Hyperparameter Name	Range of Values
body_learning_rate	From 5e-6 till 5e-5
head_learning_rate	From 1e-3 till 1e-2
num_epochs	From 3 till 10
batch_size	Amongst [8, 16, 32, 64]
n_trials	10

Table 6: Hyperparameter search space for SetFit model training

Hyperparameter Name	Range of Values
k (no. of ICL examples)	[0, 1, 5, 10, 20]
t (retriever threshold)	[0.00001, 0.3, 0.5, 0.7]

Table 7: Hyperparameter search space for choosing ICL examples for LLM based intent detection

fixed across all experiments and keep ICL examples within a label in descending order of similarity with incoming query.

For **Monte Carlo (MC) sampling** from SetFit models for hybrid system, we look at variance of the predictions as an uncertainty estimate. Specifically, let $p_i \in P \forall i \in [1, M]$ be the predicted label with maximum score from i^{th} sample, where M is the maximum number of samples. Then, we consider the prediction to be uncertain if number of different values of $p_i \forall i \in [1, M]$ is greater than 1 or less than $M/2$. We add upper limit of $M/2$ for stability.

For **latency calculations of hybrid system**, we also add time for doing multiple forward passes sequentially through SetFit in MC sampling procedure keeping memory needs constant. Since maximum $M = 20$ in our experiments, if we consider that sampling can be done in batches, then latency of hybrid system would go further down.

For SetFit models, we calculate OOS AUCROC by considering max predicted score amongst all labels. For black box LLMs, we calculate OOS AUCROC by considering score as 1 if LLM predicts an in-scope label, 0 otherwise.

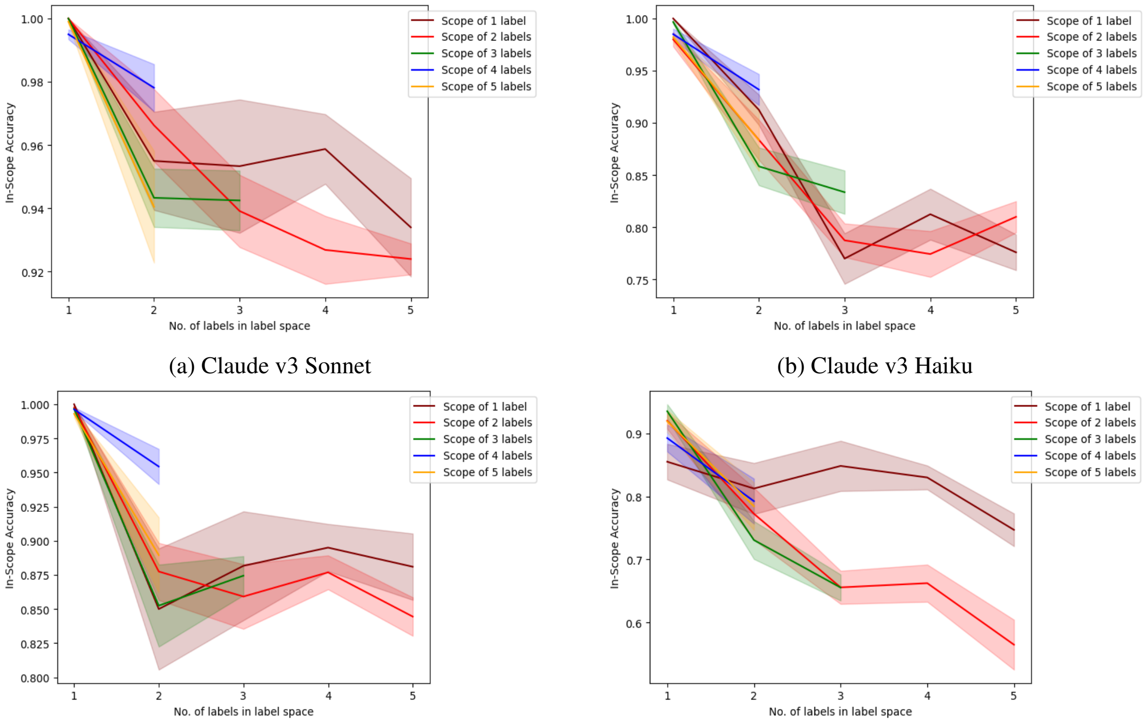
A.3 Controlled Experiment

Setup. For our controlled experiment dataset, we hand-curate 10 utterances per leaf intent, random 5 of which we use in train and other 5 we use in test for every run. We also use three paraphrases (pre-curated) of each test utterance in our test set for every run to test generalization across utterance variants. For controlled experiment, we train all SetFit models with batch size of 16 and 5 epochs. For ICL examples selection with LLMs, we use max 5 ICL examples with retriever threshold of 1e-5. Since we execute every experiment 10 times with randomly created dataset, we are unable to experiment with other hyperparameters due to compute costs. Since we do controlled experiments to develop better understanding of LLM behavior, keeping these hyper-parameters fixed is okay.

