250k):** additionally includes 50k examples for each FEN↔ASCII conversion direction.

Table 6 shows that combining representations improves next-state prediction on both FEN and ASCII relative to the corresponding single-representation baselines. However, simply mixing FEN and ASCII training data does not induce cross-representation reasoning: the FEN + ASCII model achieves only 0.016 EM on FEN→ASCII→FEN, despite strong performance on both single-representation tasks. Adding explicit FEN→ASCII→FEN training substantially improves performance on the mixed-representation task, while also improving the single-representation tasks. Finally, adding explicit

[2] https://lichess.org/page/accuracy

Table 6: **Representation mixture and learning.** Performance on next-state prediction under different data compositions. **Fmt**: format compliance; **EM**: exact match accuracy; **BA**: board accuracy.

Training	FEN→FEN			ASCII→ASCII			FEN→ASCII→FEN		
	Fmt	EM	BA	Fmt	EM	BA	Fmt	EM	BA
Best Baseline	1.000	0.613	0.930	1.000	0.934	0.988	1.000	0.912	0.981
FEN + ASCII	1.000	0.741	0.952	1.000	0.908	0.983	1.000	0.016	0.647
FEN + ASCII + F2A2F	1.000	0.805	0.965	1.000	0.937	0.989	1.000	0.917	0.982
FEN + ASCII + F2A2F + Conv	1.000	**0.860**	**0.974**	1.000	**0.953**	**0.991**	1.000	**0.939**	**0.987**

conversion supervision further improves performance across the evaluation settings. These results suggest that exposure to multiple representations alone is insufficient to induce their coordinated use; structured supervision that explicitly trains the desired cross-representation behavior is needed.

B.3 Limits of sub-skill composition

Our framework jointly trains on the target and auxiliary tasks in an aligned representation. We ask whether the same behavior can instead be obtained by training the constituent skills separately and relying on the model to compose them at inference time. To study this, we decompose the FEN→ASCII→FEN task into four sub-skills: (S1) FEN-to-ASCII conversion, (S2) ASCII-to-FEN conversion, (S3) ASCII-based reasoning, and (S4) FEN-based reasoning. We train Qwen3-8B on mixtures of these sub-skills without exposing it to the full task, and evaluate whether it can compose them at inference time.

Table 7: **Limits of sub-skill composition.** Left: models trained on constituent reasoning skills fail to compose them into the FEN→ASCII→FEN pipeline. Right: adding FEN-to-ASCII conversion to FEN-based reasoning does not induce the use of ASCII as an intermediate representation. Source indicates whether the constituent tasks use the same set of positions (Same) or disjoint sets (Diff).

Training Data	Source	Fmt	EM	BA
FEN→ASCII→FEN	–	1.000	0.912	0.981
S1 + S3 + S2	Same	0.975	0.003	0.644
S1 + S3 + S2	Diff	0.972	0.004	0.633

Training Data	Source	Fmt	EM	BA
ASCII→ASCII	–	1.000	0.934	0.988
S1 + S4	Same	0.985	0.055	0.743
S1 + S4	Diff	0.983	0.060	0.740

Table 7 shows that the constituent skills do not compose effectively at inference time. Models trained on S1 + S3 + S2 achieve near-zero exact match on the full FEN→ASCII→FEN task, despite being trained on each component of the pipeline. Similarly, augmenting FEN-based reasoning with conversion skills does not induce the use of ASCII as an intermediate representation. These results suggest that the desired representation-aligned reasoning structure must be learned end-to-end rather than assembled from independently trained sub-skills. While these results are based on Qwen3-8B and may vary across model scales, they indicate that independently learned constituent skills are insufficient to reliably induce the composed behavior at this scale.

C Commentary generation and evaluation

C.1 Data curation

We collect annotated games from classical chess books and online forums, including Chess Publishing[3]. Because the collected annotations vary in quality and factual accuracy, we apply regex- and LLM-based filtering and discard annotations that contradict objective engine evaluations (e.g., labeling a blunder as a good move). From about 25k collected examples, 10k remain after this stage. We then decompose each annotation into *atoms*, each capturing a single logical unit of observation or opinion about a move, such as a tactical consequence, positional assessment, or comparison to an alternative. We discard annotations containing only a single atom, resulting in 5k training examples.

