Legal Move Classification in Portal Legends

GitHub: https://github.com/CodenItamar/PortalLegends-LLM-Benchmark

Itamar Cohen itamar.cohen10@mail.huji.ac.il
Moral Bootbool moral.bootbool@mail.huji.ac.il
Hoshen Lugasi hoshen.lugasi@mail.huji.ac.il

Abstract

We investigate whether pretrained language models can learn rule-based reasoning from text alone. Specifically, we test whether models can classify legal and illegal moves in a board game setting without explicit access to the rules. Using a novel dataset of hand-generated examples from the custom-designed game *Portal Legends*, we fine-tuned BERT, DistilBERT, and RoBERTa for binary classification. Our results show that models are capable of internalizing aspects of the game's logical structure, with RoBERTa achieving the strongest performance on the test set. This work highlights the potential and limitations of large language models for structured reasoning tasks grounded in symbolic rules.

1 Introduction

Humans can easily understand whether a move in a game is allowed or not, following clear patterns and observations, even without explicit rules. Our goal is to teach a model to do the same, using only written descriptions of moves and game states, without showing it the rules directly. In simple terms, we want the computer to learn the game from examples.

Today, most approaches rely on either manually encoded rules or large language models (LLMs) trained on text. While LLMs can predict text well (Devlin et al., 2019; Liu et al., 2019), they often struggle when asked to follow strict, structured rules that are not common in natural language. This limitation means that computers cannot reliably decide if a move is legal in a new or unfamiliar game just by reading descriptions.

Our approach is new because we provide a small and carefully crafted dataset of game moves from *Portal Legends*, a custom board game with specific rules which the models have not trained on. By fine-tuning popular LLMs such as BERT, DistilBERT, and RoBERTa on this dataset, we test whether these models can internalize the logic of the game from text alone. We believe this will succeed because some models have shown strong ability to learn patterns in text, and the game moves provide a clear, structured pattern to learn.

If successful, this work will demonstrate that language models can acquire structured reasoning and ruling skills from examples, without explicit programming. This could help in education, game AI, and any application must learn rules from text, such as instructional manuals or legal documents.

2 Data

We created a novel dataset of 151 hand-generated examples from the board game *Portal Legends*, a turn-based game inspired by Snakes and Ladders but enhanced with bi-directional portals and character-specific abilities. Each instance consists of:

- A textual description of the board state.
- A candidate move and its resulting outcome.
- A binary label indicating whether the move is legal or illegal.

All data was manually generated to ensure consistency and correctness. Attempts to use generative models such as GPT-4 failed to reliably capture the full complexity of the rules.

The dataset is split into training, validation, and test sets as follows: 90 training, 30 validation, and 31 test examples. The dataset contains 79 legal moves and 72 illegal moves, distributed among three characters with distinct abilities: Double Jumper, Gold Miner, and Mystic. Each instance is further annotated with binary scenario features marking salient mechanics, including actor's power usage, portal involvement, die misuse, problematic board states, multi-character contexts, backward

movement, overloaded tiles, and moves marked as winning.

Character	Total	Legal	Illegal
Double Jumper	60	32	28
Gold Miner	60	30	30
Mystic	31	17	14
Total	151	79	72

Table 1: Distribution of examples by character and move legality in Portal Legends.

Character	Special Power
Double Jumper	May split move on even rolls
Gold Miner	Rolls two dice in 1–4 range
Mystic	Can re-roll once per round

Table 2: Special powers associated with each character in *Portal Legends*.

3 Methods

We fine-tuned three pretrained transformer-based models: BERT-base (Devlin et al., 2019), DistilBERT (Sanh et al., 2019), and RoBERTa-base (Liu et al., 2019).

3.1 Preprocessing

All textual descriptions of game states and candidate moves were tokenized using the HuggingFace Transformers library. We applied a maximum sequence length of 128 tokens, truncating longer descriptions and padding shorter ones. The tokenizer used was model-specific (e.g., `BertTokenizer` for BERT and DistilBERT, `RobertaTokenizer` for RoBERTa).

3.2 Training Setup

Each model was fine-tuned for binary classification (legal vs. illegal). We used the AdamW optimizer with a learning rate of 2×10^{-5}, linear learning-rate warmup, and weight decay. Batch size was set to 8. Training was performed for a maximum of 20 epochs with early stopping based on validation loss. Model checkpoints with the best validation performance were saved for evaluation.

3.3 Evaluation Metrics

Performance was measured using accuracy, precision, recall, and F1-score. In addition, we conducted per-class accuracy (legal vs. illegal), per-character accuracy (Gold Miner, Mystic, Double Jumper), and per-feature accuracy (e.g., die misuse, portal usage, special abilities) to better understand failure modes.

4 Results

We evaluated the fine-tuned models on the Portal Legends test set. Table 3 (located on the dedicated tables page) presents accuracy, precision, recall, and F1-score for each model.

