A Confidence-based Acquisition Model for Self-supervised Active Learning and Label Correction

Carel van Niekerk, Christian Geishauser, Michael Heck, Shutong Feng, Hsien-chin Lin, Nurul Lubis, Benjamin Ruppik, Renato Vukovic, and Milica Gašić

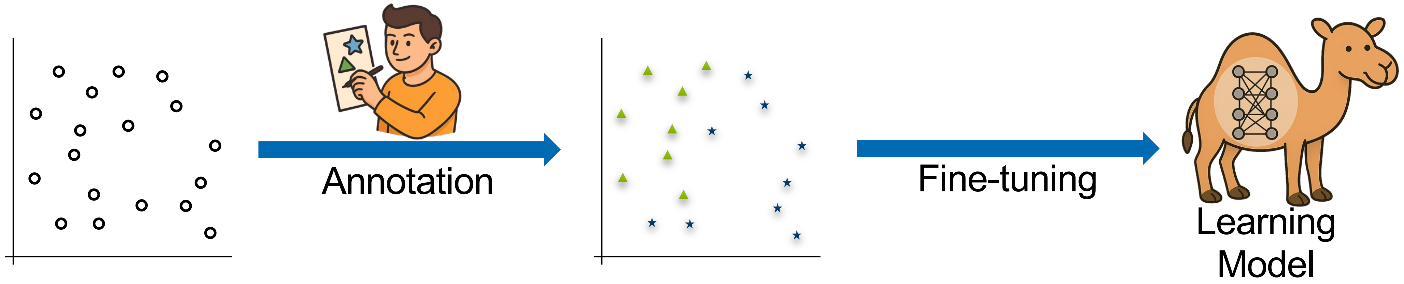
Contact Me

Implementation

Active Learning

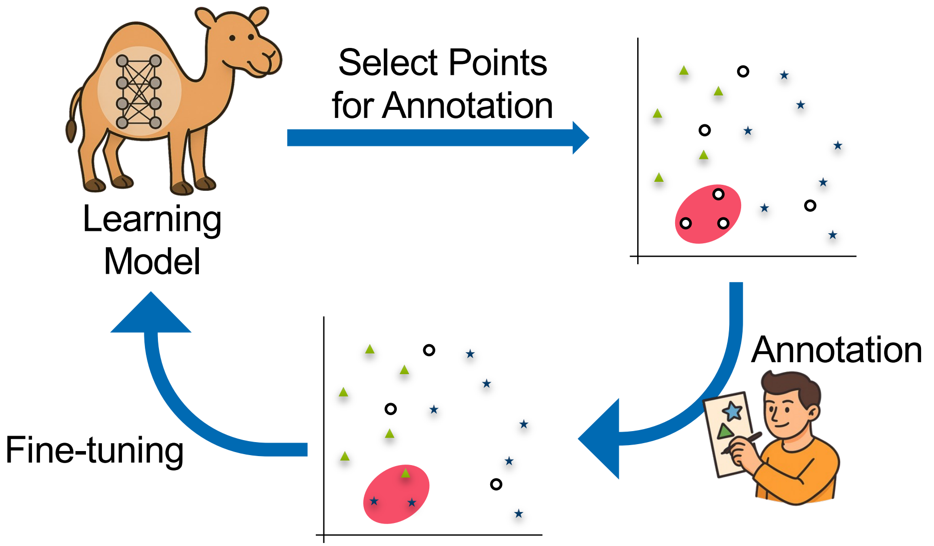
hhu
Heinrich Heine
Universität Düsseldorf

- In **Passive Learning** **large annotated corpora** are collected for tasks such as machine translation, dialogue modelling, etc.



Active Learning

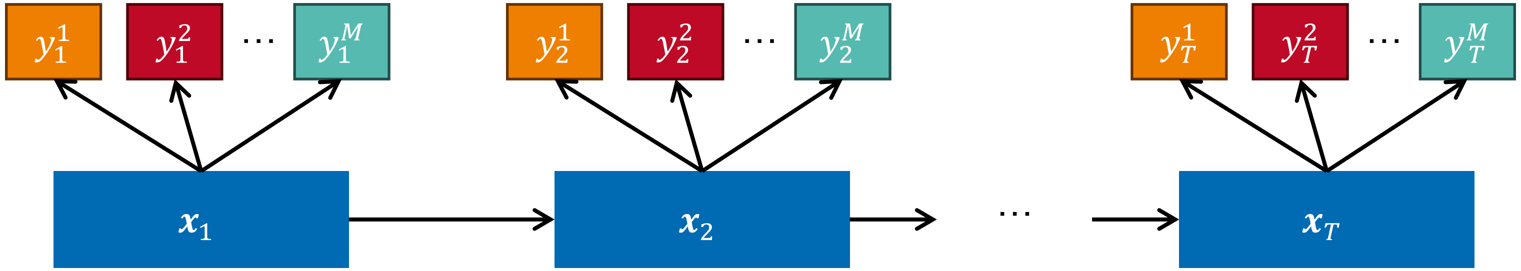
- In **Passive Learning** **large annotated corpora** are collected for tasks such as machine translation, dialogue modelling, etc.
- In **Active Learning**, the learning **model selects the most beneficial datapoints** to learn, reducing the annotation effort.



Sequential Multi-Output Problem

- **Sequential Multi-Output** problems require a label at each timestep for each output category.

- Expert labels can be very **expensive** and crowd labels very **noisy**.