[3] https://www.chesspublishing.com

Table 8: **Tool suite.** We implement a set of tools that enables the LLM judge to verify the factual correctness of diverse chess claims through chained tool calls.

Category	Tool	Purpose
Position Query	`get_piece_at`(fen, sq)	Return the piece on a given square
	`get_legal_moves`(fen)	List all legal moves in the position
	`is_check`(fen)	Determine whether the side to move is in check
	`get_material`(fen)	Compute material balance / piece counts
	`get_squares`(fen, piece, color)	Find locations of a piece type
Attack / Defense	`get_attacks`(fen, sq)	List squares attacked by a piece on a square
	`get_attackers`(fen, sq)	List pieces attacking a square
	`count_attackers_defenders`(fen, sq)	Count attacking and defending pieces on a square
	`is_pinned`(fen, sq)	Check whether a piece is pinned
	`check_threat`(fen, move)	Verify existence of a tactical or strategic threat
Geometry	`check_ray_alignment`(sq_a, sq_b)	Test alignment along ranks, files, or diagonals
	`get_squares`(fen, piece, color)	Retrieve squares occupied by specific pieces
	`check_open_file`(fen, file, color)	Check if a file is half-open
Variations	`try_variation`(fen, moves)	Simulate a move sequence
Engine Analysis	`eval_move`(fen, move)	Evaluate move quality (e.g., win probability loss)
	`get_engine_eval`(fen, move)	Get Stockfish evaluation of the move and best move
	`get_top_moves`(fen, n)	Return top-n engine-recommended moves
	`compare_moves`(fen, m_a, m_b)	Compare relative quality of two moves

Multi-Tool Verification Example

Generated commentary: "Qe4 centralizes the queen and creates a dual threat: it attacks the bishop on b4 and prepares to infiltrate Black's position via the e-file."

Candidate atom 1: Qe4 centralizes the queen and creates a dual threat: it attacks the bishop on b4 and prepares to infiltrate Black's position via the e-file. [× Incorrect - 1/3 sub-claims verified]

✓ "Qe4 centralizes the queen" — `get_piece_at`("e4"), `get_squares`("queen"), confirm queen moved from b1 to central e4. **Verified.**

× "Qe4 attacks the bishop on b4" — `get_attacks`("e4") → `['b1', 'c2', 'c4', 'd3', ...]` does not include b4. **Incorrect.**

× "Qe4 prepares to infiltrate via the e-file" — `get_attacks`("e4") shows e-file blocked by Black's e5 pawn; no penetration possible. **Incorrect.**

Result: Tool-based verification catches hallucinations (false claims about attacking b4 and e-file infiltration) that pure LLM evaluation would miss.

Finally, an LLM rewrites the atoms into fluent, self-contained commentary. Each training example contains both the list of atoms in `<atom>` tags and the resulting commentary in `<commentary>` tags.

For evaluation, we use 50 examples curated by human players with substantial chess expertise.

C.2 Tool-augmented LLM judge

Our tool-augmented LLM judge evaluates chess commentaries in two stages: (1) atomic claim decomposition and (2) tool-augmented verification.

Stage 1: Atomic claim decomposition. Each generated commentary is decomposed into *candidate atoms*, where each atom represents a position-specific factual statement. To enable fine-grained verification, each atom is further decomposed into one or more verifiable sub-claims, corresponding to individual facts that can be checked independently. For example, the atom "Qe4 centralizes the queen and attacks the bishop on b4" is decomposed into two sub-claims: (1) "Qe4 centralizes the queen" and (2) "Qe4 attacks the bishop on b4". We additionally assign each sub-claim a set of tags (e.g., *quality*, *positional*, *move continuation*) using a lightweight classifier. These tags provide task-specific guidance for the verification.

Stage 2: Tool-augmented verification. Each sub-claim is verified independently by the judge LLM using specialized prompts conditioned on the extracted tags, with mandatory access to chess tools. Verification is grounded in deterministic tool outputs, rather than the judge's own chess knowledge. A candidate atom is marked *correct* only if all of its sub-claims are verified as true; otherwise, it is marked incorrect. We implement 18 specialized tools covering positional queries, attack and defense relations, board geometry, variation legality and resulting positions, and engine-based analysis (Table 8). All tools operate on FEN positions and return structured JSON outputs, enabling fine-grained verification of complex claims.

