Thank you to all the reviewers for their detailed reviews. We address specific concerns below.

Reviewer 1 *[test set contains common classes]* Thanks for pointing this out - we should have made this clearer: we are not claiming that an in-production test set would only contain common classes, but rather that the loss defined in line 118 gives zero weight to rare classes, which is mathematically equivalent to not having them in the test set. This amounts to saying, "my classifier should be equally good on all classes, except the extremely rare ones which we deem to matter at all". So if a word has a niche sense in some small community we do not penalize the classifier for not correctly classifying that sense; for example, if we know that an NLP system is not designed for technical conversations between mathematicians, we might not mind if our word sense system fails to recognize "group" as an algebraic structure so we give it zero loss if it fails on such a class in production.

[how effective is the BERT embedding] We agree that a thorough analysis of the word sense distribution in contextualized embeddings would be interesting, but it is beyond the scope of this project. That said, we can make some qualitative comments: at the start of this project we evaluated the embeddings by hand-labelling a couple of words and found BERT does a reasonably good job of separating classes (we explicitly leverage this observation by assuming we have a distance metric). Additionally, because we use a linear classifier on pre-trained BERT embeddings, one can also get some indication of how well separated the classes are by checking the accuracy for the oracle guided learning approach.

[experimental section must be broadened] See the response to Reviewer 2 below.

Reviewer 2 *[results from a single dataset]* As you point out, dataset availability is a challenge. We did perform an experiment along the lines of what you suggest with the Skew MNIST synthetic dataset in appendix C.1. We would be happy to include a similar experiment on CIFAR-100 if you think it would be valuable (note this in your final review if this is the case).

That said, we should note that this paper does more experiments than is typical in active learning: while the evaluated words share a data generating process, each word amounts to a different active learning problem. Typical active learning papers evaluate 10-15 active learning problems, whereas we have 21 words and the Skew MNIST dataset.

[How does the method hold up to violations on assumptions regarding embedding quality?] Good question - this was the motivation behind the the experiment in Appendix C1 which degrades the quality of the embedding on the Skew MNIST dataset. In short - we found that EGAL's performance degrades to that of the standard approaches as the embedding quality degrades (see paper for details).

Reviewer 3 Thank you for picking up those typos - we will correct them in the final draft. Regarding costs - that's a good point, we've assumed the cost of obtaining rare labels is driven by their rarity, but that the costs of labelling an individual example is uniform. We will make this clear in the camera ready.

Reviewer 4 *[Why is the sampling strategy switched to uncertainty sampling?]* Because the search phase searches the neighbourhood of the exemplar, once an example has been found, any additional examples from the target class will typically be very close in embedding space and hence provide relatively little marginal value. We treat the exemplars as out of distribution examples—example usage of a word sense from WordNet will typically differ from "in the wild" text in a Reddit corpus—in order to ensure that the final classifier is not biased by any covariate shift between example usages and actual usage[1]. Of course we need some similarity between the example usage and the "in the wild" usage for the exemplars to be projected into similar parts of embedding space, but this approach allows for differences between the distribution of the exemplars and that of the actual usage without introducing any bias into the final classifier.

[Cosine distance] Good suggestion, thanks!

[Average non-contextized word embedding] We only experimented with BERT embeddings. Performance clearly depends on the quality of the embedding space, so this is an important practical consideration; WordNet embeddings would also be far cheaper to compute. We will experiment with this.

We will incorportate your minor suggestions and include a more complete description of how λ_y is computed in the text. Thank you for an extremely thorough review.

[1] Note that in skew label distributions of the type we study, the classifier will typically see very few examples of the rare class even with the EGAL active learning strategy, so individual training examples can have a relatively large influence over the final decision boundary.