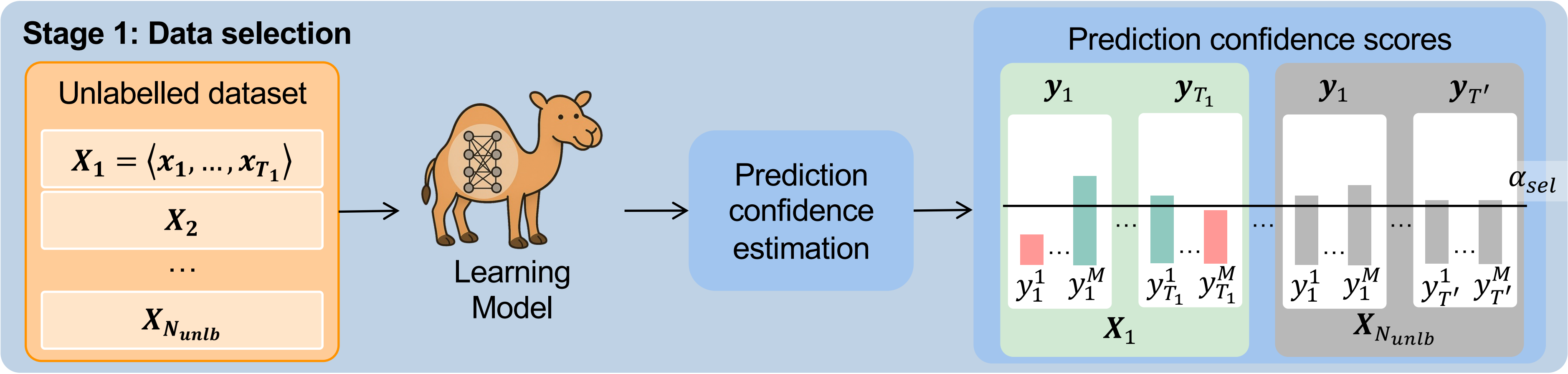
Our Approach

Heinrich Heine Universität Düsseldorf

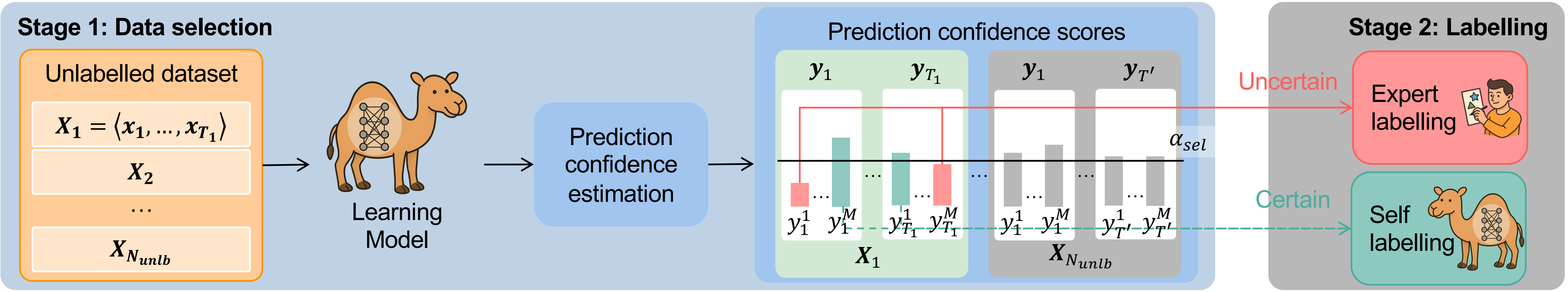
- **CAMEL:**
  - **C**onfidence-based **A**cquisition **M**odel
  - for **E**fficient self-supervised active **L**earning

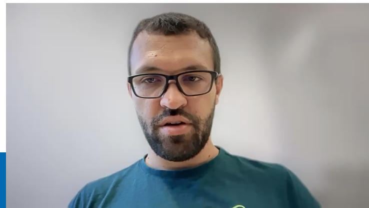
CAMEL

Heinrich Heine
Universität Düsseldorf

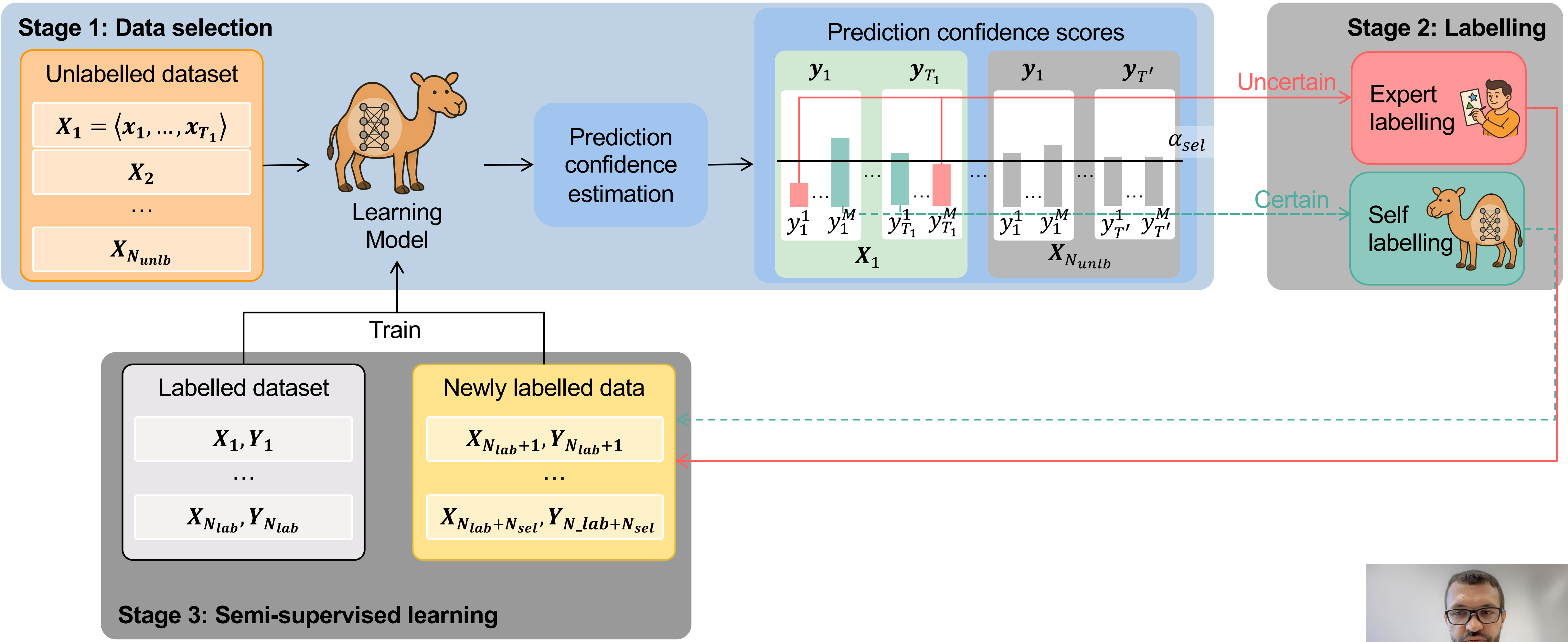


CAMEL

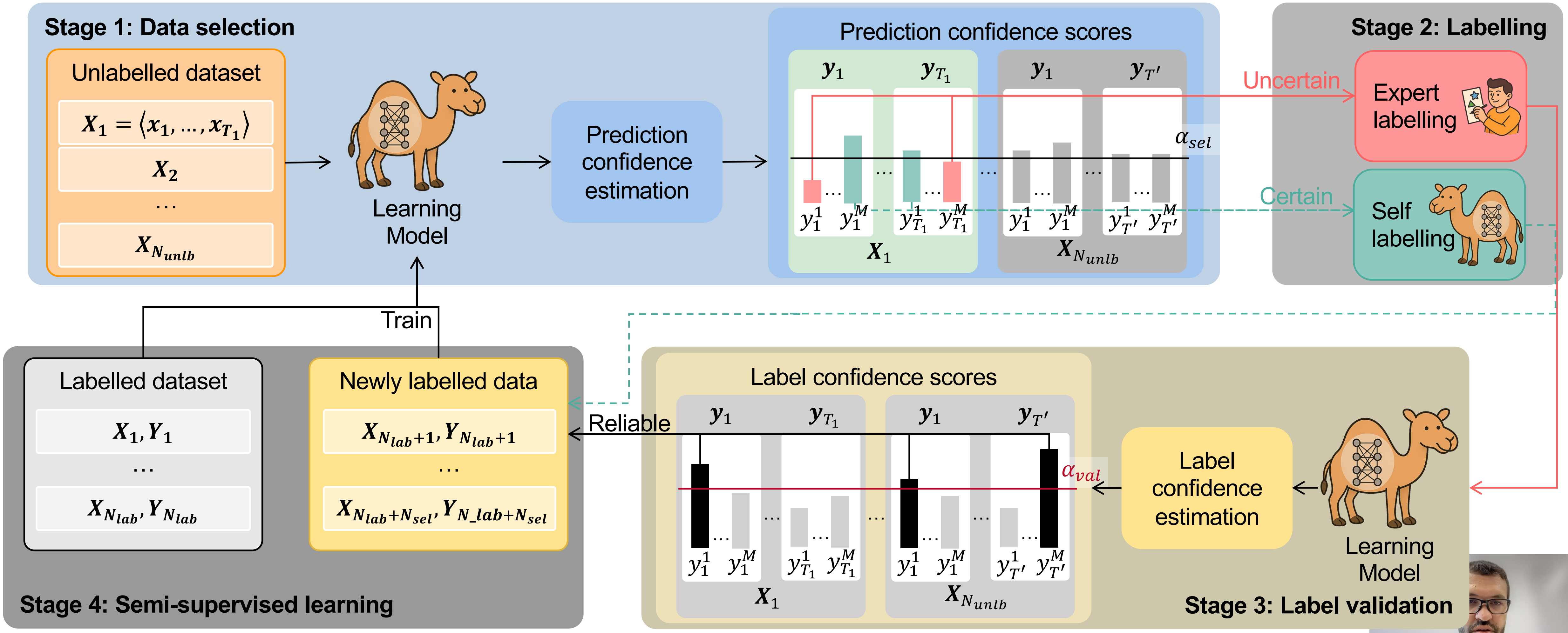

Stage 1: Data selection
Unlabelled dataset
$X_1 = \langle x_1, \ldots, x_{T_1} \rangle$
X_2
…
$X_{N_{unlb}}$
Learning Model
Prediction confidence estimation
Prediction confidence scores
y_1
y_{T_1}
$y_{T'}$
α_{sel}
y_1^1
y_1^M
$y_{T_1}^1$
$y_{T_1}^M$
$y_{T'}^1$
$y_{T'}^M$
X_1
$X_{N_{unlb}}$
Uncertain
Certain
Stage 2: Labelling
Expert labelling
Self labelling



