Bridging the Gap between Expert and Language Models: Concept-guided Chess Commentary Generation and Evaluation

Jaechang Kim[1], **Jinmin Goh**[2,3], **Inseok Hwang**[1,2], **Jaewoong Cho**[4], **Jungseul Ok**[1,2]

[1]Graduate School of AI, POSTECH
[2]Department of Computer Science and Engineering, POSTECH
[3]Department of Chemical Engineering, POSTECH
[4]KRAFTON

Abstract

Deep learning-based expert models have reached superhuman performance in decision-making domains such as chess and Go. However, it is under-explored to explain or comment on given decisions although it is important for model explainability and human education. The outputs of expert models are accurate, but yet difficult to interpret for humans. On the other hand, large language models (LLMs) can produce fluent commentary but are prone to hallucinations due to their limited decision-making capabilities. To bridge this gap between expert models and LLMs, we focus on chess commentary as a representative task of explaining complex decision-making processes through language and address both the generation and evaluation of commentary. We introduce Concept-guided Chess Commentary generation (`CCC`) for producing commentary and GPT-based Chess Commentary Evaluation (`GCC-Eval`) for assessing it. `CCC` integrates the decision-making strengths of expert models with the linguistic fluency of LLMs through prioritized, concept-based explanations. `GCC-Eval` leverages expert knowledge to evaluate chess commentary based on informativeness and linguistic quality. Experimental results, validated by both human judges and `GCC-Eval`, demonstrate that `CCC` generates commentary which is accurate, informative, and fluent.

1 Introduction

Artificial intelligence (AI) has achieved superhuman performance in various decision-making tasks, particularly in abstract strategy games like chess and Go. Milestones such as Deep Blue's victory over the world chess champion (Campbell et al., 2002) and AlphaGo's defeat of top human Go players highlight AI's capabilities in solving complex problems (Silver et al., 2017). While these expert models deliver highly accurate decisions, they often lack interpretability, which is critical for human education and trust in AI systems. The strategic insights and rationales behind decisions are often explained through natural language commentary (Chernev, 2003; Polgar, 2014). Large language models (LLMs) exhibit their outstanding performance in generating fluent natural language. However, LLMs often struggle with hallucinations due to their limited capability in complex decision-making and lack of domain-specific knowledge.

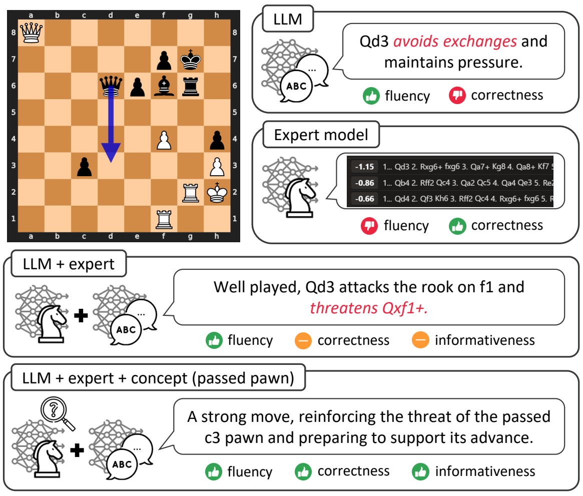

a b c d e f g h
8 7 6 5 4 3 2 1
LLM
Qd3 *avoids exchanges* and maintains pressure.
ABC
fluency
correctness
Expert model
-1.15 1... Qd3 2. Rxg6+ fxg6 3. Qa7+ Kg8 4. Qa8+ Kf7
-0.86 1... Qb4 2. Rff2 Qc4 3. Qa2 Qc5 4. Qa4 Qe3 5. Re
-0.66 1... Qd4 2. Qf3 Kh6 3. Rff2 Qc4 4. Rxg6+ fxg6 5. R
fluency
correctness
LLM + expert
Well played, Qd3 attacks the rook on f1 and *threatens Qxf1+.*
fluency
correctness
informativeness
LLM + expert + concept (passed pawn)
A strong move, reinforcing the threat of the passed c3 pawn and preparing to support its advance.
fluency
correctness
informativeness

Figure 1: *Comparison of chess commentary generation methods.* The red color indicates incorrect information.

We aim to bridge the gap between expert and language models. Specifically, we focus on the task of chess commentary generation to explain given decisions. Although chess is a resourceful testbed with extensive dataset and study (Zang et al., 2019; Lee et al., 2022; Feng et al., 2023), the chess commentary generation has two main challenges: (i) producing accurate and insightful commentary, which requires deep chess knowledge and linguistic ability, and (ii) developing evaluation metrics to assess commentary quality, which is overlooked in previous research.

Although language models can generate fluent natural language, they lack the chess-specific

---

Correspondence to: Jungseul Ok (jungseul@postech.ac.kr)

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), pages 9497–9516
April 29 - May 4, 2025 ©2025 Association for Computational Linguistics

knowledge required for chess commentary generation. Even a model (Feng et al., 2023) trained on chess-related data struggles in reasoning and understanding complex positions. One promising approach is to integrate expert models with language models. However, prior attempts (Zang et al., 2019; Lee et al., 2022) directly feeding the decision-making process of expert models to language models are inadequate because the decision-making process is hard to interpret for language models.

To address them, we introduce an effective approach using concept-based explanations of expert models. By extracting and prioritizing concepts that the expert model focuses on, we guide the language model to concentrate on the most important aspects of the game. This results in commentary that is both linguistically fluent and strategically insightful. Figure 1 illustrates previous approaches and our approach. Our experiments demonstrate that our approach achieves human-level correctness in commentary generation, while outperforming baselines and human-generated comments in informativeness (relevance, completeness) and linguistic quality (clarity, fluency).

Evaluating chess commentary generation is another challenge task. Previous works (Jhamtani et al., 2018; Zang et al., 2019; Lee et al., 2022) rely on similarity-based metrics such as BLEU, which are insufficient due to the inherently diverse nature of commentary. Different commentators may focus on distinct aspects of a position, such as attack strategies or defensive plans. In tasks like summarization or translation, which share the same challenges, LLM-based evaluation metrics (Zhong et al., 2022; Liu et al., 2023) are proposed to assess multiple dimensions. We adopt `G-Eval` (Liu et al., 2023) by incorporating expert model guidance for chess knowledge. We measure the commentary's informativeness (relevance, completeness) and linguistic quality (clarity, fluency). Through our experiments, we show that our proposed method correlates well with human judgments, offering a more reliable metric for commentary evaluation.

Our contributions are as follows:

- We propose an approach that integrates expert models with LLMs through concept-based explanations, facilitating transparent decision-making in chess commentary generation.
- We develop a prioritization mechanism that highlights important concepts and an LLM inference technique that enables the model to understand moves with concept guidance.
- We introduce and validate an LLM-based evaluation metric to assess the quality of chess commentary across multiple dimensions.

2 Related work

Chess commentary generation Chess commentary generation is generating a comment for a chess move. Jhamtani et al. (2018) first address the task by utilizing web-crawled data to form a chess commentary dataset, framing commentary generation as a sequence prediction problem. Building on this, Zang et al. (2019) incorporate domain-specific chess knowledge using internal chess models, improving quality and contextual relevance of generated comments. Lee et al. (2022), integrate BART (Lewis et al., 2020) and an external chess engine for more reliable move evaluation. However, their system classifies moves into predefined categories (e.g., excellent, good, inaccuracy, mistake, blunder), without deeper understanding of the model decision-making process. In contrast, we leverage concept-based explanation to extract chess concepts from an expert model to understand the rationale behind the decision.

