From Metadata to Metagame: Predicting Card Power and Synergy in Pokémon TCG

Agam Bansal, Thomas Lee, Mei Qu

April 13, 2025

1 Abstract

The Pokémon Trading Card Game (TCG) is a dynamic and competitive collectible card game featuring thousands of unique cards, frequent expansions, and an evolving set of tournament-legal cards. In such a rapidly shifting environment, understanding card synergy and predicting usage trends have become vital for competitive success. This research applies modern machine learning techniques to historical tournament data and card metadata to identify synergies between cards and forecast their future usage in competitive play. Our models aim to quantify relationships between cards, assess their potential utility in evolving metagames, and provide players with a data-driven edge. These findings have applications in competitive analytics, deck-building tools, and strategic planning (Xiao et al. (2023), Bertram et al. (2024)).

Keywords: Pokémon TCG, synergy prediction, competitive analytics, card power forecasting, machine learning

GitHub repository

2 Introduction

The Pokémon TCG operates within a dynamic ecosystem, where card legality, power balance, and strategy are in constant flux. The Standard format, used in most official tournaments, rotates annually—retiring old expansions and introduces new ones–forcing players to continuously reevaluate the competitive landscape. Within this shifting framework, the synergy between cards plays a pivotal role. Even cards with average standalone value can become central to winning strategies if paired correctly. Despite this, the majority of synergy discovery and usage forecasting remains based on community testing, tier lists, and anecdotal evidence. Although effective to some extent, these methods are limited in scalability and speed. Fortunately, the Pokémon TCG community has curated extensive datasets, including a full list of cards and metadata (1999–2023), and thousands of top-performing decklists from official tournaments (2011–2023). These resources provide a foundation for rigorous analysis. By applying natural language processing to parse card effects and structured machine learning models to evaluate patterns in deck composition, we aim to model the complex web of card relationships, predicting both the synergy between any two cards and the likelihood of a card's appearance in future competitive decks, which allows us to bring quantitative clarity to a traditionally qualitative domain (Zuin et al. (2020)).

3 Literature Review

Our project builds upon a rich body of literature in card synergy modeling, gaming NLP, and power forecasting by applying state-of-the-art transformer-based regression techniques to the Pokémon TCG. Prior work in synergy prediction, such as Q-DeckRec and studies leveraging tournament co-occurrence data (Chen et al. (2018)), laid the foundation for quantifying strategic interactions, but often focused on win-rate optimization or heuristic-based pairing. In contrast, we formalize synergy as a predictive task grounded in real tournament data, enabling scalable and objective measurement. Similarly, advances in natural language processing (NLP) for games—ranging from LSTM-based card analysis in Magic: The Gathering to transformer-based planning in Slay the Spire—demonstrated the predictive power of gameplay text. We extend this by fusing structured Pokémon card metadata with BERT and RoBERTa-based embeddings, allowing our models to capture nuanced interplay between text-based abilities and game-

play statistics. Furthermore, while earlier approaches to power forecasting relied on subjective ratings or classification proxies, we predict actual deck inclusion frequencies to serve as an objective measure of power, employing robust evaluation metrics like MAE, F1@100, MRR, and Average Precision. By adapting transformers for regression—using Huber loss, residual-based sample weighting, and Bayesian hyperparameter tuning—our work advances prior methodologies and introduces a novel framework for evaluating card value and synergy in evolving meta-games.

4 Approach

4.1 Data

To support our analysis of card synergy and usage rate prediction, we consolidated two major datasets from the competitive Pokémon TCG ecosystem:

1. **All Pokémon TCG Cards (1999–2023)**
   This dataset contains over 17,000 cards released from 1999 through 2023. For each card, metadata includes card type (e.g., Pokémon, Trainer, Energy), subtype (e.g., Basic, Stage 1, Item), attack names and damage values, HP, abilities, evolutionary line, energy costs, rarity, and effect text. These structured and unstructured attributes were essential for capturing the characteristics of individual cards and understanding how they might interact with each other in a deck.

2. **Competitive Decklists (2011–2023)**
   We obtained thousands of competitive decklists from top-ranking players in sanctioned tournaments across multiple years. Each decklist included the cards played, deck archetype, tournament name and format, and the deck's final placement. These decklists provide real-world insights into which cards were used together frequently and which became central to winning strategies. These co-occurrence patterns serve as ground truth labels for both our synergy and usage prediction tasks.