CAMEL


Stage 1: Data selection
Unlabelled dataset
$X_1 = \langle x_1, \dots, x_{T_1} \rangle$
X_2
…
$X_{N_{unlb}}$
Learning Model
Prediction confidence estimation
Prediction confidence scores
y_1
y_{T_1}
y_1
$y_{T'}$
α_{sel}
y_1^1 y_1^M $y_{T_1}^1$ $y_{T_1}^M$
y_1^1 y_1^M $y_{T'}^1$ $y_{T'}^M$
X_1
$X_{N_{unlb}}$
Stage 2: Labelling
Uncertain
Expert labelling
Certain
Self labelling
Train
Labelled dataset
X_1, Y_1
…
$X_{N_{lab}}, Y_{N_{lab}}$
Newly labelled data
$X_{N_{lab}+1}, Y_{N_{lab}+1}$
…
$X_{N_{lab}+N_{sel}}, Y_{N\_lab+N_{sel}}$
Stage 3: Semi-supervised learning

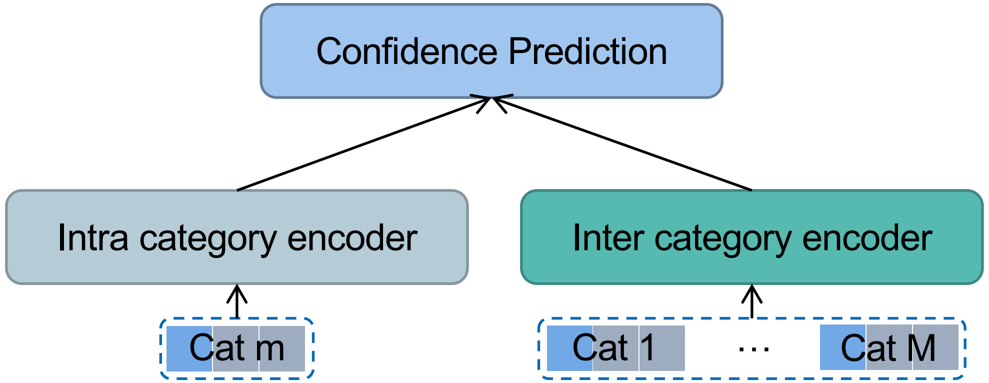
CAMELL – CAMEL with Label validation



Confidence estimation

The Model

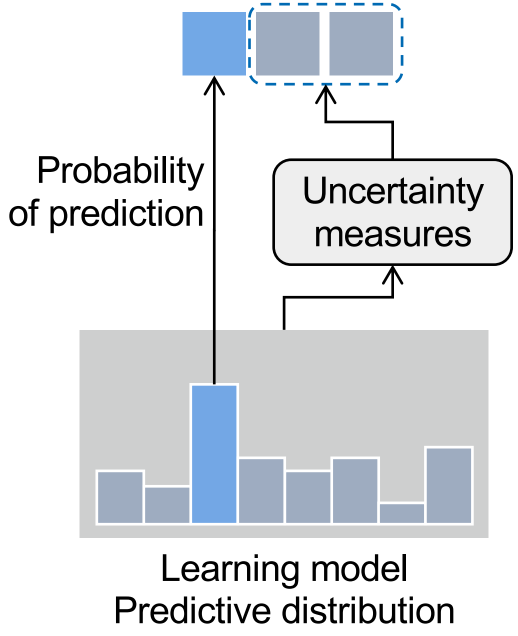
- Incorporates intra-category features to capture category specific uncertainty.
- Incorporates inter-category features to capture the correlation between categories.
- The combined intra- and inter-category encodings used for predicting the confidence.
- Objective: Predict whether the prediction of the learning model is correct.



Confidence estimation

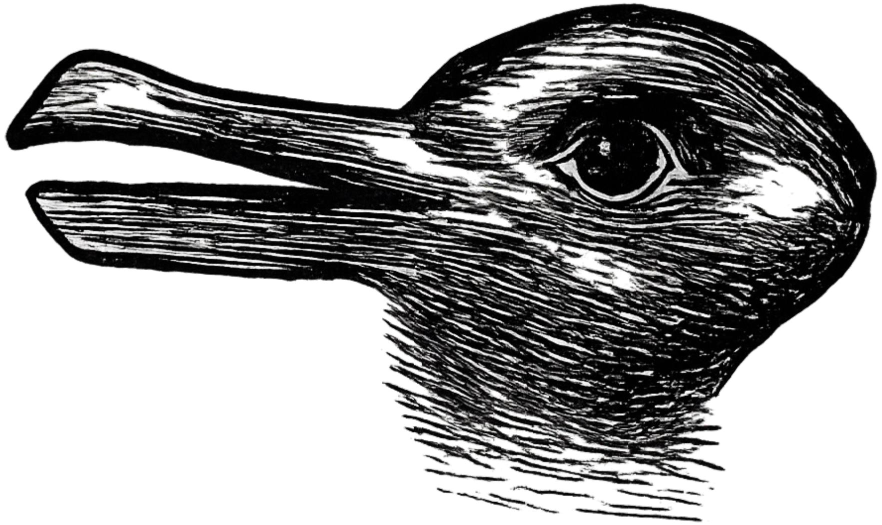
The uncertainty features

- Probability of the prediction / label.
- Uncertainty features extracted from the predictive distribution of the learning and noisy models:
  - Total Uncertainty (Entropy)
  - Knowledge Uncertainty