Not limited to chess commentary, Feng et al. (2023) fine-tune an LLM on chess-related data, to leverage chess skills, not only the linguistic ability. However, we demonstrate that its understanding of chess knowledge is inferior to `GPT-4o` (OpenAI, 2023) (Section 4.4).

Concept-based explanation in chess Concepts are high-level abstractions commonly shared within a community, enabling efficient communication. In chess, concepts such as "king safety" (i.e., all potential threats against the king) condense complex strategies into understandable terms, allowing players to communicate effectively without lengthy explanations. These concepts are understandable to both humans and language models, serving as a bridge between human intuition and neural networks. Concept-based explanations aim to make a model interpretable by aligning its internal decision-making process with these shared concepts, assuming that such concepts are linearly embedded in the representation space (Kim et al., 2018; Alain and Bengio, 2016; McGrath et al., 2022). This assumption is validated in chess domains (Pálsson and Björnsson, 2023; McGrath et al., 2022) for chess expert models like

`Stockfish` (Romstad et al.), `AlphaZero` (Silver et al., 2018), and their open-source versions, such as `LeelaChessZero` (Authors, 2024).

Prioritization of concepts Yuksekgonul et al. (2023) train a post-hoc concept bottleneck model, and the classifier following the concept bottleneck model is directly interpreted as the global importance of concepts for a class. However, they focus on finding global concept importance per class, without addressing the varying significance of concepts for individual inputs. We address prioritization of concepts for individual inputs, or local importance, to determine the influence of each concept in specific situations.

Evaluation of natural language generation Classical evaluation metrics for natural language generation (NLG) are based on similarity. Common metrics are BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004). However, these metrics fail to assess content quality (Reiter and Belz, 2009) and syntactic correctness (Stent et al., 2005), and are insufficient to measure the reliability of NLG systems. Zhang* et al. (2020); Zhao et al. (2019) compare the similarity in the text embedding space, to adequately measure semantic similarity.

Recently, beyond the similarity, Yuan et al. (2021); Mehri and Eskenazi (2020) assess generated natural language in multiple dimensions, and Zhong et al. (2022); Liu et al. (2023) evaluate in multiple dimensions using language models. The idea of using LLMs for evaluation is common, and the evaluation methods are known to be aligned with human evaluation, sometimes more than agreements among human evaluators (Rafailov et al., 2024; Chen et al., 2023). The LLM-based evaluators are focused on summarization and translation tasks. Regarding evaluation in chess commentary, they still lack the domain-specific knowledge required for evaluating chess commentary.

Evaluating chess commentary is challenging due to its diverse nature, where commentaries on the same move may vary significantly depending on the focus, such as attack strategies, defensive plans, or comparison with other moves. Chess knowledge is essential for evaluating the correctness and relevance of these commentaries. Previous chess commentary researches (Jhamtani et al., 2018; Zang et al., 2019; Lee et al., 2022) use classical metrics such as BLEU, ROUGE, or perplexity, but these metrics fall short for chess commentary, as they do not evaluate with domain-specific knowledge. While manual evaluation by human experts remains ideal, we propose an automatic evaluation method leveraging an LLM with chess knowledge.

3 Method: generation and evaluation

We propose two methods to address chess commentary generation (Section 3.1) and chess commentary evaluation (Section 3.2).

3.1 Concept-guided commentary generation

We propose Concept-guided Chess Commentary generation (`CCC`), which is a method for generating chess commentary by leveraging a chess expert model and its concept-based explanations. The method involves two key steps: 1) extracting concept vectors from a chess expert model (Section 3.1.1); and 2) generating commentary via an LLM using prioritized concepts that explain the given position and movement (Section 3.1.2). Figure 2 provides an overview of the proposed method.

3.1.1 Concept vector extraction

To make a chess expert model interpretable, we extract concept vectors that correspond to key concepts in chess. We follow a common approach (Kim et al., 2018; Yuksekgonul et al., 2023) involving two steps: preparing a dataset for concept learning and extracting concept vectors by training a linear classifier. The concepts we focus on are adopted from `Stockfish 8`, a classical chess engine that can evaluate positions for their relevance to specific concepts (see Table 6). We collect 200,000 chess positions from the Lichess open database [1] and use `Stockfish 8` to assign a score reflecting how strongly each position relates to these concepts. We then label the top 5% of positions with the highest scores as positive samples and the bottom 5% with the lowest scores as negative samples. This process results in a dataset of 20,000 positions for each concept, split equally between positive and negative samples. We employ `LeelaChessZero T78`, an open-source neural network-based chess model similar to `AlphaZero` for extracting concept vectors. For the representation space, we use the final layer before policy and value heads (layer 40). We then train a linear Support Vector Machine (SVM) (Cortes and Vapnik, 1995) to classify these samples. The resulting normal vector of the SVM classification boundary serves as the concept vector, and the distance from

---

[1] `https://database.lichess.org/#evals`

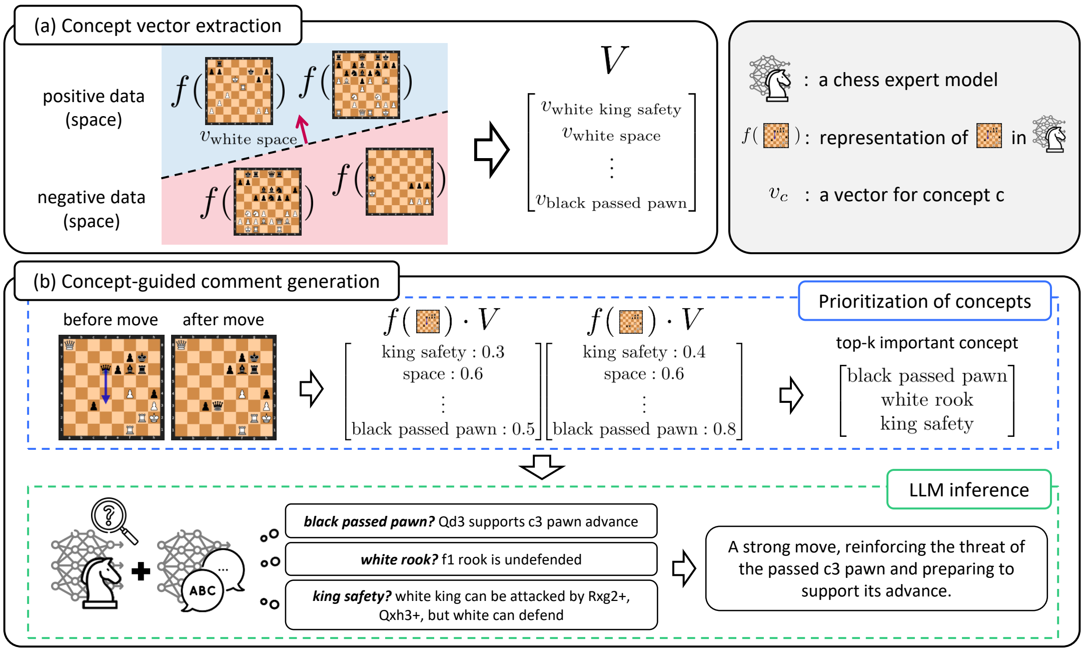

(a) Concept vector extraction
positive data (space)
negative data (space)
$v_{\text{white space}}$
V
$v_{\text{white king safety}}$
$v_{\text{white space}}$
$v_{\text{black passed pawn}}$
: a chess expert model
$f(\ )$: representation of in
v_c : a vector for concept c
(b) Concept-guided comment generation
before move
after move
$f(\ )\cdot V$
king safety : 0.3
space : 0.6
black passed pawn : 0.5
king safety : 0.4
space : 0.6
black passed pawn : 0.8
Prioritization of concepts
top-k important concept
black passed pawn
white rook
king safety
LLM inference
ABC
black passed pawn? Qd3 supports c3 pawn advance
white rook? f1 rook is undefended
king safety? white king can be attacked by Rxg2+, Qxh3+, but white can defend
A strong move, reinforcing the threat of the passed c3 pawn and preparing to support its advance.

