We want to thank the reviewers for their helpful comments. The main purpose of this work is to lay the foundation for future research in Diplomacy, starting with the no-press setting. We regret that the page limit necessarily meant that we could not provide all the empirical details in the paper itself. In this response, we provide clarification regarding the **dataset release**, **tournament evaluation**, **architectural design**, **input representation**, and **other insights**.

Dataset Release: The dataset will be made available to any interested researchers. We are unable to make it available as a simple downloadable file due to privacy concerns (the data comes from users of an online Diplomacy service). We are working with the data owners on anonymizing the data to enable wider release. We will edit the sentence in the paper referring to the dataset availability to say "Researchers can access to the dataset by contacting user@domain.com."

Tournament Evaluation: In this work, we conducted **two forms of tournament settings**. In the first setting, we sample two models A and B, use A as one randomly chosen power and B plays the other six powers (1v6). In the second setting, we uniformly sample all models per game per power (TrueSkill). The bulk of Table 4 is from the first setting while our TrueSkill scores were computed with the second setting. We apologize for the confusion and can clarify this in the main text and add details regarding TrueSkill in an appendix. Importantly, **the TrueSkill scores support our claim that both RL DipNet (27.4) and SL DipNet (28.1) can beat the strongest rule-based approach, Albert (24.5)**, which is consistent with the one-vs-six results. In terms of whether RL DipNet or SL DipNet is better, we considered both settings (1v6 and TrueSkill). The two models have TrueSkill score within one standard deviation (0.62), but the one-vs-six result suggests RL DipNet is superior. We agree with R3's concerns about whether 1v6 is a comprehensive evaluation and R2's comment that it is difficult to draw meaningful conclusions regarding which is better. We considered a range of metrics, but many, including ELO and Glicko are designed for 2-player games and generalizing them to a 7-player game like Diplomacy is non-trivial. 1v6 has the benefit that it is more efficient to compute than TrueSkill, while TrueSkill is a well-studied and robust off-the-shelf evaluation metric. We believe that developing an efficient and accurate multi-agent evaluation is an open research question and Diplomacy offers an exciting testbed for this direction.

Architectural Design: We agree with R3 that there are a lot of non-trivial modeling choices in our architecture. We showed the effect of major design choices in our ablation study in Tables 2 and 3, e.g. **the effect of without order history**, **the effect of feeding an averaging embedding instead of a location-specific embedding to the decoder**, etc. We have also conducted more fine-grained experiments. We tried **different numbers of graph convolution layers (GCN)**, and we found that after 8 layers of GCN there is no further improvement. We think this could be related to the fact that in the standard map of Diplomacy the most distant locations are connected by a paths of length 8. We also tried using **order history with more than one year**, but it did not give much improvement in terms of accuracy. We also tried **different sequential orders** during decoding, and found that using a sequential order that does not jump across the map prevents performance dropping when decoding a long sequence with multiple units. We hypothesize that this is because a unit's orders are more influenced by nearby units than distant ones. We also tried **different decoding granularity**. Instead of decoding each unit order as an atomic option (e.g. 'A PAR H'), we can decode as a sequence (e.g. ['A', 'PAR', 'H']). We call the first one unit-based and the latter token-based. We found that although the token-based model had better performance in terms of token accuracy, it produced fewer correct unit orders overall. We apologize for writing some of the claims without referring to the evidence, like "orders from the last movement phase are enough to infer the current relationship between the powers". We will properly modify it and provide more results in appendix.

Input Representation Our input representation is a result of both empirical findings and domain knowledge. Because we want to take advantage of the game adjacency map, we treat each location on the map as a node and try to encode the unit information as a node feature. We also have special treatment for the coast. E.g., if a fleet is on a coast, we will also encode that fleet's information on the parent location, because in Diplomacy occupying any coast of the location is equivalent to occupying the location. We also tried **different ways of encoding history information**. Previously we encoded the history by appending the board state from the previous year, but that did not help much. We hypothesize that the order history should be more informative than board state history because order history captures additional information (e.g., attacks that failed, supports and convoys.). This information can convey relationships between powers. Given this, we encode history by embedding the previous orders on the map as an extra node feature.

Other In terms of RL training, our main insight is that reward shaping based on gaining and losing supply centers is helpful in training. Both R2 and R3 also raise the question of whether there is an effective way to detect **signaling support orders**, particularly in no-check games. Although we have not done such an analysis yet, one simple method to see how common they are would be to compare the fraction of orders that are invalid in no-press and press games. Since signaling orders with invalid syntax are predominantly used in no-press games, we should see a large gap in the measurements. However a case-by-case analysis is still needed to differentiate signalling from mere mistakes.