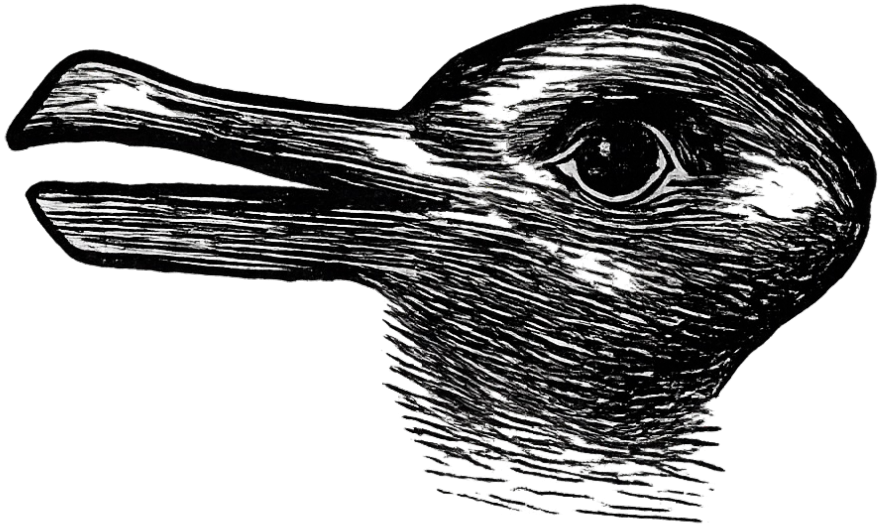
Uncertainty

Heinrich Heine
Universität Düsseldorf



Joseph Jastrow, 1899

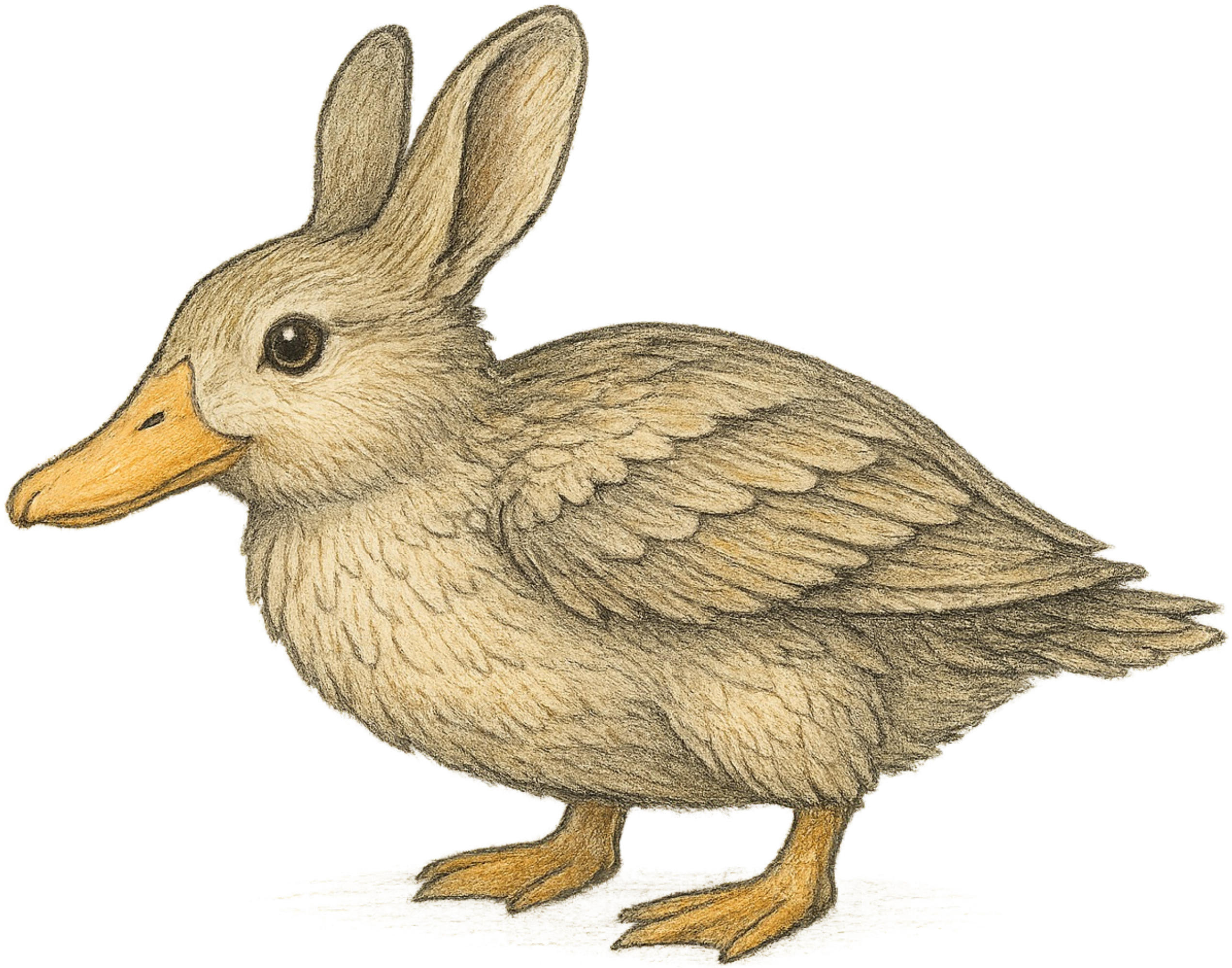
Uncertainty



Duck	Rabbit
0.5	0.5

Joseph Jastrow, 1899

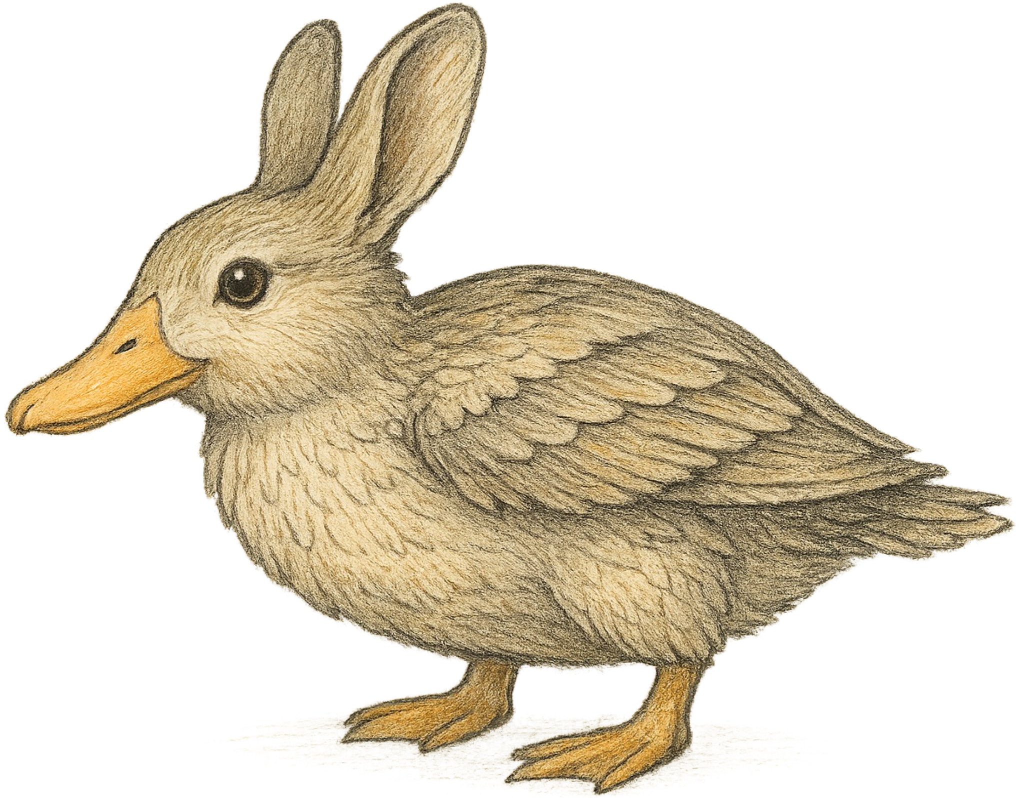
Uncertainty



ChatGPT 4o, 2025

Uncertainty

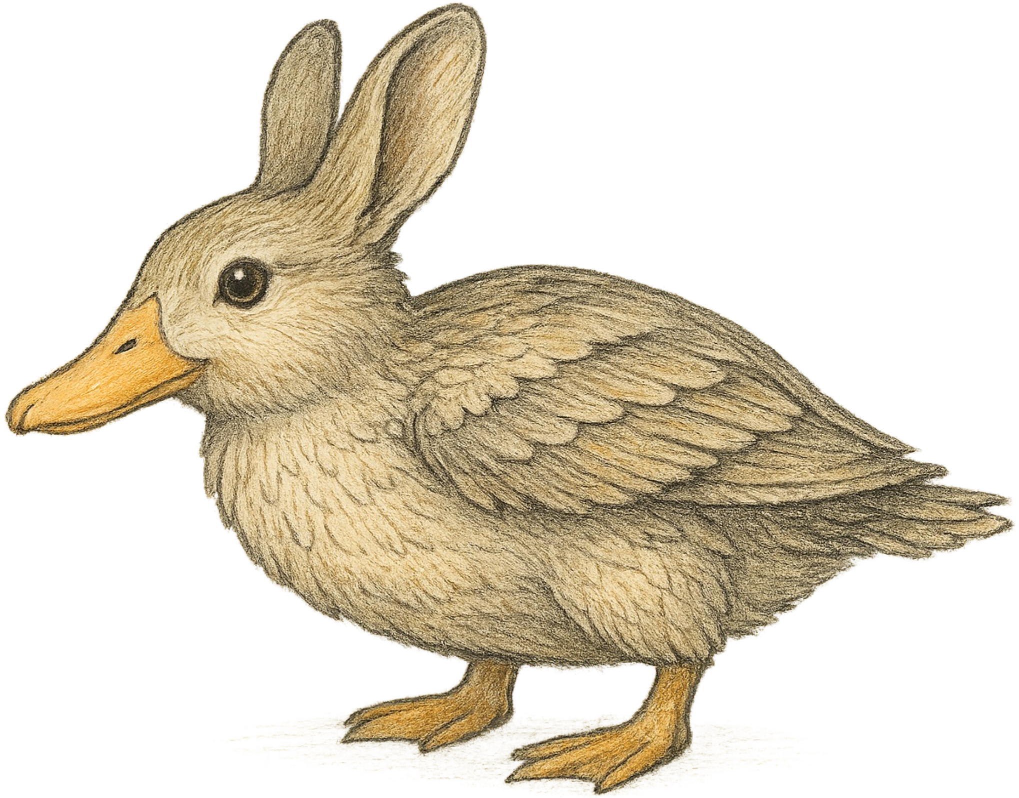
hhu
Heinrich Heine
Universität Düsseldorf



Duck	Rabbit
0.7	0.3

ChatGPT 4o, 2025

Uncertainty



Duck	Rabbit	Dubbit
0.08	0.02	0.9

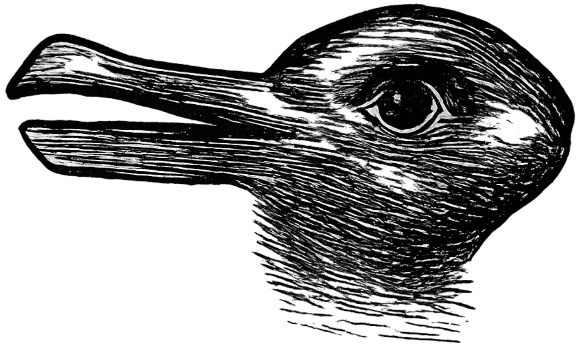
ChatGPT 4o, 2025

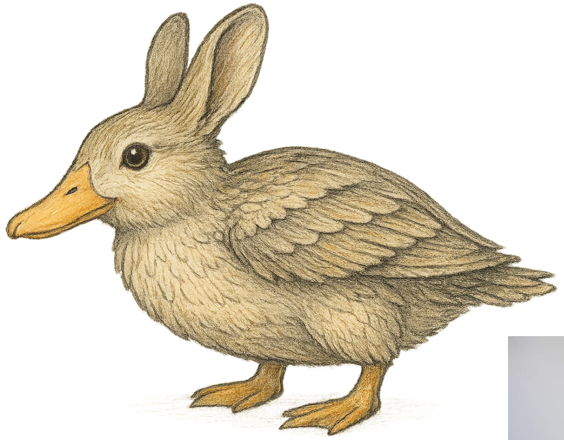
Uncertainty

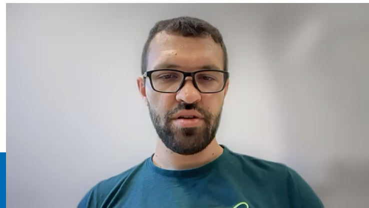
- Uncertainty in Machine Learning models stem from two main sources:

- **Data** Uncertainty: Uncertainty which stems from **ambiguity in the data**.

- **Knowledge** Uncertainty: Uncertainty which stems from a **lack of knowledge** within the model.