Figure 2: *Overview of* `CCC`*, consists of (a) extracting concept vectors and (b) generating concept-guided commentary.*

this boundary determines the concept score for any input position. This score quantifies how strongly a given board state aligns with the extracted concept.

3.1.2 Chess comment generation with an expert model and extracted concepts

Prioritization of concepts Given a chess position and a specific move, our goal is to identify the concepts most relevant to explaining that movement. For the chess position, we compute the score for each concept by taking the dot product between the expert model representation of the position and the extracted concept vectors. These concept scores reflect how strongly each concept is reflected in the current position. To prioritize concepts, we compare the concept scores before and after the move. By analyzing the differences between pre-move and post-move scores, we identify which concepts are most influenced by the move. This allows us to assign priority to the concepts that explain the impact of the move.

Commentary generation via LLM We generate chess commentary using an LLM and a chess expert model. Although a language model understands chess-specific notations and terms, it lacks the ability to perform chess-specific reasoning and complex analysis, which can result in hallucination. By integrating chess expert model output, the LLM determines whether to focus on advantageous aspects or disadvantageous aspects. However, since the chess expert model output is based on scalar values, it still generates incorrect comments. Concept-based explanation guides the LLM to focus on critical aspects. Figure 3 is a typical example of a concept-guided comments.

To enhance the reasoning ability of LLM, we employ few-shot prompting, Chain-of-Thought (CoT) prompting (Wei et al., 2022), and chess-specific information. This approach provides the LLM with a deeper understanding of chess positions, and prevents potential use of wrongly prioritized concepts. Additionally, we enumerate all existing attacks towards opponent pieces to prevent mentioning of non-existing pieces or illegal moves.

3.2 Automatic evaluation of commentary

Our evaluation approach, termed `GCC-Eval`, modifies and extends `G-Eval` to better address the specific challenges of evaluating chess commentary. The core components of `GCC-Eval` are: (i) Multi-dimensional evaluation by an LLM. (ii) Expert model evaluation for chess knowledge. (iii) Auto-CoT for score-only output. (iv) weighted summation for non-integer scores. Note that our contributions are on the first and second aspects to ensure accurate chess commentary evaluation, focusing on informativeness and linguistic quality.

Evaluation dimensions The evaluation covers four dimensions: relevance, completeness, clarity, and fluency. While clarity and fluency are general linguistic measures, relevance and completeness require a deep understanding of chess. To address this, we employ an expert model to augment the

LLM's capabilities when scoring relevance and completeness. This integration ensures that the commentary is not only linguistically sound but also informative from a domain-expert perspective.

LLM generates a single-integer evaluation between one to five after Auto-CoT. The probability of each score represents the confidence of LLM for the score. We use a weighted summation of score probabilities as follows:

$$\mathrm{score}(x) = \sum_{s \in \{1,2,3,4,5\}} s \times p(s|x). \quad (1)$$

This method allows to produce non-integer scores in LLM-based evaluation, capturing subtle nuances in the evaluation that would be missed by integer-only scoring schemes. The scoring prompts, including the expert evaluation and Auto-CoT reasoning, are described in Appendix A.

4 Experiments

4.1 Experimental settings

Dataset We evaluate our model using *Chess Commentary* dataset introduced by Jhamtani et al. (2018). This dataset contains full chess games accompanied by user-generated commentary on specific moves, collected from an online chess forum[2]. Following the train/valid/test split introduced by Jhamtani et al. (2018), we use only the test set for our experiments. Since the absence of preprocessing code, we manually align the raw data with pre-processed data to ensure fair comparison with `GAC`. Additionally, we exclude comments that covering multiple moves for simplicity in analysis.

Baselines We compare the experimental results within several methods:

- **reference**: These are reference texts from the `GameKnot` dataset.
- **`GAC`** (Jhamtani et al., 2018): An LSTM model trained on the `GameKnot` dataset for generating chess commentary.
- **`GPT-4o`** (OpenAI, 2023): The unmodified version of `GPT-4o`, accessed via OpenAI API, with a temperature setting of 0.1 to avoid noisy outputs. For detailed discussion of comparison of LLMs, refer to Section 4.4.
- **`GPT-4o + expert`**: This is the same `GPT-4o` model but augmented with evaluations from a chess expert model. Note that Lee et al. (2022) use BART with a chess expert model and `GPT-4o + expert` is superior because it uses more powerful language model and a sufficient expert model.

Reproducibility For better reproducibility, we release our source code [3].

4.2 Human evaluation

Human evaluation settings We conducted a manual human evaluation to assess the quality of the generated chess commentary. All participants were fluent in English, both in reading and writing, and possessed sufficient chess knowledge to evaluate the comments reliably. Specifically, we recruited five participants from the university community and social networks. Each participant held a Chess.com rapid rating above 1500, placing them within the top 0.49% among chess players[4]. The average rating among participants was 1776. Participants were compensated with an amount equivalent to 73 USD, which matched the local minimum wage for 10 hours of work. The evaluation plan was approved by our university's Institutional Review Board (IRB number: PIRB-2024-E028). All participants were informed about the purpose of the research and the intended use of the evaluation data, and consent was obtained before participation. The evaluation results were anonymized and did not contain any personal information. To enhance the quality of the evaluation process, we conducted two rounds of pilot tests with two participants who were not involved in the main evaluation. These pilot studies helped us clarify potentially ambiguous expressions, adjust the response format, estimate the time required, and optimize the timing of mandatory breaks based on participant feedback.

The evaluation followed a within-participant design. For each move, participants evaluated commentary generated by five methods (four baselines and `CCC`), with the order of methods randomized. A total of 50 moves were evaluated, resulting in 250 comments assessed by each participant. The evaluation took approximately four hours to complete, including breaks. Participants were required to take breaks after every 50 comments (i.e., approximately 10-minute breaks after 40 minutes of evaluation) to ensure sustained focus. During the evaluation, participants were presented with

---

[2] `https://gameknot.com/`

[3] `https://github.com/ml-postech/concept-guided-chess-commentary`

[4] retrieved in Oct 2024, from `https://www.chess.com/leaderboard/live/rapid`

Comment generation methods	**Correctness**	**Relevance**	**Completeness**	**Clarity**	**Fluency**	**Words per comment**
Reference	0.62	*0.52	*0.30	*0.60	*0.62	*15.6
`GAC` (Jhamtani et al., 2018)	**0.63**	*0.46	*0.15	*0.66	*0.64	*8.9
`GPT-4o`	*0.36	*0.49	*0.40	$\underline{^*0.72}$	*0.84	*27.1
`GPT-4o + expert`	*0.43	$\underline{^*0.56}$	$\underline{^*0.49}$	$\underline{^*0.72}$	$\underline{^*0.85}$	*26.2
`GPT-4o + expert + concept` (`CCC`, ours)	$\underline{0.60}$	**0.67**	**0.59**	**0.80**	**0.91**	28.5

Table 1: *Average scores of human evaluation.* Bold and underlined text indicate the best and second-best methods in each column, respectively. Numbers are rescaled to the range $[0, 1]$. * denotes the difference from `CCC` is significant (i.e., p-value < 0.001).