C.3 Error analysis

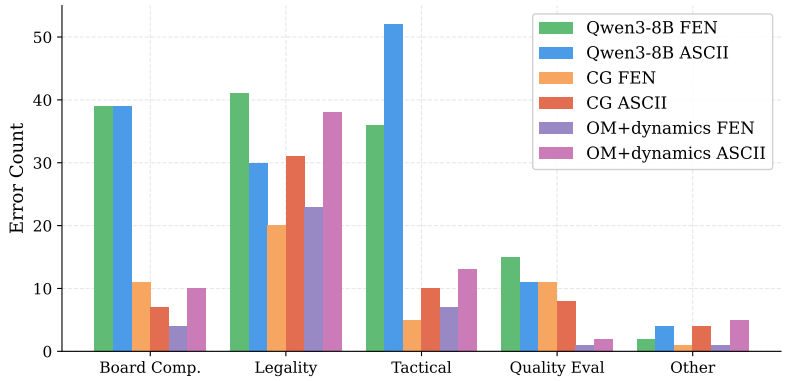

Figure 5: **Error analysis for commentary generation.** Distribution of error types across six models. Board comprehension errors reflect failures in static board parsing; Legality errors reflect failures in multi-step board state tracking; Tactical errors are those regarding downstream consequences of piece localization; Quality errors reflect failures in position understanding.

To understand the sources of factual correctness improvements in Table 3, we analyze the distribution of error types across training conditions (Figure 5). We consider the following error categories:

- **Board comprehension errors**: Failures to accurately identify the static board configuration, including piece identity, location, and structural properties such as open or blocked files.
- **Legality errors**: Failures to track board state across moves, resulting in illegal sequences.
- **Quality evaluation errors**: Incorrect assessments of move strength or positional advantage, including claims that contradict engine analysis or invalid comparisons between positions.
- **Tactical reasoning errors**: Incorrect reasoning about dynamic consequences, such as asserting nonexistent threats or claiming that a move is forced when legal alternatives exist.

We highlight two observations. First, the base model exhibits a high rate of board comprehension errors in both representations, accompanied by downstream legality and tactical reasoning errors. Commentary generation fine-tuning substantially reduces board comprehension and tactical reasoning errors, indicating improved static board understanding across representations and auxiliary supervision conditions. Second, ASCII-aligned auxiliary supervision increases the frequency of illegal move sequences relative to the base model. This shift coincides with more extensive move continuations in generated commentary, suggesting that auxiliary supervision may encourage more detailed but harder-to-maintain multi-step reasoning. Thus, while auxiliary supervision improves factual reasoning overall, tracking board state across longer move sequences remains a limitation.

D Reasoning examples

D.1 Optimal-move prediction

Optimal-move prediction

```
<think>
The current position is: 2rqr1k1/pp1bppb1/3p1np1/4n2p/3NP2P/1BN1BP2/PPPQ2P1/1
    K1R3R w - -.
It is White's turn to move.

Let me analyze White's candidate moves:

First, consider Bh6.
If played, a natural continuation might be:
Bh6
... Nc4
Bxc4
... Rxc4
Bxg7
Overall, White comes out with a slight edge (about +0.46 pawns).
In this variation, White's win rate is about 54.22%.

Another possible choice is Bh6.
If played, a natural continuation might be:
Bh6
... Nc4
Bxc4
... Rxc4
Bxg7
Overall, White comes out with a slight edge (about +0.46 pawns).
In this variation, White's win rate is about 54.22%.

Another possible choice is g4.
If played, a natural continuation might be:
g4
... hxg4
h5
... Nxh5
Bh6
Overall, White comes out with a slight edge (about +0.73 pawns).
In this variation, White's win rate is about 56.68%.

Based on the above analysis, the best move is g4.
It gives White a 56.68% win rate.
</think>

<answer>
g4
</answer>
```

D.2 Surface task

Surface task

```
<think>
The board in FEN is:
8/3k1PK1/7R/8/8/8/4r3/8 w - -

Output the rank field for rank 2 exactly as written, with no other text.
```

```
Rank 2 is the 7th rank field in the board string.
Reading it off gives: 4r3
</think>

<answer>
4r3
</answer>
```

