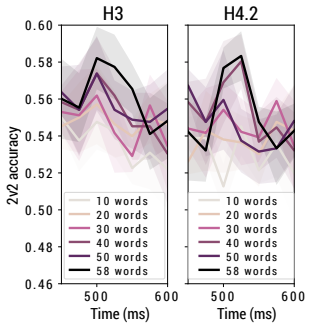
We thank the reviewers for the thoughtful comments and attempt to address their questions, space permitting.

Methodological contributions [all reviewers]. We would like to clarify our *methodological* contributions and their significance as follows: **(1)** we provide the first methodology that can predict brain recordings as a function of *both* the observed stimulus and question task. This is important because it will not only encourage neuroscientists to formulate mechanistic computational hypotheses about the effect of a question on the processing of a stimulus, but also enable neuroscientists to test these different hypotheses against each other by evaluating how well they can align with brain recordings. While we have implemented and compared several hypotheses for this effect, and have found some to be better than others, parts of the MEG recordings remain to be explained by future hypotheses. We hope neuroscientists will build on our method to formulate and test such future hypotheses. We will make our code publicly available to facilitate this. **(2)** we perform all learning in a zero-shot setting, in which neither the stimulus nor the question used to evaluate the learned models is seen during training (i.e. not just as the specific stimulus-question pair but also in combination with any other question/stimulus). Note that this is not the case in previous work that examines task effects, and we are the first to demonstrate how zero-shot learning can be applied successfully to this question. This is important for scientific discovery because it can test the generalization of the results beyond the experimental stimuli and tasks.

Effect size [all reviewers]. We acknowledge that the magnitudes of the presented effects (i.e. accuracies, differences between hypotheses) are small, due to a limited amount of data and the underlying difficulty of analyzing single-trial MEG data. The accuracies we observe are on par with other reported single-trial MEG accuracies[36]. Other work has mitigated the low signal-to-noise ratio of single-trial MEG by averaging the recordings corresponding to different repetitions of the same stimulus[30], or grouping 20 examples together for a 20v20 classification task[36]. Neither is an option for us because our data does not contain multiple repetitions of the same question-stimulus pair, and our zero-shot setting would require us to hold out a large portion of our training set if we were to evaluate on 20 stimulus-question pairs. In the absence of these options, we have taken careful precautions to validate our results (by evaluating our models on held-out data in a cross-validated fashion) and evaluated the significance of the model performances and differences between them, and corrected for multiple comparisons. We trust that the effects we have shown to be significant are indeed true, but we agree that there may be effects that we are not able to reveal due to limited power and hope that neuroscientists will apply our methods in the future to larger datasets with multiple repetitions.

Optimizing H4.2 [R1]. Following R1's suggestion, we trained H4.2 with increasing amounts of data. We also tested H3, which is a simpler model that we expect to learn with fewer samples. To the right, we show that H3 continues to improve as we add more examples, up to the maximum (i.e. 1044 samples = 58words$\times 18$ questions). This suggests that even this simpler model may benefit from more training data. H4.2 also appears to improve with more samples, however it is less clear whether the performance peak has been reached or whether this is due to the difficulty of the optimization problem. Here we present results for times where we expect the two models to perform the best ($450 - 600$ms, Fig.4), and will include the full figures in the paper. For the H4.2 optimizer, we chose Adam because it is a common choice for non-linear problems, such as H4.2. We agree that further investigation into different optimizers, such as L-BFGS, may further improve the H4.2 results. We thank R1 for the suggestion and will test other optimizers and include the results in the appendix.



Connection to word processing literature [R2]. While we have focused the related work on studies that are methodologically similar to ours or investigate task effects, we agree that relating the results to word processing theories will strengthen the paper and thank R2 for the suggestion. We will incorporate this in the discussion section.

NeurIPS fit [R2]. We believe our contributions are of interest to both neuroscience and ML researchers, and NeurIPS is the foremost venue where these communities come together. We will include a discussion of the following directions that we hope our work inspires: **(1)** *new NLP architectures or training algorithms*: it is common to train an NLP model to perform well at a specific task by tuning all model parameters, including the word embeddings. A more brain-aligned method, as inspired by our findings that the question task affects the late stages of processing, would keep some computation task-independent. Such a model may exhibit less catastrophic forgetting when learning new tasks. **(2)** *better evaluation methods for AI models*: our finding that a question representation based on human judgements outperforms BERT may interest the ML community that works on evaluating and incorporating spatial reasoning and other common sense abilities into deep neural networks. **(3)** *new models for task-stimulus interaction in the brain*: the growing computational neuroscience community at NeurIPS is particularly suited for designing such models.

Other types of regularization [R3]. Previous work has found that as long as regularization parameters are properly selected via cross-validation, as is done in our case, different regularization techniques lead to similar results for brain prediction (Wehbe et al. 2015, Ann.Appl.Stat.). We also found that, when the Lasso regularization parameters were carefully tuned, Lasso led to similar results to Ridge regression. However, the Lasso results were much more sensitive to tuning and the optimization was slower. We will incorporate this justification for using Ridge regression in the paper.

Additional comments [R1]. We used custom-built functions that are provided in the supplementary to analyze the MEG data and MNE to visualize Fig. 5 (which we will properly cite). We agree that source localization would improve the understanding of the task effect location, and we leave this to future neuroscientific research.