Types of errors	`GPT-4o`	`GPT-4o + expert`	`GPT-4o + expert + concept`
Referring illegal move or non-existing pieces	0.46	0.28	0.20
Wrong understanding of tactical/immediate advantage	0.46	0.40	0.26
Wrong understanding of positional/long-term advantage	0.28	0.26	0.28
Wrong evaluation of the move/position	0.32	0.30	0.34

Table 2: *Error rates in different causes of incorrectness.* Note that the questions allow multiple answers per question. Error types in the lower rows require more comprehensive reasoning.

Metrics	κ
Correctness	0.5393
Relevance	0.2448
Completeness	0.2449
Clarity	0.1782
Fluency	0.2328

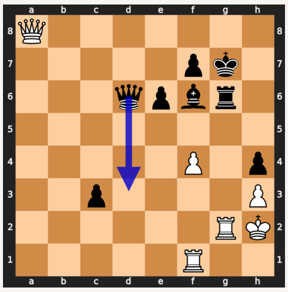
Table 3: *Inter-annotator agreements of human evaluation measured by Fleiss' kappa(κ).*

a chessboard displaying a specific move, marked with a blue arrow. Alongside the moves, the corresponding commentaries were provided. Each participant was asked to rate the commentary across six questions: five evaluation metrics and one question for categorizing the type of incorrectness when applicable. The evaluated metrics were correctness, relevance, completeness, clarity, and fluency. Relevance and completeness assessed how informative and insightful the comment was, while clarity and fluency evaluated its linguistic naturalness. Relevance, completeness, clarity, and fluency were assessed using a five-point Likert scale, while correctness was evaluated using a three-point Likert scale, as the correctness of a comment was closer to a binary decision rather than a scaled question. For clear presentation, the scores were rescaled to a range of 0 to 1. Appendix C provides the instructions and questions used for human evaluation.

Main results Table 1 presents the results of the human evaluation. Our proposed method, `CCC`, achieves the highest scores in all metrics except correctness, where it ranks second. Also, `CCC` outperforms the reference comments in every metric except correctness, and the correctness is also comparable to the reference. The reference comments, collected from online sources, often contain grammatical mistakes and informal language, underscoring the limitation of similarity-based evaluation metrics. This highlights the need for evaluation metrics beyond similarity, especially when the quality of the reference comments is suboptimal. The use of expert models and concept guidance contribute significantly to the overall performance improvement, as evidenced by the higher scores across most metrics. While `GPT-4o + expert` shows only a slight improvement in correctness, it generates more detailed explanations, which in some cases lead to minor factual inaccuracies in the details, as illustrated in Figure 3. Although `GAC` exhibits the highest correctness slightly outperforming `CCC`, we observe that `GAC`'s higher correctness comes at the cost of lower details of the explanations; their explanations tend to be brief and thereby less informative in general, leading to lower scores in completeness and shorter the comment lengths.

Detailed analysis Table 2 provides a detailed analysis of the types of errors. The usage of the expert model and concept reduces simple errors, but errors requiring comprehensive understanding remain within the margin of error.

To validate the consistency of the human evaluation, we calculate inter-annotator agreement using



- **Reference comment**
  Exchanging rook will give more oportunities to attack the king and find the mate.
- `GAC` (**Jhamtani et al., 2018**)
  Black moves his queen to a safe square.
- `GPT-4o`
  Qd3 is a solid move, *avoiding exchanges* and maintaining pressure. Black's queen is now targeting several critical squares and creating threats. White needs to find accurate defensive moves to stay in the game.
- `GPT-4o + expert`
  well played, Qd3 attacks the rook on f1 and *threatens Qxf1+*. White now faces a tough defense, expected move is g2-g6 to create counter-play.
- `GPT-4o + expert + concept` (Pawns, Black Passedpawns, White Kingsafety)
  Qd3 is a strong move, reinforcing the threat of the passed c3 pawn and preparing to support its advance. It also keeps pressure on White's position, maintaining the initiative.

Figure 3: *Examples of generated comments. Red text* denotes incorrect information, and blue text denotes important concepts and affected counterparts.

Metrics	Correctness		Relevance		Completeness		Clarity		Fluency	
	ρ	τ	ρ	τ	ρ	τ	ρ	τ	ρ	τ
BLEU-1	0.17	0.10	-0.06	0.02	-0.25	-0.07	-0.08	0.01	-0.36	-0.16
ROUGE-1	0.04	-0.01	-0.19	-0.08	-0.29	-0.18	-0.18	-0.14	-0.29	-0.17
ROUGE-2	0.15	0.03	-0.10	-0.03	-0.16	-0.05	-0.02	0.03	-0.15	0.00
ROUGE-L	0.08	0.01	-0.18	-0.08	-0.29	-0.18	-0.16	-0.14	-0.29	-0.17
`GCC-Eval`	-	-	0.40	0.24	0.56	0.39	0.44	0.23	0.55	0.38

Table 4: *Correlations between human and automatic evaluations.* ρ and τ denotes Pearson correlation and Kendall's tau correlation, respectively.

Fleiss' Kappa (Fleiss and Cohen, 1973) of ranks across different methods. Table 3 reports the agreement of the participants. The agreement for correctness is 0.54, indicating moderate agreement. This is notably higher than for other metrics, suggesting that correctness is more indisputable for chess experts, compared to more subjective qualities like relevance or fluency.

4.3 Automatic evaluation

To perform an automatic evaluation of generated chess commentaries, we employ our proposed metric, `GCC-Eval`. This metric is designed to assess both linguistic quality and domain-specific relevance in chess commentary. To validate its reliability, we calculate the correlation between `GCC-Eval` scores and human evaluations using the same dataset from prior human evaluation studies. As shown in Table 4, `GCC-Eval` consistently shows a higher correlation with human assessments across all evaluation criteria compared to traditional metrics, such as BLEU and ROUGE, which rely on surface-level similarity measures with reference comments. We further apply `GCC-Eval` to evaluate the performance of different chess commentary generation methods. The results in Table 5 indicate that `CCC` outperforms the baselines in all `GCC-Eval` metrics, showcasing the effectiveness of integrating domain-specific expertise and concept-based explanations.

4.4 Other experiments

Chess skills and knowledge of language model While LLMs can generate linguistically sound commentary, they lack the deep, inherent understanding of chess strategies. Integrating expert models like chess engines compensates for this limitation, ensuring that the LLM's output is both fluent and grounded in expert knowledge. To verify the chess skill level of LLMs, we use mate-in-one chess problems and evaluate how the models solve them, in Table 7. `GPT-4o` solves 57% of problems, while other language models are below 12%, even though `ChessGPT` (Feng et al., 2023) is fine-tuned on chess-related documents. When the expert model evaluation result is given in prompt, the LLM solves 95% of the problems, which is not surprising because the expert model evaluation includes the answer.