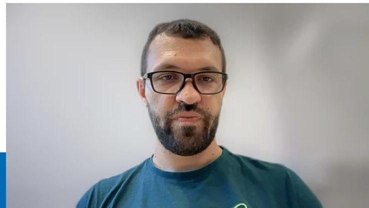




Uncertainty Estimation

- The **source of uncertainty** can be distinguished using an ensemble of models.
  - The **entropy** in the predictive distribution is the **total uncertainty**.
  - The **variation** between ensemble member predictions is the **knowledge uncertainty**.
  - Data uncertainty is the difference between these two.
  - This is **computationally expensive** especially for language models.

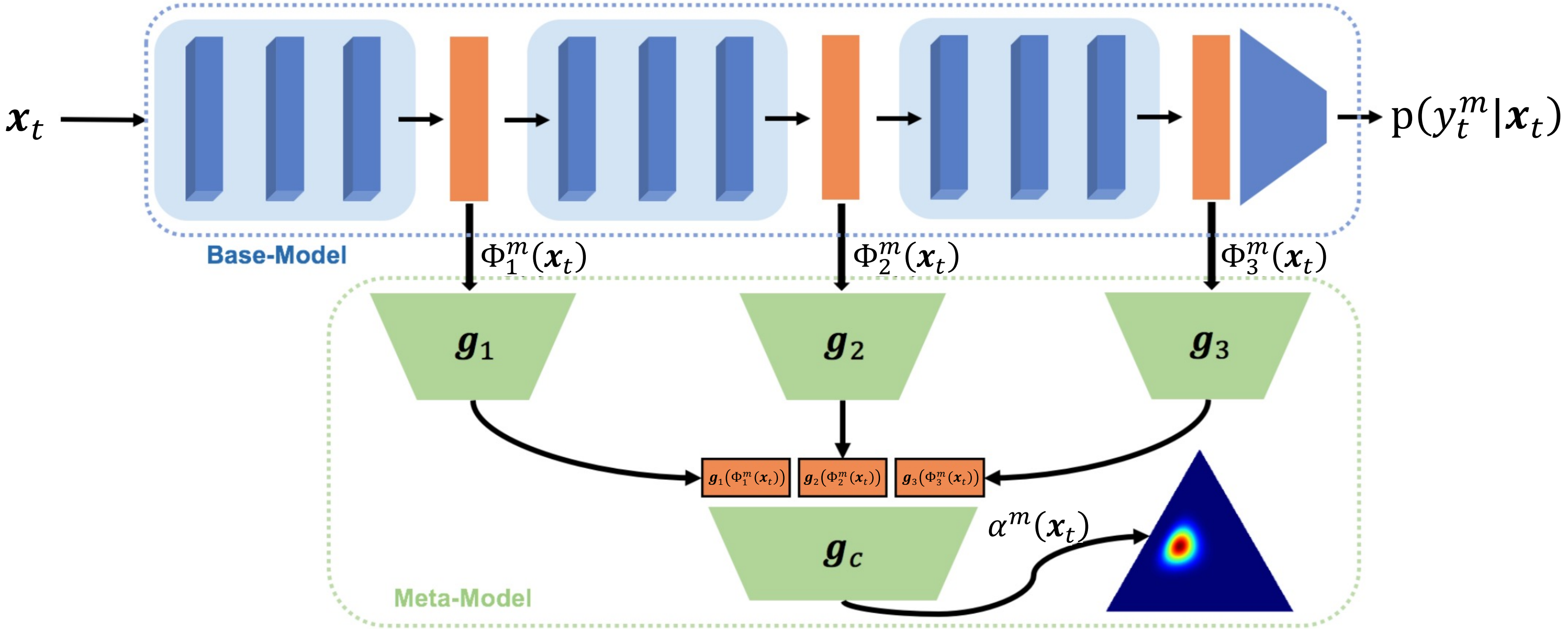
- Alternative: Learning a **Dirichlet distribution**, $\mathrm{Dirichlet}(\boldsymbol{\alpha})$, over the probability simplex.
  - The **mean predictive distribution** provides a measure of **total uncertainty**.
  - The **variation** under this distribution provides a measure of **knowledge uncertainty**.