Each card includes structured attributes (e.g., HP, retreat cost) and unstructured text (e.g., attack descriptions, abilities), making it ideal for hybrid modeling with NLP techniques. We split the dataset by "Pokémon" and "trainer" card types, which differ significantly in both gameplay rules and text structure—warranting separate evaluation and model training.

4.2 Power Level Computation

One of the biggest challenges in finding the strongest cards is that currently there is no formula for assigning quantitative power levels to cards. However, we can look at card usage rates in tournaments as a heuristic for power level. Players will always choose the strongest cards to put in their decks when competing in tournaments, so a card that appears in many decks must be a very strong card. Each year, a new set of cards becomes legal for play while old cards are phased out. We compute the usage rates for cards in each year separately to train and evaluate our models.

4.3 Baseline

We created two constant models as baseline assessments to compare against the performance of our models. The first constant model predicted the mean of the training data, and the second constant model predicted the median of the training data. These simple baseline models are appropriate for our task because of the distribution of our target variable. Among the thousands of cards that are legal for play in a given year, only a handful are strong enough to see any competitive play, so these models perform very well in terms of L1 and L2 loss. It is important for our models to perform comparably to these baselines to accurately reflect the reality of the highly skewed card power level distributions while still being able to identify the strongest cards.

Metric	Baseline (mean)	Baseline (median)
Trainer MAE	**0.0575**	**0.0379**
Pokémon MAE	**0.0056**	**0.0027**

Table 1: Baseline Model Performance Metrics

4.4 Modeling

Building on the limitations of traditional machine learning models, we adopted a hybrid modeling strategy that combined traditional machine learning approaches with modern transformer-based deep learning models. This iterative process involved exhaustive experimentation with text encodings, model structures, and hyperparameter adjustments akin to grid-searching for optimal arrangement frameworks.

4.4.1 Data Preprocessing and Feature Construction

Figure 1: The card "Comfey", with explanation of its card features

We concatenated prominent text fields such as the name, abilities, attacks, and rules pertaining to the card into a singular string representing the card. Using HuggingFace tokenizers (RoBERTa-base and bert-base-uncased), we padded and truncated sequences to 512 tokens for GPU compatibility. Furthermore, we modified the model for regression by removing the classification head and adding a single neuron output which used a sigmoid activation to constrain predictions within the [0, 1] range.

4.4.2 RoBERTa (`RoBERTa-base`) Adaptation

For RoBERTa, we configured the architecture for regression by changing `num_labels` = 1 to allow for the prediction of continuous values. We also substituted the default MSE loss with Huber loss due to its greater robustness against label noise and outlier values, which are often present in subjective power estimates. To enhance learning further, we applied dynamic error-based weighting, which increased the weighting of harder-to-predict samples by their residuals, thus motivating the model to focus on more challenging cases. We also incorporated early stopping based on validation MAE, where training was ceased after three steps of no improvement, to control overfitting.

4.4.3 BERT (`bert-base-uncased`) Adaptation

In parallel, we fine-tuned a BERT model with the same architecture and training strategy as RoBERTa to allow for a straightforward comparison with RoBERTa. This consistency in approach allowed for a fair evaluation of the transformer models on the same dataset and task. Both models outperformed traditional regressors in capturing complex interactions within the text, highlighting their ability to contextualize the power levels embedded in card abilities and descriptions.

4.4.4 Fine-Tuning

Further improvements were introduced after evaluating our initial models. As a more advanced measure, we incorporated optimization algorithms to fine tune the training dynamics by adjusting the learning rate, warm up steps, and weight decay (Frazier (2018)). Moreover, we augmented the RoBERTa model by adding two dense layers on top of the transformer backbone (Liu et al. (2019)). The first dense layer, consisting of 512 neurons with ReLU activation, introduced non-linear transformations of the hidden states, while the second layer, with 128 neurons, further refined these representations before feeding into the output layer. This architectural enhancement was based on the hypothesis that additional layers could help the model capture more nuanced patterns in the data, although the impact was not consistent across all configurations of training setup.