Comment generation methods	Relevance	Completeness	Clarity	Fluency
Reference	0.51	0.25	0.47	0.72
`GAC` (Jhamtani et al., 2018)	0.47	0.14	0.39	0.81
`GPT-4o`	0.79	0.48	0.85	0.95
`GPT-4o + expert`	0.81	0.49	0.75	0.90
`GPT-4o + expert + concept` (`CCC`, ours)	**0.89**	**0.54**	**0.88**	**1.00**

Table 5: *Automatic evaluation results using* `GCC-Eval`. Numbers are rescaled to the range $[0, 1]$.

Concepts	Accuracy	Precision	Recall
Material	0.93	0.93	0.94
Imbalance	0.80	0.73	0.93
Pawns	0.84	0.81	0.90
White Knights	0.91	0.87	0.96
Black Knights	0.91	0.87	0.97
White Bishop	0.77	0.73	0.87
Black Bishop	0.75	0.71	0.83
White Rooks	1.00	1.00	1.00
Black Rooks	0.99	0.99	1.00
White Queens	0.74	0.71	0.79
Black Queens	0.81	0.84	0.77
White Mobility	0.99	0.99	1.00
Black Mobility	0.98	0.96	0.99
White Kingsafety	0.96	0.97	0.94
Black Kingsafety	0.94	0.96	0.91
White Threats	0.93	0.90	0.96
Black Threats	0.93	0.90	0.97
White Space	1.00	1.00	1.00
Black Space	1.00	1.00	1.00
White Passedpawns	0.98	0.98	0.98
Black Passedpawns	0.92	0.91	0.94

Table 6: *Test accuracy, precision and recall of concept-based explanations.*

While `GPT-4o + expert` includes the answer in the prompt, `GPT-4o + concept` also shows significant improvement of 17.2%p, with only a simple hint that there is a mate. It implies that a proper concept serves as a powerful hint for the precise analysis of the position. For more detailed explanation, refer to Appendix E.

Reliability of the concept-based explanation We assess the reliability of the concept-based explanations. Table 6 shows that the average accuracy of the extracted chess concepts is 0.91, demonstrating that the model effectively identifies and utilizes key domain-specific concepts. This further supports the idea that concept-based explanations serve as reliable source for guiding the LLM in generating chess comments.

Interactive commentary generation We also explore the potential of `CCC` for generating interactive and context-aware chess commentary. By augmenting the LLM with the decision-making capabilities of an expert model, it responds to flexible user questions, providing deeper insights beyond simple commentary on a move. The questions can be strategic intentions, long-term plans, and potential threats in a given chess position. An example of these interactive commentary capabilities and corresponding results are found in Appendix D. These experiments demonstrate that `CCC` is capable of generating not only accurate move annotations, but also high-quality interactive chess insights that meet different requirements of different users.

Language models	LLM	LLM + expert	LLM + concept (mate-in-one)
`GPT-4o`	0.564	0.982	0.736
`GPT-4o-mini`	0.014	0.988	0.031
`GPT-3.5-turbo`	0.036	0.988	0.056
`ChessGPT` (Feng et al., 2023)	0.118	0.563	0.175

Table 7: *LLM chess skill evaluation on mate-in-one problems.*

5 Discussions

Language model as an explanation form Our work shows that the `CCC` framework effectively transfers AI-driven chess knowledge to human users. Beyond concept-based explanation, language models can act as a crucial medium between the expert model's internal reasoning and the end-user. This connection facilitates more intuitive and understandable feedback than traditional explanation methods like saliency-based, which suffer from issues of inconsistency and unreliability. By employing language-based form of explanation, the transparency of the explanation can be improved, making the evaluation of the model's reliability more straightforward.

Fine-tuning with `GCC-Eval` We validate that `GCC-Eval` is well-correlated with human evaluation. One promising direction to improve the quality of chess commentary is to incorporate `GCC-Eval` as a training objective, replacing human evaluator. By optimizing models to directly align with this evaluative criterion, we can better ensure that the generated commentary meets the standards of human chess experts. This approach offers a potential pathway toward more robust and human-aligned commentary systems in future applications.

Applicability of `CCC` While we focus on chess in this paper as a first step, our approach is applicable to other domains, such as science (Tshitoyan et al., 2019; Sprague et al., 2019), medical imaging (Yeche et al., 2019), and games(Das et al., 2023) where concept-based explanation methods have been actively studied. well-established concept extractors and relevant datasets in those domains make the domains suitable for applying our method.

Results with PGN chess notation We also verify that `gpt-4o` understands PGN notation, not only FEN notation. Inferring from `CCC` results with PGN notation, the differences from that with FEN notation are not noticible, but we cannot validate it in large scale because our dataset does not contain the game progresses which is required for using PGN notation. Appendix F provides the results with PGN notation.

6 Conclusions

In this paper, we propose methods for chess commentary generation (`CCC`) and evaluation (`GCC-Eval`). `CCC` integrates expert and language models through concept-based explanations, utilizing techniques such as prioritization, few-shot learning, and Chain-of-Thought prompting to align effectively with expert knowledge. `CCC` either surpasses or matches the quality of human-generated commentary, demonstrating the capability of LLMs to express expert-level understanding and potentially enhance learning for human users. We also present `GCC-Eval`, a multi-dimensional evaluation framework that incorporates chess-specific knowledge to assess chess commentary. The strong correlation between human evaluation and `GCC-Eval` validates the robustness. These findings underscore promising future research directions, including using a language model as an explanation method and using `GCC-Eval` fine-tuning chess commentary generation models.

7 Limitations

Use of proprietary LLMs We release the source code used in our experiments. However, since we employed proprietary LLMs including `GPT-4o` (from July to October 2024), it can be limited to fully reproduce the results. Nonetheless, the proposed framework remains adaptable and can be further enhanced with the integration of more advanced LLMs. In addition, it is also interesting to further investigate the efficacy of our framework with smaller LLMs.

Educational purpose / comment for beginners The main audience for commentary is often beginners and or those with less knowledge than the commentator. In the human evaluation in Section 4.2, we assess the commentary in the view of expert chess players. Another human evaluation involving novice players can assess the educational impact of the comments. For the same purpose, Chen et al. (2023) propose counterfactual simulatability, as an automatic evaluation metric of the improvement of students.

Beyond chess commentary Although we focus on the chess commentary generation, our method can be extended to other tasks, that require comprehensive decision-making abilities and have an expert model. Empirical experiments in other tasks require finding the appropriate tasks and corresponding expert models.

More concepts Although we use concepts from `Stockfish 8`, there are other useful concepts such as fork, pin, double-pawn or open-file. We do not use the concepts because of insufficient concept labels, but they could be valuable, as the concept "mate-in-one" improves chess skill in Table 7.

Differences between concept evaluation function and extracted concept In our work, we extract the concept vectors from an expert model. Although using oracle concept evaluation functions is relatively more accurate, there are two key reasons for using the extracted concepts. First, recent findings (Schut et al., 2023) emphasize that expert models often possess super-human knowledge, capturing patterns and strategies not easily interpretable by humans. It implies the extracted concepts can cover the comprehensive knowledege of model, even if the humans do not understand and an oracle concept evaluation function is not present. Second, when the model has defects, the extracted concepts are used to find the cause of failure. These two aspects facilitate us to use extracted concepts.