Results. Table 8 shows example queries from each intent from controlled experiment dataset. From controlled experiments, Fig 5 and Fig 6 show change in In-Scope accuracy and OOS Recall with number of labels in label space and scope of labels, respectively.

Level 1 class	**Level 2 class**	**Example Utterance**
Product Recommendation	Static Product Attribute based	show laptop with 8gb RAM
Product Recommendation	Similarity/Comparison with other products based	show laptop comparable to the Dell XPS 13
Product Recommendation	Compatibiliy with other products based	show laptop bags compatible with Dell XPS 15
Product Recommendation	Offers based	show laptop with HDFC bank EMI offers
Product Recommendation	Customer Reviews/Ratings based	show laptops whose battery life is highly praised by users
Product Recommendation	Budget based	show laptops under 50k
Product Recommendation	Purpose/Usecase based	show laptops suitable for graphic design work
Product Recommendation	Warranty/Return policy based	show laptops with hassle-free return options
Product Recommendation	Delivery ETA based	show laptops that can be delivered within the next week
Product Recommendation	Past sales based	show the most popular laptop models recently
Product Evaluation	Static Product Attribute based	does this laptop have 8gb RAM
Product Evaluation	Similarity/Comparison with other products based	is this laptop comparable to the Dell XPS 13
Product Evaluation	Compatibiliy with other products based	are these laptop bags compatible with Dell XPS 15
Product Evaluation	Offers based	does this laptop have HDFC bank EMI offers
Product Evaluation	Customer Reviews/Ratings based	are these laptops whose battery life is highly praised by users
Product Evaluation	Budget based	are these laptops under 50k
Product Evaluation	Purpose/Usecase based	are these laptops suitable for graphic design work
Product Evaluation	Warranty/Return policy based	do these laptops have hassle-free return options
Product Evaluation	Delivery ETA based	can these laptops be delivered within the next week
Product Evaluation	Past sales based	are these the most popular laptop models recently

Table 8: Example utterance for each leaf intent from controlled experiment dataset used to understand behavior of LLM based intent detection.



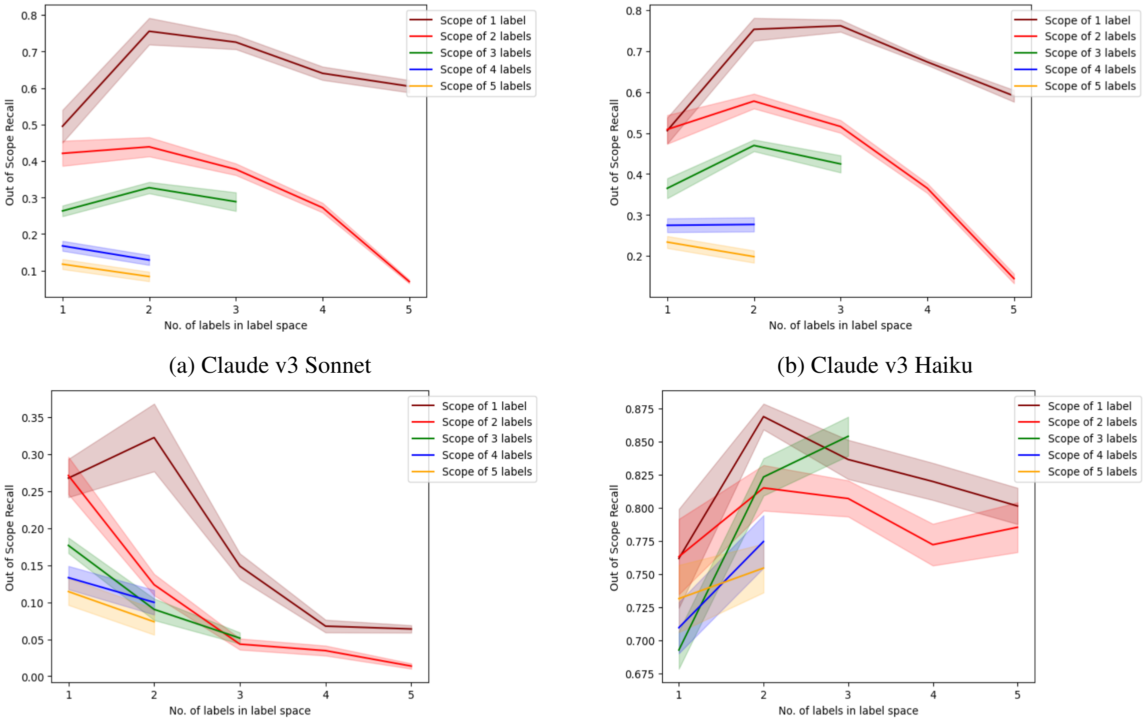
(a) Claude v3 Sonnet

(b) Claude v3 Haiku

(c) Mistral 7B

(d) SetFit + Neg Aug

Figure 5: Change in In-Scope accuracy with number of labels in label space and scope of labels.



(a) Claude v3 Sonnet

(b) Claude v3 Haiku

(c) Mistral 7B

(d) SetFit + Neg Aug

Figure 6: Change in OOS Recall with number of labels in label space and scope of labels.