RoBERTa achieved the highest overall performance, with an accuracy of 0.806 and the best balance between precision (0.917) and recall (0.688), resulting in the highest F1-score of 0.786. BERT achieved moderate performance, with high precision (0.727) but lower recall (0.500), indicating it was cautious and missed some illegal moves. DistilBERT had high recall (0.875) but low precision (0.560), suggesting it over-predicted illegal moves.

Overall, these results indicate that while all models can learn aspects of the game's rules from text alone, RoBERTa is most effective at balancing false positives and false negatives.

Model	Test Acc.	Precision	Recall	F1
BERT	0.645	0.727	0.500	0.593
DistilBERT	0.581	0.560	0.875	0.683
RoBERTa	0.806	0.917	0.688	0.786

Table 3: Performance of baseline models on the Portal Legends dataset (test set). RoBERTa achieves the best overall balance between precision and recall.

5 Error Analysis

A detailed quantitative and qualitative analysis revealed systematic differences across models.

5.1 Overall Performance

On the test set of 31 examples, BERT made 11 errors (accuracy 0.645), DistilBERT made 13 errors (0.581), and RoBERTa only 6 errors (0.806). In terms of F1, RoBERTa achieved the strongest balance (0.786) compared to DistilBERT (0.683) and BERT (0.593).

5.2 Error Types

- **BERT**: Tended to miss illegal moves. Out of 31 cases, it produced 8 false negatives and 3 false positives. This yielded a precision of 0.727 but a recall of only 0.50, indicating that BERT was cautious and often failed to catch rule violations.

- **DistilBERT**: The opposite trend. It over-predicted illegal moves, with 11 false positives and only 2 false negatives. This led to high recall (0.875) but poor precision (0.560), effectively flagging too many legal moves as illegal.

- **RoBERTa**: Achieved the best trade-off. It made only 1 false positive and 5 false negatives, with precision 0.917 and recall 0.688. Errors were fewer and mostly due to missed illegalities rather than over-prediction.

5.3 By Class

BERT achieved 0.80 accuracy on illegal moves but only 0.50 on legal ones, showing a mild bias towards predicting illegality. DistilBERT, paradoxically, achieved 0.875 on legal but only 0.27 on illegal cases, reflecting its tendency to over-flag. RoBERTa reached 0.93 accuracy in illegal moves and 0.69 in legal ones, achieving the best overall balance.

5.4 By Character

The Mystic character was consistently the hardest to classify: BERT: 28.6% accuracy, DistilBERT: 29% accuracy, RoBERTa: 57% accuracy. In contrast, Gold Miner was the easiest (BERT 83.3%, DistilBERT 58%, RoBERTa 92%), while Double Jumper fell between (BERT 66.7%, DistilBERT 75%, RoBERTa 83%). Mystic's re-roll ability remained a primary failure point, highlighting difficulties in compositional reasoning about conditional rules.

5.5 By Features

Performance also varied by feature type:

- **Character ability usage:** Models were more accurate when character powers were explicitly mentioned (68%) versus omitted (55%). DistilBERT, however, struggled more with abilities (50%).

- **Dice misuse:** Easy for RoBERTa (100%) and BERT (80%), but DistilBERT only achieved 20%.

- **Problematic board states:** RoBERTa (100%) and BERT (100%) handled them well, DistilBERT poorly (33%).

5.6 By Input Length

The length of the text was another important factor. Longer descriptions (330 tokens or more) were correlated with poor accuracy (down to 0–33%), especially for BERT and DistilBERT. RoBERTa was more robust, maintaining high accuracy in both short (260-290 tokens) and long inputs (up to 381).

6 Discussion and Conclusion

The comparison between models highlights important trade-offs. BERT achieved high precision (0.73) but low recall (0.50), suggesting that it is cautious and often misses illegal moves. DistilBERT showed the opposite pattern, with very high recall (0.88) but low precision (0.56), capturing most illegal moves but generating many false alarms. RoBERTa struck the best balance, with high precision (0.92) and reasonably high recall (0.69), generating the highest F1 score (0.79). These patterns indicate that different architectures have distinct biases when learning structured reasoning from text.

The Mystic was hardest to classify due to having fewer examples and a more complex, conditional power.

While our results are promising, the small size of the dataset and the fully hand-generated examples limit generalization and linguistic diversity. Future work should expand the dataset with more complex board states, additional characters, and larger-scale models (e.g., T5 or LLaMA). Moreover, few-shot or in-context prompting approaches could complement fine-tuning and reduce reliance on manual data creation. Successfully addressing these limitations could enable language models to learn structured, rule-based reasoning from text, with potential applications in educational tools, game AI, and instruction following in broader domains.

6.1 Limitations and Future Work

The relatively small dataset (151 examples) limits generalization. Furthermore, examples were fully hand-generated, which ensures rule consistency but restricts linguistic diversity. Future work should expand the dataset with more characters, more complex board states, and larger-scale models (e.g., T5, LLaMA). Few-shot or in-context prompting may also be explored to complement fine-tuning and reduce reliance on manual data creation.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.