5 Results

The following results depict our models' performance on each MAE score over the respective dataset. F1 score provides a balanced measure of the model's performance, combining both precision and recall to give a comprehensive evaluation of its ability to correctly classify the most powerful cards. The Mean Reciprocal Rank (MRR) further highlights how quickly the model ranks the relevant cards, with higher MRR values indicating that high-power cards are found early in the list of predictions (Hoyt et al. (2022)). Finally, Average Precision (AP) quantifies the precision at each rank across the top 100 predictions, offering a detailed view of the model's ability to rank relevant cards highly and maintain precision throughout the predictions. A summary of our results is shown in Table 2.

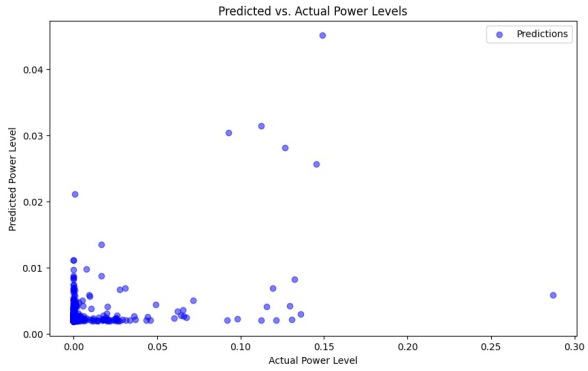


Figure 2: Best Pokémon Model Predicted vs. Actual Power Levels

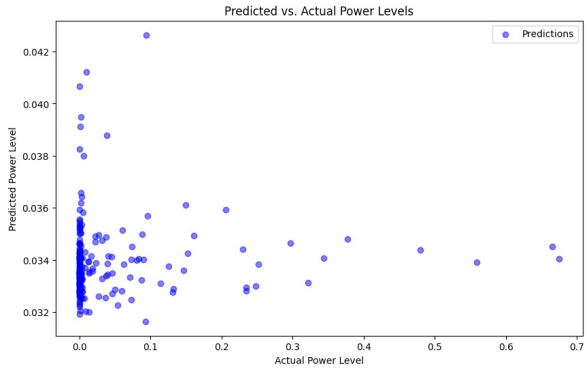


Figure 3: Best Trainer Model Predicted vs. Actual Power Levels

These results reveal that the BERT and RoBERTa flavored models performed considerably better on Pokémon cards than on Trainer cards. This performance difference can be better understood by examining the distribution of power levels across card types (Table 3).

Key Observations:

1. Trainer cards have significantly higher average power levels (0.0356 vs. 0.0025)
2. Trainer cards show much greater variability (std: 0.0979 vs. 0.0156)
3. The highest power level for Trainer cards (0.6744) is more than double that of Pokémon cards (0.2870)
4. The dataset is imbalanced (1426 Pokémon cards vs. 192 Trainer cards)

These distribution characteristics explain why the model struggles more with Trainer cards and why the baseline (mean prediction) achieves good overall metrics despite not capturing card-specific attributes.

5.1 Card Text Attention

The analysis task that we were most interested in is whether the language models learned something meaningful about the combination of text that makes a card powerful. Although the models do not have an understanding of the rules of the game, we were curious if it was able to discover patterns that made cards strong. For each of the fine-tuned BERT and RoBERTa models, we extracted the attention outputs for a sample of the higher-powered cards and computed the average attention for each token across the layers. For both models, we can see that each token attends mostly to their neighboring tokens as evidenced by the darker diagonal band in the attention heatmap with some notable exceptions. One difference between the RoBERTa attention output and the BERT attention output is that RoBERTa gives a lot of attention to the periods in the card text. Like other card games, punctuation can have a huge impact on how the cards are read, so it makes sense that the model thinks they are important as well. Two examples are discussed below, and more examples are shown in the appendix (Figure 8, 9).

Key Observations:

1. The model heavily attends to "Pokémon" references in Trainer cards
2. Energy costs (##ret for retreat, ##ener for energy) are important for both card types.
3. Type information (##type) is significant for Pokémon cards.
4. The model finds meaningful patterns in punctuation and formatting.
5. For powerful Pokémon like Tyranitar V, the model recognizes the significance of certain mechanics (e.g., "strike" indicating Single Strike or Rapid Strike abilities).