Acknowledgement

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government Ministry of Science and ICT (MSIT) (RS-2019-II191906, Artificial Intelligence Graduate School Program (POSTECH); IITP-2024-RS-2024-00437866; RS-2024-00509258, Global AI Frontier Lab), and a grant from KRAFTON AI.

References

Guillaume Alain and Yoshua Bengio. 2016. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*.

The LCZero Authors. 2024. Leelachesszero. Available at https://lczero.org/.

Murray Campbell, A. Joseph Hoane, and Feng hsiung Hsu. 2002. Deep blue. *Artif. Intell.*, 134:57–83.

Yanda Chen, Ruiqi Zhong, Narutatsu Ri, Chen Zhao, He He, Jacob Steinhardt, Zhou Yu, and Kathleen McKeown. 2023. Do models explain themselves? counterfactual simulatability of natural language explanations. *arXiv preprint arXiv:2307.08678*.

I. Chernev. 2003. *Logical Chess : Move By Move: Every Move Explained*. Rizzoli.

Corinna Cortes and Vladimir Naumovich Vapnik. 1995. Support-vector networks. *Machine Learning*, 20:273–297.

Devleena Das, Sonia Chernova, and Been Kim. 2023. State2explanation: Concept-based explanations to benefit agent learning and user understanding. *Advances in Neural Information Processing Systems*, 36:67156–67182.

Xidong Feng, Yicheng Luo, Ziyan Wang, Hongrui Tang, Mengyue Yang, Kun Shao, David Mguni, Yali Du, and Jun Wang. 2023. Chessgpt: Bridging policy learning and language modeling. *arXiv preprint arXiv:2306.09200*.

Joseph L Fleiss and Jacob Cohen. 1973. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and psychological measurement*, 33(3):613–619.

Harsh Jhamtani, Varun Gangal, Eduard Hovy, Graham Neubig, and Taylor Berg-Kirkpatrick. 2018. Learning to generate move-by-move commentary for chess games from large-scale social forum data. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1661–1671, Melbourne, Australia. Association for Computational Linguistics.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. 2018. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *International Conference on Machine Learning*, pages 2668–2677. PMLR.

Andrew Lee, David Wu, Emily Dinan, and Mike Lewis. 2022. Improving chess commentaries by combining language models with symbolic reasoning engines. *arXiv preprint arXiv:2212.08195*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.

Thomas McGrath, Andrei Kapishnikov, Nenad Tomašev, Adam Pearce, Martin Wattenberg, Demis Hassabis, Been Kim, Ulrich Paquet, and Vladimir Kramnik. 2022. Acquisition of chess knowledge in alphazero. *Proceedings of the National Academy of Sciences*, 119(47):e2206625119.

Shikib Mehri and Maxine Eskenazi. 2020. Usr: An unsupervised and reference free evaluation metric for dialog generation. *arXiv preprint arXiv:2005.00456*.

OpenAI. 2023. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Aðalsteinn Pálsson and Yngvi Björnsson. 2023. Unveiling concepts learned by a world-class chess-playing agent. In *IJCAI*, pages 4864–4872.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

J. Polgar. 2014. *Judit Polgar Teaches Chess 3 Â a Game of Queens*. Judit Polgar teaches chess. Quality Chess UK LLP.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

Ehud Reiter and Anja Belz. 2009. An investigation into the validity of some metrics for automatically evaluating natural language generation systems. *Computational Linguistics*, 35(4):529–558.

Tord Romstad, Marco Costalba, Joona Kiiski, and Gary Linscott. Stockfish chess engine. `https://stockfishchess.org`.

Lisa Schut, Nenad Tomasev, Tom McGrath, Demis Hassabis, Ulrich Paquet, and Been Kim. 2023. Bridging the human-ai knowledge gap: Concept discovery and transfer in alphazero. *arXiv preprint arXiv:2310.16410*.

David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, Timothy Lillicrap, Karen Simonyan, and Demis Hassabis. 2018. A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, 362(6419):1140–1144.

David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas baker, Matthew Lai, Adrian Bolton, Yutian Chen, Timothy P. Lillicrap, Fan Hui, L. Sifre, George van den Driessche, Thore Graepel, and Demis Hassabis. 2017. Mastering the game of go without human knowledge. *Nature*, 550:354–359.

Conner Sprague, Eric B Wendoloski, and Ingrid Guch. 2019. Interpretable AI for deep learning-based meteorological applications. In *American Meteorological Society Annual Meeting*.

Amanda Stent, Matthew Marge, and Mohit Singhai. 2005. Evaluating evaluation methods for generation in the presence of variation. In *Proceedings of the 6th international conference on Computational Linguistics and Intelligent Text Processing*, pages 341–351.

Vahe Tshitoyan, John Dagdelen, Leigh Weston, Alexander Dunn, Ziqin Rong, Olga Kononova, Kristin A Persson, Gerbrand Ceder, and Anubhav Jain. 2019. Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature*, 571(7763):95–98.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Hugo Yeche, Justin Harrison, and Tess Berthier. 2019. Ubs: A dimension-agnostic metric for concept vector interpretability applied to radiomics. In *Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support: Second International Workshop, iMIMIC 2019, and 9th International Workshop, ML-CDS 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 17, 2019, Proceedings 9*, pages 12–20. Springer.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. *Advances in Neural Information Processing Systems*, 34.

Mert Yuksekgonul, Maggie Wang, and James Zou. 2023. Post-hoc concept bottleneck models. In *The Eleventh International Conference on Learning Representations*.

Hongyu Zang, Zhiwei Yu, and Xiaojun Wan. 2019. Automated chess commentator powered by neural chess engine. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5952–5961, Florence, Italy. Association for Computational Linguistics.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M Meyer, and Steffen Eger. 2019. Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 563–578.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022. Towards a unified multi-dimensional evaluator for text generation. *arXiv preprint arXiv:2210.07197*.

Appendix

A Details for `GCC-Eval`

The scoring prompt, which includes the expert evaluation and Auto-CoT reasoning, is illustrated in Figure A1.

B Reproduction of baselines

Although Jhamtani et al. (2018) provide the source code, necessary pre-processing files are missing. For fair comparison, we align the raw files with the pre-processed files to compare reference text and chess move to the generated comments. For the same reason, we cannot reproduce Zang et al. (2019). Lee et al. (2022) do not share the source code, and as our baseline `GPT-4o + expert` shares the same idea with it, we do not reproduce it.

C Human evaluation examples

Figure A2 and Figure A3 are instructions and questions we used for human evaluation.

D Interactive commentary

Figure A4 shows an example of interactive comments, starting from `CCC`. The initial chess commentary is generated by `CCC`. If there are parts of the generated comments that are unclear or difficult to understand, users can engage with the system by asking follow-up questions to clarify any ambiguous or complex parts of the commentary. Similarly, they can request additional insights, such as alternative moves or a deeper analysis of the current game position.

This interactive approach enhances knowledge transfer between the AI and users, making expert-level chess understanding more accessible. By enabling two-way communication, the functionality of LLMs is extended, transforming the model from a static generator of text into an interactive learning tool that adapts to the needs and curiosity of the user. This capability promotes a more engaging and educational experience in chess commentary, expanding the role of LLMs in expert domains.