D.3 Static state task

Static state task

```
<think>
The board in FEN is:
2rqr1k1/pp1bppb1/3p1np1/4n2p/3NP2P/1BN1BP2/PPPQ2P1/1K1R3R w - -

Using the standard 1/3/3/5/9 piece values and ignoring kings, which side has
    more material, and by how much?

White material: 38. Black material: 38.
The material is equal.
</think>

<answer>
equal
</answer>
```

D.4 Next-state prediction

FEN → FEN (short reasoning)

```
<think>
I need to apply the move Nxe5 to the position
r2qkbnr/ppp1pppp/8/3Pn3/6b1/2N2N2/PPP2PPP/R1BQKB1R w KQkq -.

Current position analysis:
- It is White's turn.
- The move Nxe5 moves the knight from f3 to e5.
- This captures the knight on e5.

After applying the move:
- The piece layout becomes:
  r2qkbnr/ppp1pppp/8/3PN3/6b1/2N5/PPP2PPP/R1BQKB1R
- It is now Black's turn (b)
- Castling rights: KQkq
- En passant square: -

The resulting board position is:
r2qkbnr/ppp1pppp/8/3PN3/6b1/2N5/PPP2PPP/R1BQKB1R
</think>

<answer>
r2qkbnr/ppp1pppp/8/3PN3/6b1/2N5/PPP2PPP/R1BQKB1R
</answer>
```

FEN → FEN (long reasoning)

```
<think>
I need to apply the move Nxe5 to the position
r2qkbnr/ppp1pppp/8/3Pn3/6b1/2N2N2/PPP2PPP/R1BQKB1R w KQkq -.

Current position analysis:
- It is White's turn.
- The position has 30 pieces on the board.

Current FEN position rank by rank:
Rank 8: r2qkbnr -- 6 piece(s): r, q, k, b, n, r
Rank 7: ppp1pppp -- 7 piece(s): p, p, p, p, p, p, p
Rank 6: 8 -- empty rank
Rank 5: 3Pn3 -- 2 piece(s): P, n
Rank 4: 6b1 -- 1 piece(s): b
Rank 3: 2N2N2 -- 2 piece(s): N, N
Rank 2: PPP2PPP -- 6 piece(s): P, P, P, P, P, P
Rank 1: R1BQKB1R -- 6 piece(s): R, B, Q, K, B, R

Move analysis:
- The White knight on f3 moves to e5.
- Source square f3 (rank 3, file f) will be emptied.
- Destination square e5 (rank 5, file e) will be occupied.
- This captures the black knight on e5.

Affected ranks: rank 5, rank 3
- Rank 5: 3Pn3 -> 3PN3 (e5 changes from n to N)
- Rank 3: 2N2N2 -> 2N5 (N removed from f3)

Resulting position verification:
Rank 8: r2qkbnr
Rank 7: ppp1pppp
Rank 6: 8
Rank 5: 3PN3 [changed]
Rank 4: 6b1
Rank 3: 2N5 [changed]
Rank 2: PPP2PPP
Rank 1: R1BQKB1R

After applying the move:
- The piece layout becomes:
  r2qkbnr/ppp1pppp/8/3PN3/6b1/2N5/PPP2PPP/R1BQKB1R

The resulting board position is:
r2qkbnr/ppp1pppp/8/3PN3/6b1/2N5/PPP2PPP/R1BQKB1R
</think>

<answer>
r2qkbnr/ppp1pppp/8/3PN3/6b1/2N5/PPP2PPP/R1BQKB1R
</answer>
```

ASCII → ASCII (short reasoning)

```
<think>
I need to apply Nxe5 to the current position.
It is White's turn.

Current board:
r . . q k b n r
p p p . p p p p
. . . . . . . .
```

```
. . . P n . . .
. . . . . . . b .
. . N . . N . .
P P P . . P P P
R . B Q K B . R

Move Nxe5: The knight moves from f3 to e5.
It captures the knight on e5.

Updated board:
r . . q k b n r
p p p . p p p p
. . . . . . . .
. . . P N . . .
. . . . . . . b .
. . N . . . . .
P P P . . P P P
R . B Q K B . R

The resulting board position is shown above.
</think>

<answer>
r . . q k b n r
p p p . p p p p
. . . . . . . .
. . . P N . . .
. . . . . . . b .
. . N . . . . .
P P P . . P P P
R . B Q K B . R
</answer>
```