5.1.1 Example 1: Ultra Ball 10

This is a strong card because it allows a player to search their deck for any Pokémon card of their choosing and that you are allowed to play as many item cards as you would like in a turn. When looking at the attention heatmaps outputted by BERT and RoBERTa, these keywords are correctly given more attention. In the BERT model, the pair of words with the highest attention is "play" and "cards" from the sentence "you may play as many item cards as you like during your turn." The word "search" correctly attends the most to the word "deck," demonstrating an understanding of the card text. In the RoBERTa model, the words in the last sentence ("you may play as many item cards as you like during your turn") are given more attention than the words in the beginning of the sentence. Cards like "hand" and "deck" were also given more attention compared to the other cards.

5.1.2 Example 2: Comfey 1

The card text on Pokémon cards are generally more complex than the text on item cards. One example of this is Comfey, which has a middling attack but a powerful ability. Looking at the RoBERTa attention output for Comfey, we can see there are multiple regions of attention, indicating that our model was able to successfully understand the different parts of the text without being prompted to do so. These observations are highlighted in Figure 7.

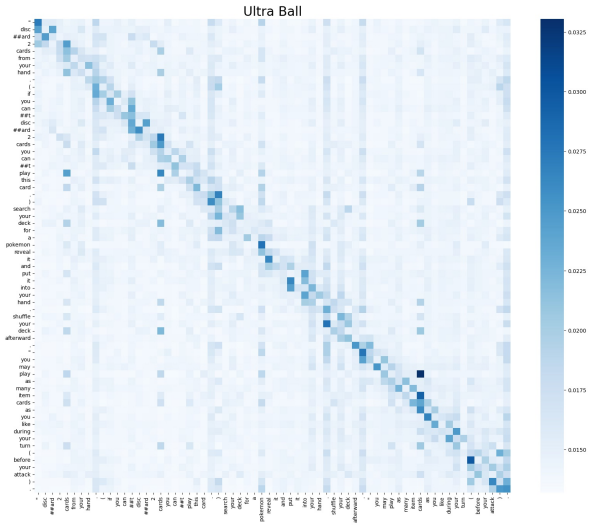


Figure 4: Ultra Ball BERT Attention

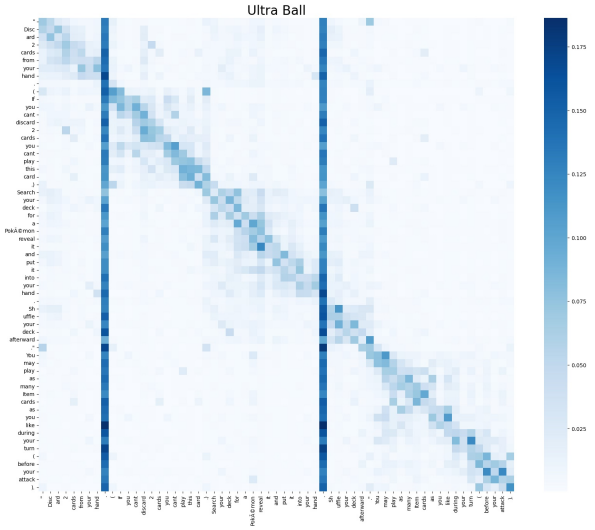


Figure 5: Ultra Ball RoBERTa Attention

6 Additional Card Synergy Task

In addition to our work on finding the most powerful cards, we also conducted an exploratory task on discovering synergies between cards. Pokémon cards do not exist in a vacuum. The best performing decks will always utilize interactions between cards to create combinations that are much stronger than the individual cards themselves. Some of the synergies are very clearly intended by the game developers 10, but most synergies require some experience with the game to identify and understand. With thousands of Pokémon cards, it becomes a very challenging task to discover

which cards synergize well with one another.

6.1 Data

Although it is clear that some cards have much stronger synergy with one another than others, there has not been any prior work done to quantify the level of synergy between cards. We had to created a procedure to label our data. We used a strategy inspired by TF-IDF. For each pair of cards, we define their level of synergy as the frequency of the two cards appearing in the same deck divided by the individual frequencies of each card. This step takes into account that some cards are used often in all decks and are generically good cards instead of being uniquely synergistic with another card.

6.2 Model

For this task, we used a BERT model fine-tuned on card synergy data for years before 2023 and evaluated the model on card synergy data for 2023. In order to output pairwise synergies between each card, we added an additional dense layer of size equal to the total number of cards legal across all years with a sigmoid activation. We decided on using a BERT model because it performed well for the power-level prediction task. Validation testing motivated not adding any additional hidden layers except for the necessary output layer.