Uncertainty Estimation

Heinrich Heine
Universität Düsseldorf

- The **variation between embeddings** at different layers allows accurate post-hoc uncertainty learning.



Post-hoc Uncertainty Learning Using a Dirichlet Meta-Model, Shen, et. al. 2023

Uncertainty Estimation Objective

- ELBO Objective:

$$\mathcal{L}\left(\boldsymbol{\theta}^{(\mathrm{meta})}; \mathcal{D}\right)$$

- Expected **Likelihood**: $= \mathbb{E}_{\mathrm{p}(\boldsymbol{x},y|\mathcal{D})}\left[\mathbb{E}_{\mathrm{p}\left(\boldsymbol{\pi}|\boldsymbol{x},\boldsymbol{\theta}^{(\mathrm{meta})}\right)}[-\log \mathrm{p}(y|\boldsymbol{\pi})]\right]$
- **KL Penalty**: $+\lambda\mathbb{E}_{\mathrm{p}(\boldsymbol{x},y|\mathcal{D})}\left[D_{KL}\left[\mathrm{p}\left(\boldsymbol{\pi}|\boldsymbol{x},\boldsymbol{\theta}^{(\mathrm{meta})}\right) || \mathrm{p}(\boldsymbol{\pi}|\boldsymbol{\beta})\right]\right]$

- Challenge: We require an **informative** prior $\mathrm{Dirichlet}(\boldsymbol{\beta})$.

Dynamic Priors

Heinrich Heine Universität Düsseldorf

- Train an ensemble of E models on a **small subset** of the training data.
- Using predictions from this ensemble to produce the prior using **Sterlings Approximation**:

$$\boldsymbol{\beta} = \beta_0(\boldsymbol{x})\tilde{\boldsymbol{\pi}}(\boldsymbol{x}), \text{ where}$$

$$\tilde{\boldsymbol{\pi}}(\boldsymbol{x}) = \frac{1}{E}\sum_{e=1}^{E} \pi_k^{(e)}(\boldsymbol{x})$$ (Mean – Total Uncertainty)

$$\beta_0(\boldsymbol{x}) = \frac{K-1}{2\sum_{k=1}^{K} \tilde{\pi}_k(\boldsymbol{x}) d_k(\boldsymbol{x})}, \text{ where}$$

$$d_k(\boldsymbol{x}) = \log \tilde{\pi}_k(\boldsymbol{x}) - \frac{1}{E}\sum_{e=1}^{E} \log \pi_k^{(e)}(\boldsymbol{x})$$ (Variation – Knowledge Uncertainty)

- This prior is used in the first active learning step.
- After this the predicted posterior of the previous active learning step is used as a prior in order to **update the beliefs of the model**.

Experiments

Heinrich Heine
Universität Düsseldorf

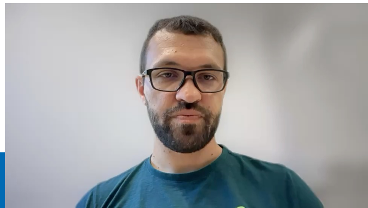
Machine Translation

- Model: `Ensemble T5-small` encoder-decoder transformer.

- Uncertainty Estimation:
  - Total Uncertainty: Entropy of the predictive distribution.
  - Knowledge Uncertainty: Mutual information between predictive distribution and ensemble members.

- Dataset: `WMT17 DE-EN` for German to English translation.

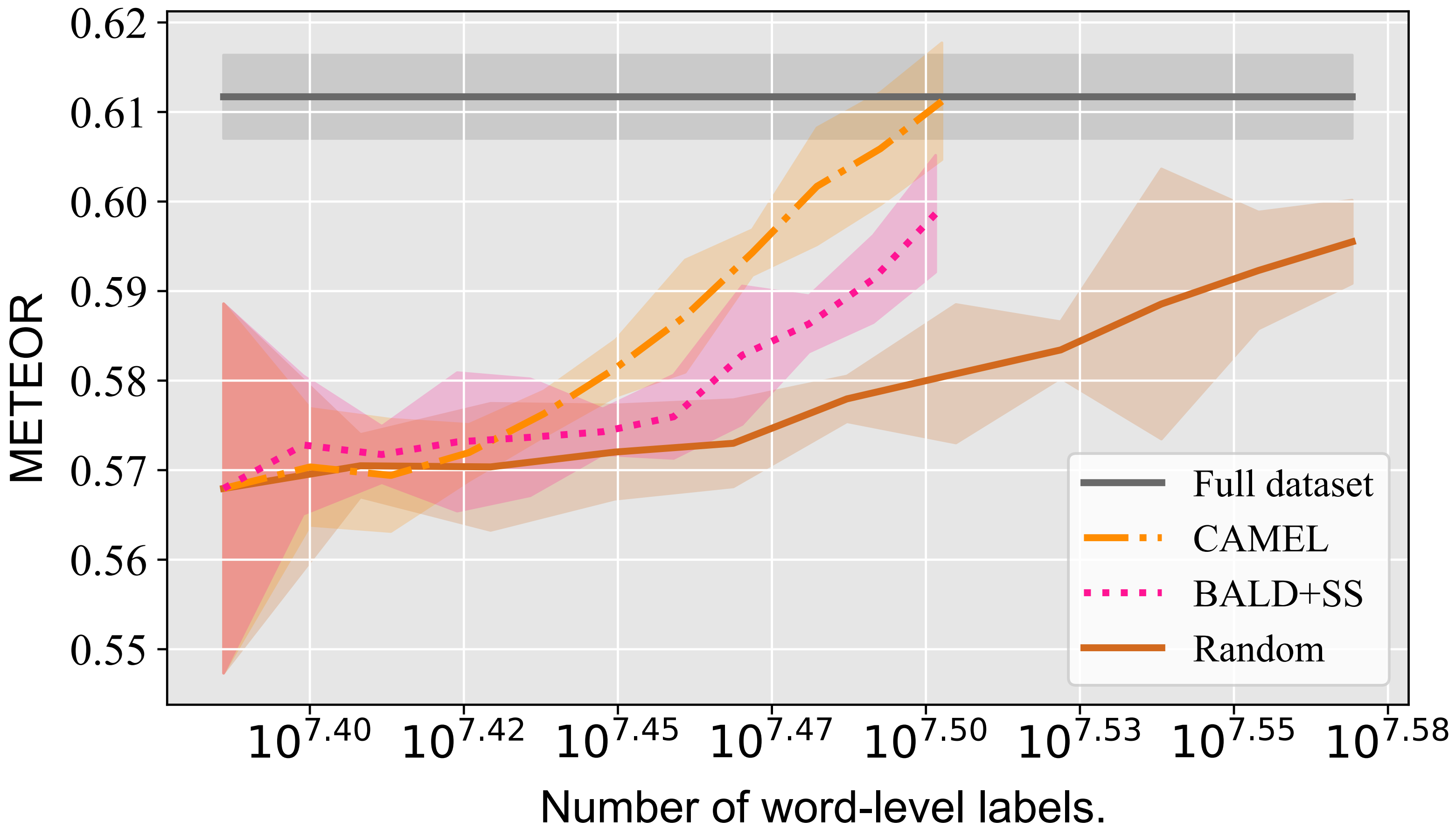
- Confidence Estimation Model - Simplified model for the single category sequential task.