E Chess skill evaluation details

We conduct chess skill evaluation for LLMs. We use mate-in-one puzzle data from database of Lichess (`https://database.lichess.org/#puzzles`). We conduct evaluation for 1,000 puzzle data. Evaluation prompts are shown in Figure A5. For `GPT-4o + expert`, we include expert model evaluation information in the prompt (Figure A5(a)). For `GPT-4o + concept`, we provide an explanation indicating that the board is in a mate-in-one situation (Figure A5(b)). For `GPT-4o`, `GPT-4o-mini`, `GPT-3.5-turbo`, and `ChessGPT`, we use a basic prompt for evaluation (Figure A5(c)).

F Results with PGN notation

We verify that LLM and `CCC` also works with PGN notation instead of FEN notation. Figure A6 presents an example result of `CCC` with FEN and `CCC` with PGN. There are no significant differences in the generated commentary quality or usage of concepts.

G Licenses of artifacts

In this study, `GPT-4o` is used in compliance with its usage policy. `ChessGPT` is used under the terms of the Apache-2.0 license. The Lichess database is used according to the Creative Commons CC0 license. As there are no specific license statements for `GameKnot` and `GAC`, we regard them as Creative Commons CC0 license.

All artifacts are used within the intended use.

[System]
You will be given single comment about a chess move.
Your task is to rate the comment on one metric.
Please make sure you read and understand these instructions carefully. Please keep this document open while reviewing, and refer to it as needed.

Evaluation Criteria:
Relevance (1-5) - Relevence of a target comment. The comment should include only information relevant to the chess move or reasoning for taking or not taking the chess move. An engine evaluation result is given as a hint.
Evaluation Steps:
1. Read the comment carefully.
2. Assess how well the comment addresses the important information about the chess move, and how relevant it is.
3. Assign a Relevance score from 1 to 5.

[User]
position: *8/3nk3/1p4pp/1N1P1p2/1bP2KP1/3P1P2/7P/8 b - - 0 0*
move: *30... Bd2+*
target comment: *Good move, Bd2+ forces the White king to move, gaining tempo and improving the position of the Black bishop.*
engine evaluation: *actual move - Bd2+ 232cp, expected reply - f4g3, best move - Bd2+ similar to actual move, second best move - Nc5 similar to actual move*
Score(1-5, score ONLY):

(a) Example prompt of relevance.

[System]
You will be given single comment about a chess move.
Your task is to rate the comment on one metric.
Please make sure you read and understand these instructions carefully. Please keep this document open while reviewing, and refer to it as needed.

Evaluation Criteria:
Completeness (1-5) - Completeness of a comment. The comment should cover all critical points on the chess board, ensuring that no important factors are overlooked. An engine evaluation result is given as a hint.
Evaluation Steps:
1. Read the comment carefully.
2. Assess how well the comment addresses the important information, and how well the comment covers the entire important information without missing any.
3. Assign a Completeness score from 1 to 5.

[User]
position: *8/3nk3/1p4pp/1N1P1p2/1bP2KP1/3P1P2/7P/8 b - - 0 0*
move: *30... Bd2+*
target comment: *Good move, Bd2+ forces the White king to move, gaining tempo and improving the position of the Black bishop.*
engine evaluation: *actual move - Bd2+ 232cp, expected reply - f4g3, best move - Bd2+ similar to actual move, second best move - Nc5 similar to actual move*
Score(1-5, score ONLY):

(b) Example prompt of completeness.

[System]
You will be given single comment about a chess move.
Your task is to rate the comment on one metric.
Please make sure you read and understand these instructions carefully. Please keep this document open while reviewing, and refer to it as needed.

Evaluation Criteria:
Clarity (1-5) - Clarity of a comment. The comment should be clear and detailed, without vague or ambiguous statements.
Evaluation Steps:
1. Read the commment carefully.
2. Assess how the comment is clear and detailed, without vague or ambiguous statements.
3. Assign a Clarity score from 1 to 5.

[User]
position: *8/3nk3/1p4pp/1N1P1p2/1bP2KP1/3P1P2/7P/8 b - - 0 0*
move: *30... Bd2+*
comment: *Good move, Bd2+ forces the White king to move, gaining tempo and improving the position of the Black bishop.*
Score(1-5, score ONLY):

(c) Example prompt of clarity.

[System]
You will be given one comment written for a chess move.
Your task is to rate the comment on one metric.
Please make sure you read and understand these instructions carefully. Please keep this document open while reviewing, and refer to it as needed.

Evaluation Criteria:
Fluency (1-5): Fluency of a comment.
1. Read the commment carefully.
2. Assess the sentences of comment is coherently organized. The comment should contain well-structured language and coherent transitions.
3. Assign a Fluency score from 1 (not readable) to 5 (very fluent).

[User]
target comment: *Good move, Bd2+ forces the White king to move, gaining tempo and improving the position of the Black bishop.*
Score(1-5, score ONLY):

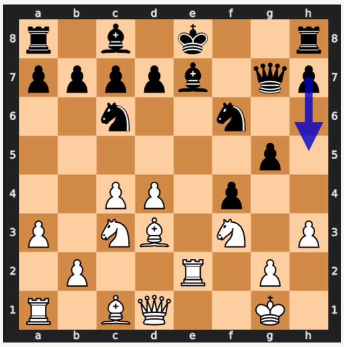
(d) Example prompt of fluency.

Figure A1: *Example prompts for* `GCC-Eval`. The *blue text* in the figure changes according to the experimental conditions.

Chess Commentary Evaluation: Instruction

About survey
This survey asks you to evaluate comments on chess moves. For each move (shown in blue arrow on a chessboard), five different comments for the same move are provided. For each comment, you are asked to evaluate it based on five metrics: **correctness, relevance, completeness, clarity, and fluency**. In what follows, we explain each metric and then provide example evaluations on three comments.

Please do not spend too much time analyzing the chessboard. **Each page is expected to be completed in one minute (although you may need some extra time at the beginning)**.



Explanation of each criteria
Please refer to the image above and read example comments with it.

Correctness:
You are asked to indicate your level of agreement to that "The commentary provides accurate analysis, ensuring that all evaluations and moves are logically and factually correct."

Example of correct sentence: Black expands kingside with h5.
Example of incorrect sentence (misuse of move/pieces): Black attacks knight.
Example of incorrect sentence (wrong understanding/description of tactical advantage): White can fork knight and bishop in the next move.
Example of incorrect sentence (wrong understanding/description of positional / long-term advantage): Black's kingside castling will improve king safety.

Relevance:
You are asked to indicate your level of agreement to that "The commentary focuses on the key aspects of the move or game without including unnecessary or unrelated information."

Example of relevant sentence: Black expands kingside with h5.
Example of irrelevant sentence: Black's c8 bishop has not activated.

Completeness:
You are asked to indicate your level of agreement to that "The commentary covers all critical aspects of the position, ensuring that no important factors are overlooked."

Example of complete comment: Black expands kingside with h5, supported by rook. But black still have weaknesses in pinned e7 bishop and c6 knight can be threatened by d5.
Example of incomplete comment: Black expands kingside with h5.

Clarity:

You are asked to indicate your level of agreement to that "The commentary is clear and detailed, without vague or ambiguous statements."

Example of clear sentence: Black is starting kingside pawn push, followed by g4 and f3.
Example of unclear sentence: Black is starting kingside pawn push.

Fluency:

You are asked to indicate your level of agreement to that “The commentary flows smoothly, with well-structured language and coherent transitions between sentences.”

Example of fluent comment: A bold pawn push, h5 aims to challenge white’s center and possibly prepare for further expansion on the kingside.
Example of not fluent comment: A pawn push, h5 to white’s Center and expansion the kingside.