6.3 Results

The model did not show an improvement to a 0 baseline (assuming no synergies) because its predictions were less conservative than reality. However, for this task, our group were most interested in whether the model was able to discover new synergies not yet discovered by players. We can see where the model disagrees with the test data by subtracting the predictions from the test data (Figure 7). Differences of 0 represent agreement between the two models, and negative values indicate synergies that the model believes exist but have not been found by players. Taking a closer look at the most negative values we can see that the model predicted many less-used combat focused cards (Mew EX and Single Strike Energy) to have high synergy with Boss's Orders 10. On the one hand, Boss's Orders is an overall very strong card, so it may only be picking up on that. On the other hand, Boss's Orders complements well with attacking cards, so the model could be learning those relationships. It will require further analysis and play testing of the cards to see if the model is picking up useful synergies. A snapshot of the cards with the highest disagreement between test data and predictions is shown in Figure 6.

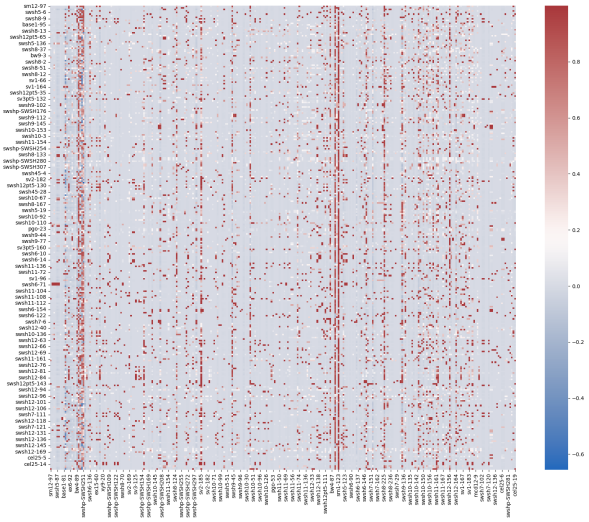


Figure 6: Difference between player-discovered synergies and model predictions

7 Conclusion

Building on the baseline is a challenging task due to "power level" being relatively subjective—synergy, context, or individual play style alters strength which breaks any form of consistency needed for models to learn. Even using more sophisticated approaches, the improvements remained inconsistent across datasets. As with many other games, standard evaluation metrics like MAE or F1@100 provide insufficient value because card interactions are far more complicated than simple hierarchies and contextual heuristics that players use to reason about the game. The predicted best answers tend to suffer from social biases rather than delivering objective truths. This highlights the challenge of tasks that require modeling with context-sensitive, subjective targets.

References

Timo Bertram, Johannes Fürnkranz, and Martin Müller. 2024. Learning with generalised card representations for "magic: The gathering".

Zhengxing Chen, Chris Amato, Truong-Huy Nguyen, Seth Cooper, Yizhou Sun, and Magy Seif El-Nasr. 2018. Q-deckrec: A fast deck recommendation system for collectible card games.

Peter I. Frazier. 2018. A tutorial on bayesian optimization.

Charles Tapley Hoyt, Max Berrendorf, Mikhail Galkin, Volker Tresp, and Benjamin M. Gyori. 2022. A unified framework for rank-based evaluation metrics for link prediction in knowledge graphs.

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.

Changnan Xiao, Yongxin Zhang, Xuefeng Huang, Qinhan Huang, Jie Chen, and Peng Sun. 2023. Mastering strategy card game (hearthstone) with improved techniques.

Gianlucca Zuin, Luiz Chaimowicz, and Adriano Veloso. 2020. Deep learning techniques for explainable resource scales in collectible card games. *IEEE Transactions on Games*, PP:1–1.

Appendix

A Model Results

Table 2: Power Level Results Summary

Dataset	Model	MAE	F1 Top-100	Mean Reciprocal Rank at Top-100	Average Precision at Top-100
Pokemon	BERT	**0.0044**	0.2500	**0.1224**	0.5537
Pokemon	RoBERTa	0.0046	**0.2700**	0.1188	**0.5741**
Pokemon	RoBERTa with dense layers	0.0078	0.1100	0.0591	0.1577
Trainer	BERT	0.0673	0.4900	**0.0700**	0.5795
Trainer	RoBERTa	0.0527	0.4700	0.0676	0.5318
Trainer	RoBERTa with dense layers	0.0595	**0.5300**	0.0451	0.4711
Trainer	RoBERTa with bayesian optimization	**0.0526**	0.5100	0.0674	**0.6057**