ASCII → ASCII (long reasoning)

```
<think>
I need to apply the move Nxe5 to the current position.

Current board:
r . . q k b n r
p p p . p p p p
. . . . . . . .
. . . P n . . .
. . . . . . . b .
. . N . . N . .
P P P . . P P P
R . B Q K B . R

It is White's turn.
There are 30 pieces on the board.

Pieces on affected ranks:
Rank 5: white pawn on d5, black knight on e5
Rank 3: white knight on c3, white knight on f3

Move analysis:
- Nxe5: White knight (N) on f3 moves to e5.
- This captures the black knight on e5.
- Square f3: N -> .
- Square e5: captured -> N
```

```
Affected ranks: 5, 3
- Rank 5: e5: n -> N
  before: . . . P n . . .
  after:  . . . P N . . .
- Rank 3: f3: N -> .
  before: . . N . . N . .
  after:  . . N . . . . .

Board comparison (before -> after):
Rank 8: r . . q k b n r  (unchanged)
Rank 7: p p p . p p p p  (unchanged)
Rank 6: . . . . . . . .  (unchanged)
Rank 5: . . . P n . . .  ->  . . . P N . . . [changed]
Rank 4: . . . . . . b .  (unchanged)
Rank 3: . . N . . N . .  ->  . . N . . . . . [changed]
Rank 2: P P P . . P P P  (unchanged)
Rank 1: R . B Q K B . R  (unchanged)

Updated board:
r . . q k b n r
p p p . p p p p
. . . . . . . .
. . . P N . . .
. . . . . . b .
. . N . . . . .
P P P . . P P P
R . B Q K B . R
</think>

<answer>
r . . q k b n r
p p p . p p p p
. . . . . . . .
. . . P N . . .
. . . . . . b .
. . N . . . . .
P P P . . P P P
R . B Q K B . R
</answer>
```

FEN → ASCII → FEN (short reasoning)

```
<think>
I need to apply Nxe5 to
r2qkbnr/ppp1pppp/8/3Pn3/6b1/2N2N2/PPP2PPP/R1BQKB1R w KQkq -.
Nxe5: The knight moves from f3 to e5.
It captures the knight on e5.

Updated board in ASCII:
r . . q k b n r
p p p . p p p p
. . . . . . . .
. . . P N . . .
. . . . . . b .
. . N . . . . .
P P P . . P P P
R . B Q K B . R

Converting changed ranks back to FEN:
  Rank 5: . . . P N . . . -> 3PN3
  Rank 3: . . N . . . . . -> 2N5
```

```
Resulting FEN:
r2qkbnr/ppp1pppp/8/3PN3/6b1/2N5/PPP2PPP/R1BQKB1R
</think>

<answer>
r2qkbnr/ppp1pppp/8/3PN3/6b1/2N5/PPP2PPP/R1BQKB1R
</answer>
```

FEN → ASCII → FEN (long reasoning)

```
<think>
I need to apply the move Nxe5 to the position
r2qkbnr/ppp1pppp/8/3Pn3/6b1/2N2N2/PPP2PPP/R1BQKB1R w KQkq -.

First, let me convert the FEN position to an ASCII board.
r2qkbnr/ppp1pppp/8/3Pn3/6b1/2N2N2/PPP2PPP/R1BQKB1R
converted rank by rank:
Rank 8: r2qkbnr -> r . . q k b n r
Rank 7: ppp1pppp -> p p p . p p p p
Rank 6: 8 -> . . . . . . . .
Rank 5: 3Pn3 -> . . . P n . . .
Rank 4: 6b1 -> . . . . . . b .
Rank 3: 2N2N2 -> . . N . . N . .
Rank 2: PPP2PPP -> P P P . . P P P
Rank 1: R1BQKB1R -> R . B Q K B . R

Current board:
r . . q k b n r
p p p . p p p p
. . . . . . . .
. . . P n . . .
. . . . . . b .
. . N . . N . .
P P P . . P P P
R . B Q K B . R

Applying Nxe5: the knight moves from f3 to e5.
It captures the black knight on e5.

Updated board:
r . . q k b n r
p p p . p p p p
. . . . . . . .
. . . P N . . .
. . . . . . b .
. . N . . . . .
P P P . . P P P
R . B Q K B . R

Now converting the ASCII board back to FEN:
Rank 8: r . . q k b n r -> r2qkbnr
Rank 7: p p p . p p p p -> ppp1pppp
Rank 6: . . . . . . . . -> 8
Rank 5: . . . P N . . . -> 3PN3
Rank 4: . . . . . . b . -> 6b1
Rank 3: . . N . . . . . -> 2N5
Rank 2: P P P . . P P P -> PPP2PPP
Rank 1: R . B Q K B . R -> R1BQKB1R

The resulting position:
```

```
r2qkbnr/ppp1pppp/8/3PN3/6b1/2N5/PPP2PPP/R1BQKB1R
</think>

<answer>
r2qkbnr/ppp1pppp/8/3PN3/6b1/2N5/PPP2PPP/R1BQKB1R
</answer>
```