This diagram compares relevance and completeness. High relevance indicates C is small, and high completeness indicates A is small.

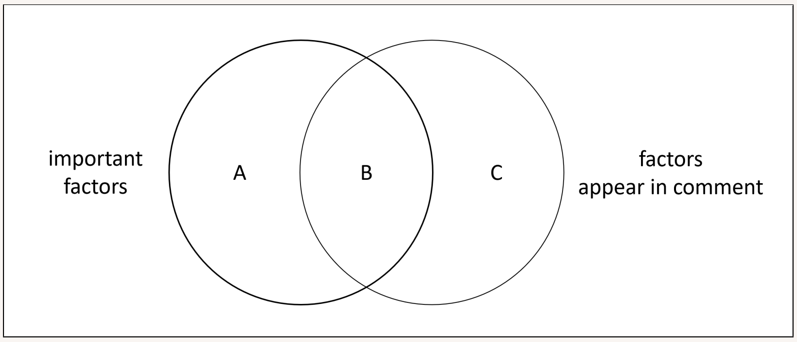
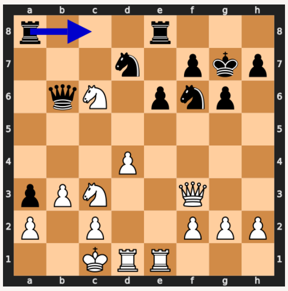


Figure A2: *Survey instruction.*

Chess Commentary Evaluation

Position and move (1/250)



Comment to be evaluated:
20... Rac8 is a solid move, reinforcing the c-file and preparing to challenge White's strong knight on c6. However, White still maintains a strong position with active knights and potential threats.

Correctness: The commentary provides accurate analysis, ensuring that all evaluations and moves are logically and factually correct.

.	1	2	3	
disagree	○	○	○	agree

If the comment is incorrect, what is the incorrect point?

- ☐ referring illegal move or non-existing pieces
- ☐ wrong understanding/usage of tactical/immergent advantage
- ☐ wrong understanding/usage of positional/long-term advantage
- ☐ wrong evaluation of the move
- ☐ etc:

Relevance: The commentary focuses on the key aspects of the position or game without including unnecessary or unrelated information.

.	1	2	3	4	5	
strongly disagree	○	○	○	○	○	strongly agree

Completeness: The commentary covers all critical aspects in the position, ensuring that no important factors are overlooked.

.	1	2	3	4	5	
strongly disagree	○	○	○	○	○	strongly agree

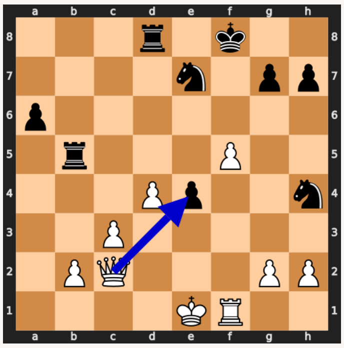
Clarity: The commentary is clear and detailed, without vague or ambiguous statements.

.	1	2	3	4	5	
strongly disagree	○	○	○	○	○	strongly agree

Fluency: The commentary flows smoothly, with well-structured language and coherent transitions between sentences.

.	1	2	3	4	5	
strongly disagree	○	○	○	○	○	strongly agree

Figure A3: *Survey example.*



[Comment of LLM:]

comment: Qxe4 is an even trade but misses a stronger opportunity with Qf2. While the move captures the passed e4 pawn, it exposes White to counterplay, particularly b5-b2, which could generate strong pressure on White's queenside. Better would have been Qf2, maintaining positional advantage and limiting Black's activity.

[Following question by user:]

After the move, can black's h4 knight survive?

[Response of LLM:]

After 26. Qxe4, Black's knight on h4 is under threat. The White queen can capture the knight with Qxh4. Black doesn't have any immediate way to defend or save the knight on h4 effectively.

Given that the knight is undefended, and White can simply take it on the next move, the knight cannot survive unless Black manages to create a very strong counterattack that would force White to deal with something else first, but that seems unlikely based on the current position.

Thus, it looks like Black's knight on h4 cannot survive and is likely lost after White's next move.

Figure A4: *Example of interactive commentary.*

[System]
You will be given a chess board, formatted with Forsyth-Edwards notation(FEN) string.
Your task is to find the best move of this board.
Please answer the best move in standard algebraic notation(SAN).
[User]
position: N6r/1p1k1ppp/2np4/b3p3/4P1b1/N1Q5/P4PPP/R3KB1R b KQ - 0 18
engine evaluation: best move - Bxc3#
Move(SAN formatted move only):

(a) Example prompt of `GPT-4o + expert`.

[System]
You will be given a chess board, formatted with Forsyth-Edwards notation(FEN) string.
Your task is to find the best move of this board. *You can make checkmate in one move.*
Please answer the best move in standard algebraic notation(SAN).
[User]
position: N6r/1p1k1ppp/2np4/b3p3/4P1b1/N1Q5/P4PPP/R3KB1R b KQ - 0 18
Move(SAN formatted move only):

(b) Example prompt of `GPT-4o + concept` "mateIn1".

[System]
You will be given a chess board, formatted with Forsyth-Edwards notation(FEN) string.
Your task is to find the best move of this board.
Please answer the best move in standard algebraic notation(SAN).
[User]
position: N6r/1p1k1ppp/2np4/b3p3/4P1b1/N1Q5/P4PPP/R3KB1R b KQ - 0 18
Move(SAN formatted move only):

(c) Example prompt of `GPT-4o`, `GPT-4o-mini`, `GPT-3.5-turbo`, `ChessGPT`.

Figure A5: *Example prompts for chess skill evaluation with mate-in-one problems.* The *blue text* in figures (a) and (b) indicates the differences from figure (c).

[PGN]

1. e4 Nf6 2. Nc3 d6 3. b3 g6 4. f4 Bg7 5. Nf3 O-O 6. Bc4 Re8 7. O-O b6 8. Ne2 Nxe4 9. d3 Bxa1 10. dxe4 Bg7 11. Ng3 Nc6 12. f5 Ne5 13. fxg6 hxg6 14. Nxe5 Bxe5 15. Bxf7+ Kg7 16. Bxg6 Rh8 17. Rf5 Bxf5 18. Bxf5 Qg8 19. Nh5+ Kf7 20. Qd5+ e6

[comment generated using PGN]

The move 21. Bxe6+ is a strong continuation for White, as it captures a pawn with check and maintains the pressure on Black's position. This move forces Black to respond to the check, giving White the initiative. The move also opens up the possibility of further attacks on Black's king, which is somewhat exposed. The White bishop on e6 is well-placed, and Black's pieces are somewhat disorganized, especially with the Black queen and rooks not ideally coordinated to defend. Overall, this move helps White maintain a significant advantage.

[FEN]

r5qr/p1p2k2/1p1pp3/3QbB1N/4P3/1P6/P1P3PP/2B3K1 w - - 0 21

[comment generated using FEN]

The move 21. Bxe6+ is a strong tactical decision by White. By capturing the pawn on e6 with check, White not only gains material but also disrupts Black's pawn structure and exposes the Black king further. This move forces Black to respond to the check, giving White the initiative. The White bishop is well-placed, and this move capitalizes on its active position. The evaluation indicates a significant advantage for White, reflecting the effectiveness of this move.

Figure A6: An example of `CCC` with PGN and FEN notation.