Table 3: Power Level Distribution Statistics

Card Type	Count	Power Level (Mean)	Power Level (Std.)	Power Level (Range)
Pokemon	1426	0.0025	0.0156	0.0000 to 0.2870
Trainer	192	0.0356	0.0979	0.0000 to 0.6744

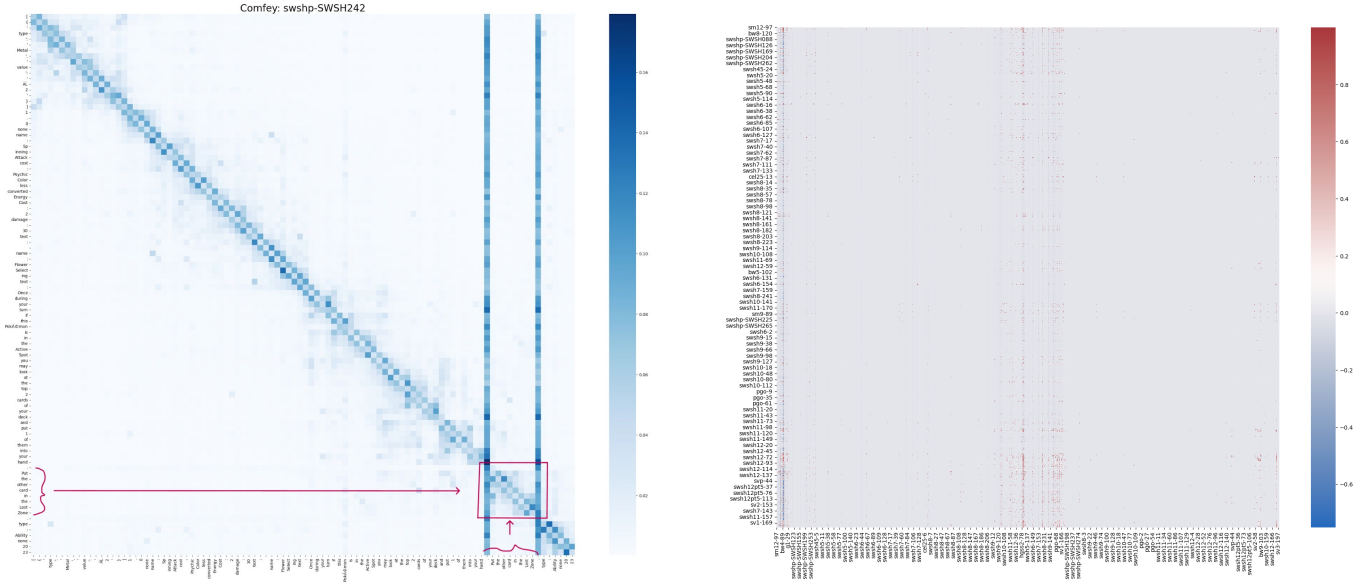
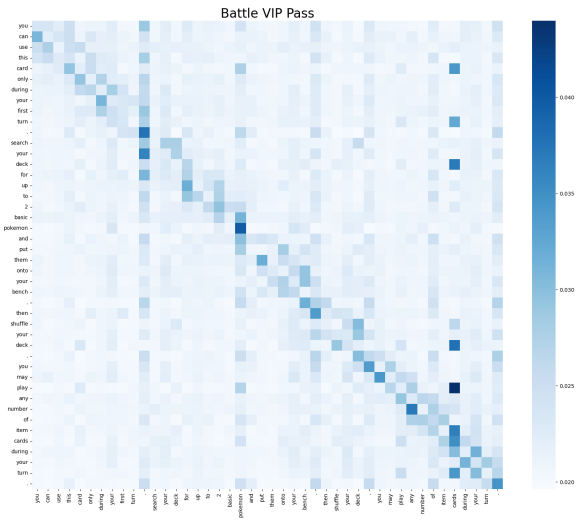
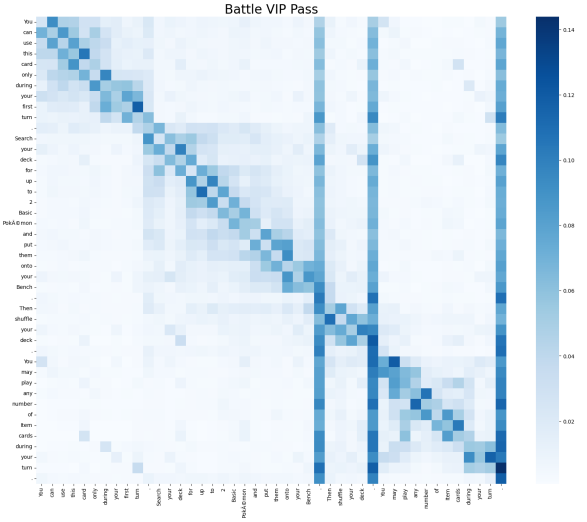


