Rebuttal

Here we provide the results of additional experiments requested by reviewers.

How well do LLMs perform in predicting precise yield values?

In addition to our regression experiments provided in Appendix A.1, we tested LLMs performance in predicting precise yield values on the USPTO-R dataset across two settings:

1. We selected Claude 3 Haiku as one of the best performing models and asked it to predict reaction yields in a few-shot approach with the following system prompt:

You are an expert chemist. Based on text descriptions of organic reactions your task is to predict their yields using your experienced reaction yield prediction knowledge. You will be provided with several examples of reactions and corresponding yields measured in %. Based on these examples, predict the yield of the given reaction. Please, answer with only one float number, do not write anything else.

2. We selected Mistral 7B embeddings as the best performing ones for the USPTO-R dataset and employed grid-serach to find the optimal hyperparameters of the XGBRegressor model.

The comparison of obtained results with baselines is provided below:

	USPTO-R	
	R2	RMSE
Yield-BERT	-0.03	25.91
T5Chem	-15.53	103.47
Egret	-0.16	27.51
DRFP + XGB	-0.03 ± 0.01	25.90 ± 0.16
Claude 3 Haiku (k=2)	-0.14 ± 0.14	27.22 ± 1.64
Claude 3 Haiku (k=4)	-0.13 ± 0.13	27.11 ± 1.46
Claude 3 Haiku (k=6)	-0.10 ± 0.08	26.82 ± 0.91
Claude 3 Haiku (k=8)	-0.07 ± 0.08	26.37 ± 1.04
Claude 3 Haiku (k=10)	-0.10 ± 0.08	26.80 ± 0.94
Mistral 7B embeddings + XGB	**0.06 ± 0.00**	**24.71 ± 0.03**

Conclusion:

Since the regression problem is challenging, LLMs do not provide impressive performance similarly to baselines. However, we would like to highlight that Mistral 7B embeddings achieve best performance among other approaches, which again emphasizes the ability of LLM embeddings to become novel SOTA reaction representations.

How chemical LLMs compare to generalist LLMs?

To compare chemical LLMs directly with generalist LLMs in the few-shot setting, we conducted additional **few-shot experiments with chemical LLMs** on the USPTO-R dataset. Despite changes in prompts and model configuration, Galactica-6.7B consistently responded with the same yield category, producing the same metrcis across all k values. The results are provided below:

	2	4	6	8	10

	2		4		6		8		10	
	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score
ChemDFM-v1.0-13B	0.47 ± 0.05	0.41 ± 0.09	0.44 ± 0.04	0.41 ± 0.18	0.49 ± 0.03	0.57 ± 0.09	0.45 ± 0.05	0.43 ± 0.12	0.50 ± 0.05	0.45 ± 0.23
Galactica-6.7B	0.50 ± 0.00	0.67 ± 0.00	0.50 ± 0.00	**0.67 ± 0.00**	0.50 ± 0.00	**0.67 ± 0.00**	0.50 ± 0.00	0.67 ± 0.00	0.50 ± 0.00	**0.67 ± 0.00**
ChemLLM-7B-Chat	0.33 ± 0.01	0.49 ± 0.01	0.32 ± 0.02	0.49 ± 0.02	0.35 ± 0.03	0.52 ± 0.03	0.4 ± 0.03	0.57 ± 0.03	0.44 ± 0.03	0.60 ± 0.03
Claude 3 Haiku	**0.52 ± 0.00**	**0.68 ± 0.00**	0.57 ± 0.02	0.59 ± 0.04	0.55 ± 0.02	0.67 ± 0.02	0.54 ± 0.03	**0.68 ± 0.01**	0.53 ± 0.01	0.65 ± 0.03
Mistral Small	0.51 ± 0.00	0.66 ± 0.00	**0.59 ± 0.03**	0.60 ± 0.04	**0.61 ± 0.02**	0.63 ± 0.04	**0.56 ± 0.03**	0.62 ± 0.02	**0.57 ± 0.03**	0.61 ± 0.05

Conclusion:

It can be observed that in the few-shot setting generalist LLMs still outperform chemical LLMs in most cases. The best accuracy=0.61 achieved by Mistral Small (k=6) is 11% higher than the best result achieved by chemical LLMs (accuracy=0.5 by Galactica-6.7B).