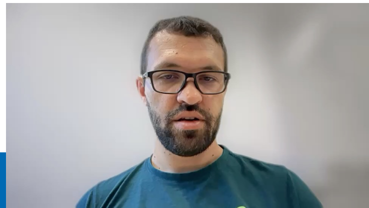


Results

hhu
Heinrich Heine
Universität Düsseldorf

Machine Translation



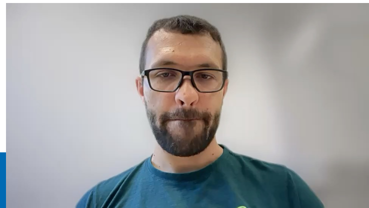


Experiments

Heinrich Heine
Universität Düsseldorf

Dialogue Belief Tracking

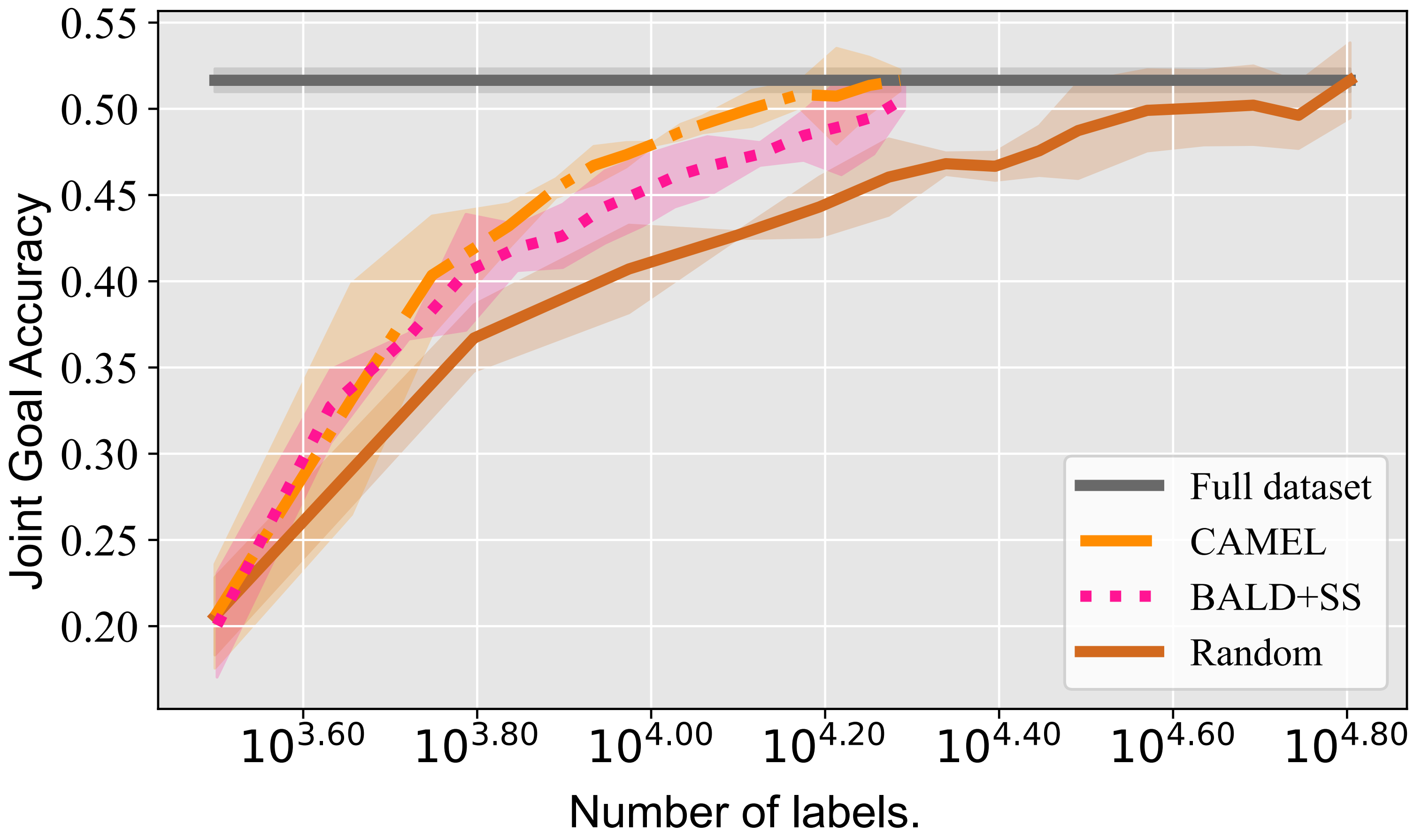
- Model
  - `Ensemble-SetSUMBT`
  - `Meta-Uncertainty SetSUMBT`
- Uncertainty Estimation:
  - Total Uncertainty: Entropy of the predictive distribution or Entropy within the Dirichlet distribution.
  - Knowledge Uncertainty: Mutual information between predictive distribution and ensemble members or knowledge uncertainty estimate using the Dirichlet distribution.
- Datasets:
  - `MultiWOZ 2.1` (Noisy version)
  - `MultiWOZ 2.4` (Cleaned version)
- Metric: Joint goal accuracy.