Figure 7: Left: Comfey attention plot; Right: Total synergy differences. Areas with no color represent agreement between true values and predictions.

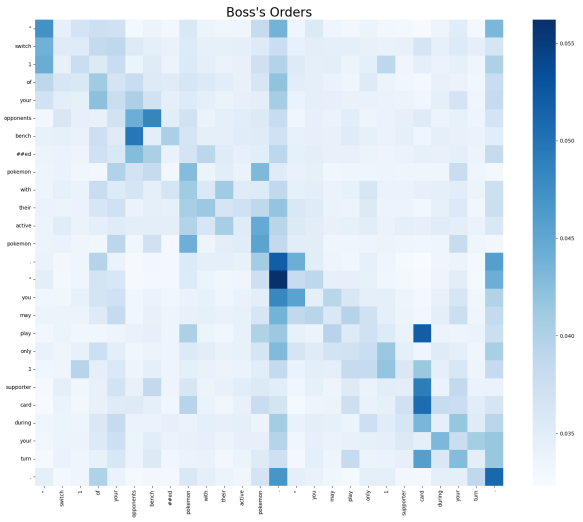
B Attention Outputs



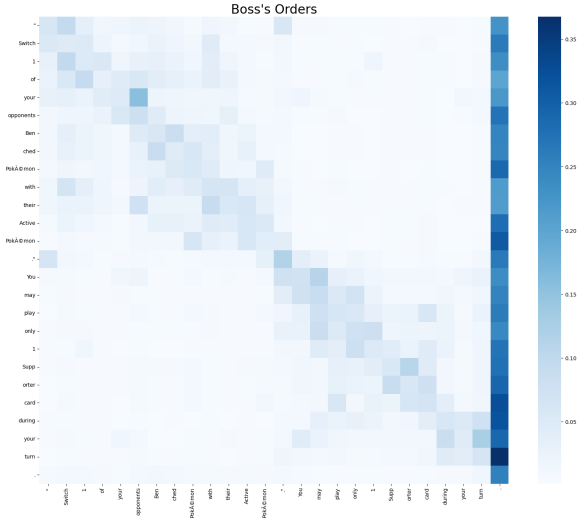
(a) BERT attention for Battle VIP



(b) RoBERTa attention for Battle VIP

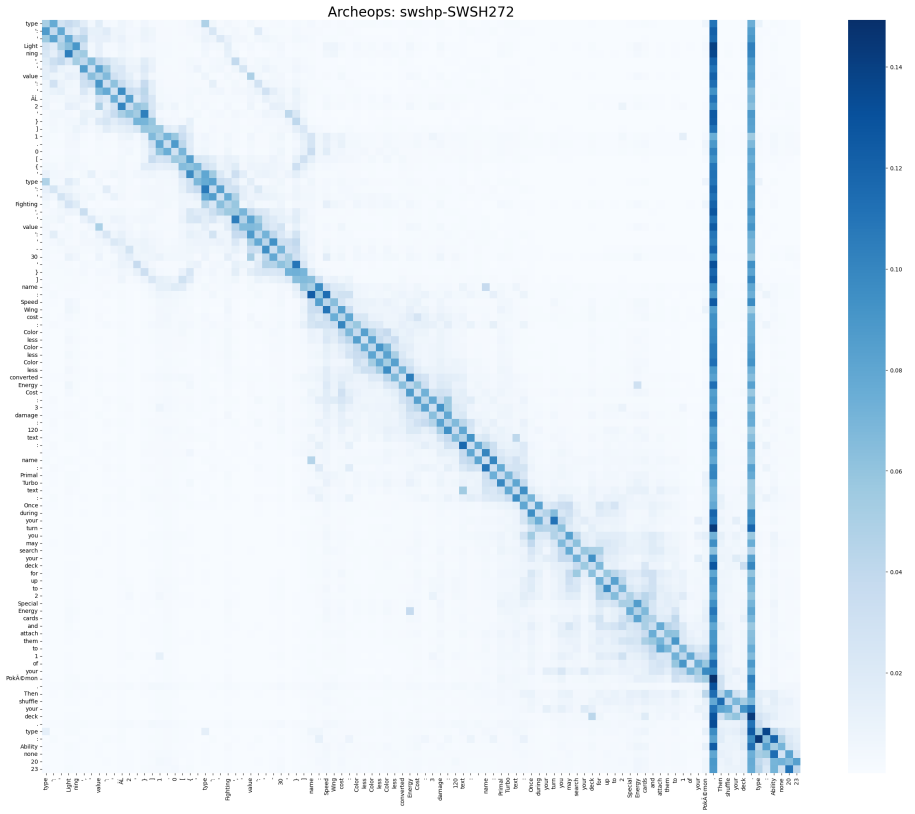


(c) BERT attention for Boss's Orders

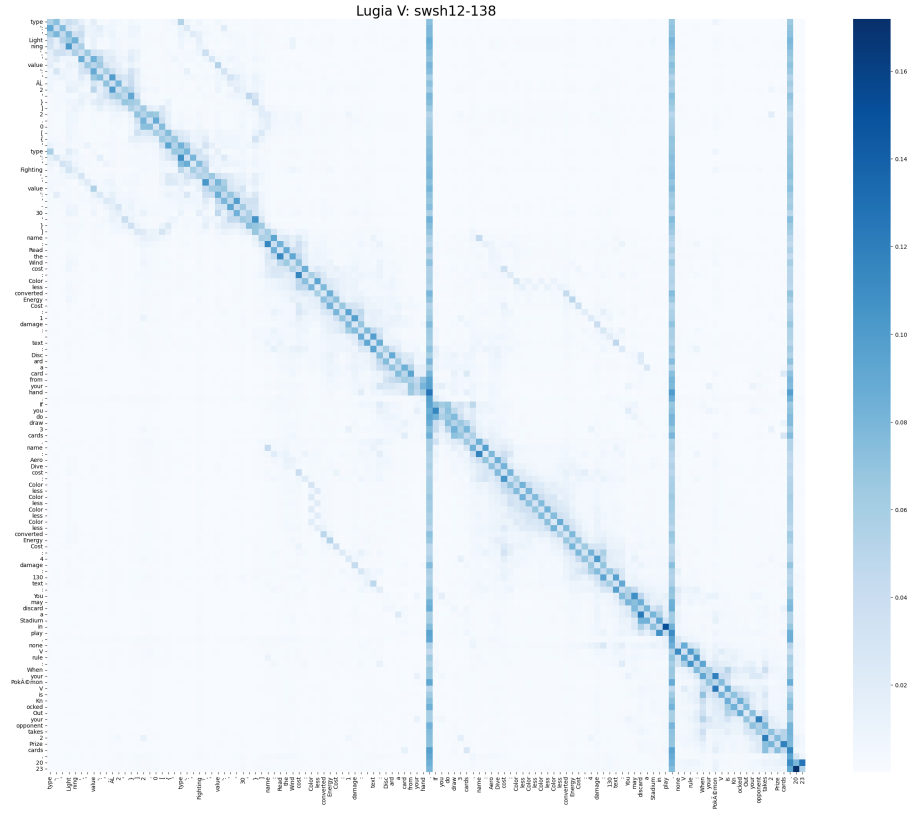


(d) RoBERTa attention for Boss's Orders

Figure 8: Comparison of BERT and RoBERTa attention for Battle VIP and Boss's Orders



(a) RoBERTa attention for Archeops



(b) RoBERTa attention for Lugia V

Figure 9: RoBERTa attention for Archeops and Lugia—two cards which appear frequently together. Some of the same attention patterns appear for both cards.

C Example Cards

(a) Ultra Ball

(b) Boss's Orders

(c) Lugia V

(d) Archeops

Mew EX

(f) Single Strike Energy

Figure 10: Collection of images for cards referenced in the paper.

Figure 11: Synergy example intended by card designers.