D.5 Legal-move prediction

Legal-move prediction

```
<think>
White pieces - scan each rank:
  Rank 4: 3NP2P. White: N(d4), P(e4), P(h4).
  Rank 3: 1BN1BP2. White: B(b3), N(c3), B(e3), P(f3).
  Rank 2: PPPQ2P1. White: P(a2), P(b2), P(c2), Q(d2), P(g2).
  Rank 1: 1K1R3R. White: K(b1), R(d1), R(h1).
White has: Knight(d4), Pawn(e4), Pawn(h4), Bishop(b3), Knight(c3), Bishop(e3),
     Pawn(f3), Pawn(a2), Pawn(b2), Pawn(c2), Queen(d2), Pawn(g2), King(b1),
    Rook(d1), Rook(h1).

Consider Bishop on e3.
Bishops move diagonally.
  f4(✓), g5(✓), h6(✓).
  d4(own knight, stop).
  f2(✓), g1(✓).
  d2(own queen, stop).

Select Bf4 - improves bishop activity.
</think>

<answer>Bf4</answer>
```

D.6 Representation mapping

FEN-to-ASCII mapping

```
<think>
I need to convert the FEN position to ASCII format.
The FEN position is: 2rqr1k1/pp1bppb1/3p1np1/4n2p/3NP2P/1BN1BP2/PPPQ2P1/1
    K1R3R

In FEN notation:
- Uppercase letters represent white pieces (K, Q, R, B, N, P)
- Lowercase letters represent black pieces (k, q, r, b, n, p)
- Numbers represent consecutive empty squares

Converting rank by rank (from rank 8 to rank 1):
Rank 8: 2rqr1k1 → . . r q r . k .
Rank 7: pp1bppb1 → p p . b p p b .
Rank 6: 3p1np1 → . . . p . n p .
Rank 5: 4n2p → . . . . n . . p
Rank 4: 3NP2P → . . . N P . . P
Rank 3: 1BN1BP2 → . B N . B P . .
Rank 2: PPPQ2P1 → P P P Q . . P .
Rank 1: 1K1R3R → . K . R . . . R

The complete ASCII representation:
</think>
```

```
<answer>
...
. B N . B P . .
P P P Q . . P .
. K . R . . . R
</answer>
```

E Limitations

While our results demonstrate the effectiveness of representation-aligned auxiliary supervision for domain adaptation, several limitations remain. First, our study focuses on chess, a controlled domain with precise semantics and well-defined representations. Although this setting allows us to isolate the effect of representation choice from the training objective, it remains to evaluate how well our findings generalize to less structured domains. Second, we operationalize representation compatibility using empirical proxies such as task performance and reasoning behavior. While these provide useful signals, they do not constitute a formal measure and may depend on the choice of probe tasks. Developing more principled measures of compatibility, e.g., based on internal representations, could strengthen the framework and improve its applicability across domains. Finally, our framework assumes access to auxiliary supervision derived from environment structure. In domains lacking such structure, applying our approach may require additional modeling effort. Determining how to construct useful auxiliary supervision in such settings is an important direction for future work.

F Broader impacts

This work studies how representation choices affect domain adaptation in language models and how environment-derived auxiliary supervision can be aligned with those representations. More effective adaptation methods may facilitate the development of models for specialized domains, including settings where reliable structured reasoning is important. At the same time, improved adaptation does not guarantee robustness or factual reliability, particularly when auxiliary supervision or the underlying environment is imperfectly specified. We view this work as a step toward more principled language model adaptation and emphasize careful evaluation before applying such methods to consequential tasks.