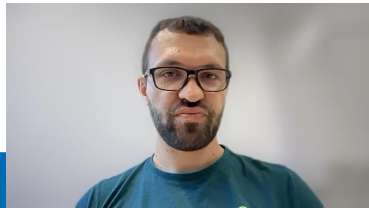


Results

hhu
Heinrich Heine
Universität Düsseldorf

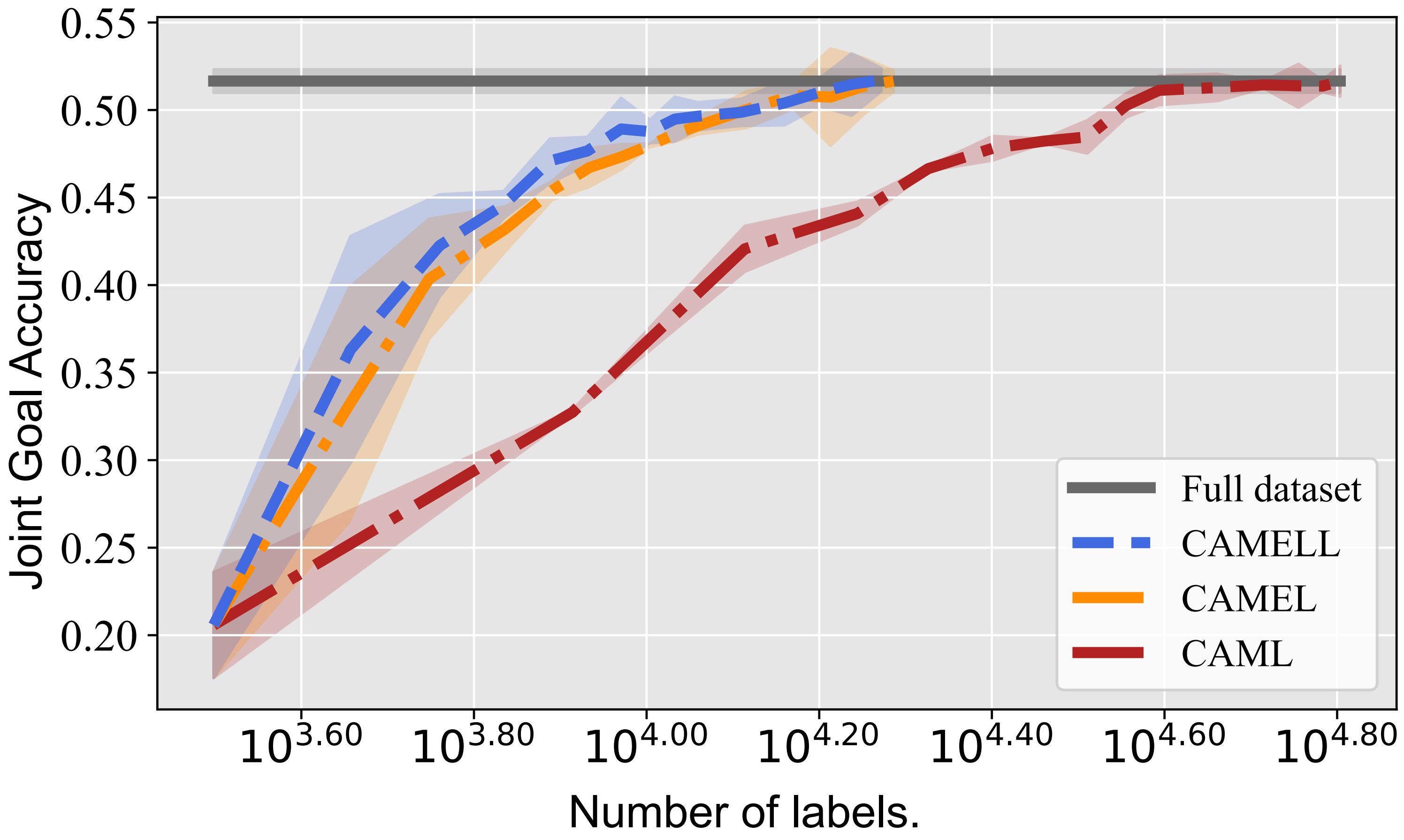
Dialogue Belief Tracking - Ensemble

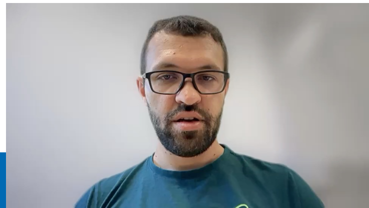




Results

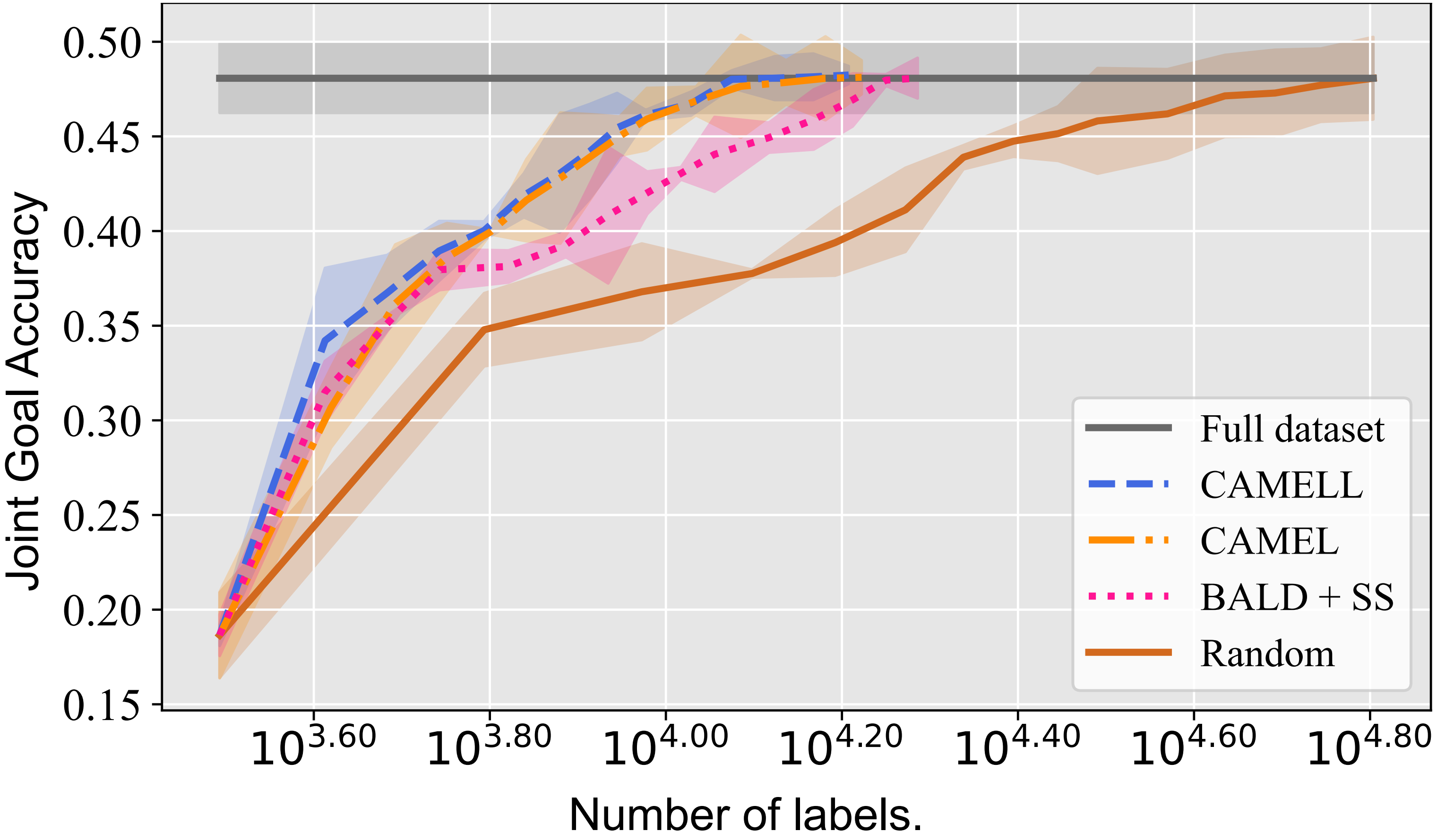
Dialogue Belief Tracking – Ensemble Ablation

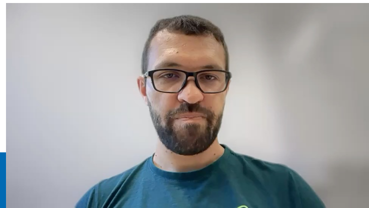




Results

Dialogue Belief Tracking – PostHoc Meta Model





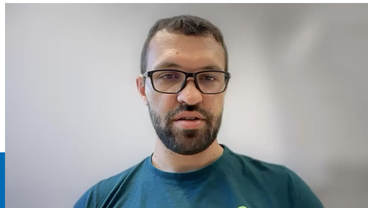
Label Correction

Label Correction Process

- Steps:
  - Select labels in the dataset with label confidence below the threshold.
  - If the prediction confidence is greater than the label threshold replace the label.
- Noisy `MultiWOZ 2.1` dataset used to train the ensemble SetSUMBT model.
- `Ensemble SetSUMBT` used for label correction.
- `Trippy` (a state-of-the-art span prediction based) DST used to evaluate the quality of the corrections.

Results

Model	Label Corr.	MultiWOZ 2.1	MultiWOZ 2.4
CE-SetSUMBT	None	51.79	61.63
	Offline	52.83	**63.32**
TripPy	None	55.28	64.45
	Offline	56.11	**66.02**

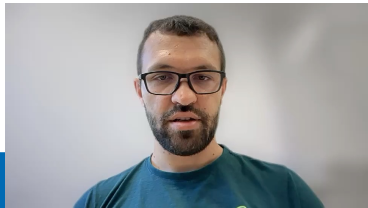


Label Correction

Heinrich Heine
Universität Düsseldorf

Examples

Conversation	MultiWOZ 2.1 Labels and Corrections
User: I would like to find a place that serves **moderately** priced **Chinese** food.	{Restaurant: {Food: Chinese, (95%) Price: Moderate, (94%) Day: Tuesday, (11%) Day: not_mentioned}} (72%)
User: I need a train leaving on **Friday** and I want to get there by **21:30**. Leaving **Broxbourne**.	{Train: {Dept.: Broxbourne, (94%) Day: Friday, (95%) Arrive by: 21:20, (1%) Arrive by: 21:30}} (83%)



Conclusion

Heinrich Heine Universität Düsseldorf

✓ Selective self-supervision **improves efficiency**.

✓ CAMELL SetSUMBT achieve 95% of a tracker's full-training dataset performance using **merely 16% of the expert-provided labels**.

✓ CAMELL can be applied to **automatically correct labels** in a dataset.