Towards a robust experimental framework & benchmark for lifelong language learning

Aman Hussain | Nithin Holla | Pushkar Mishra
Helen Yannakoudakis | Ekaterina Shutova

The Problem

Serious flaws
in
lifelong learning experiments

Our Solution

Degree-of-Belief:
a robust
experimental framework

The Problem

Model Training
$f : \mathcal{X} \to \mathcal{Y}$

Data Stream
$\{(x_i, y_i)\}_{i=1}^{n}$

Task Distributions
$\{P_j\}_{j=1}^{m}$

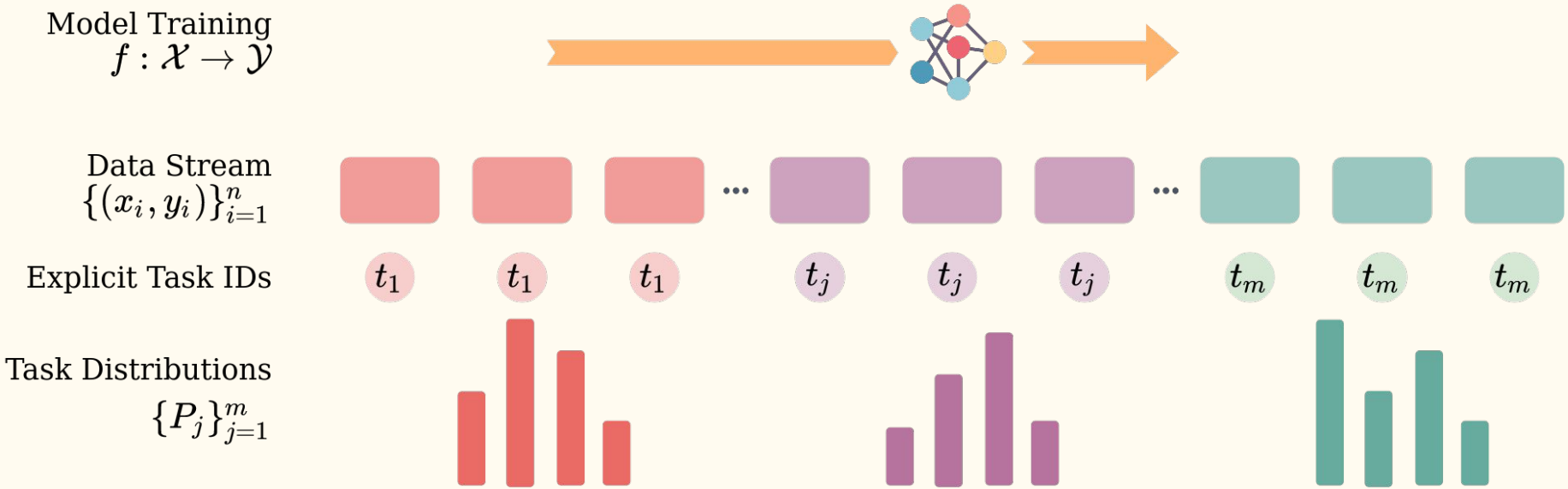
Formulation of General Lifelong Learning

Model Training
$f : \mathcal{X} \to \mathcal{Y}$

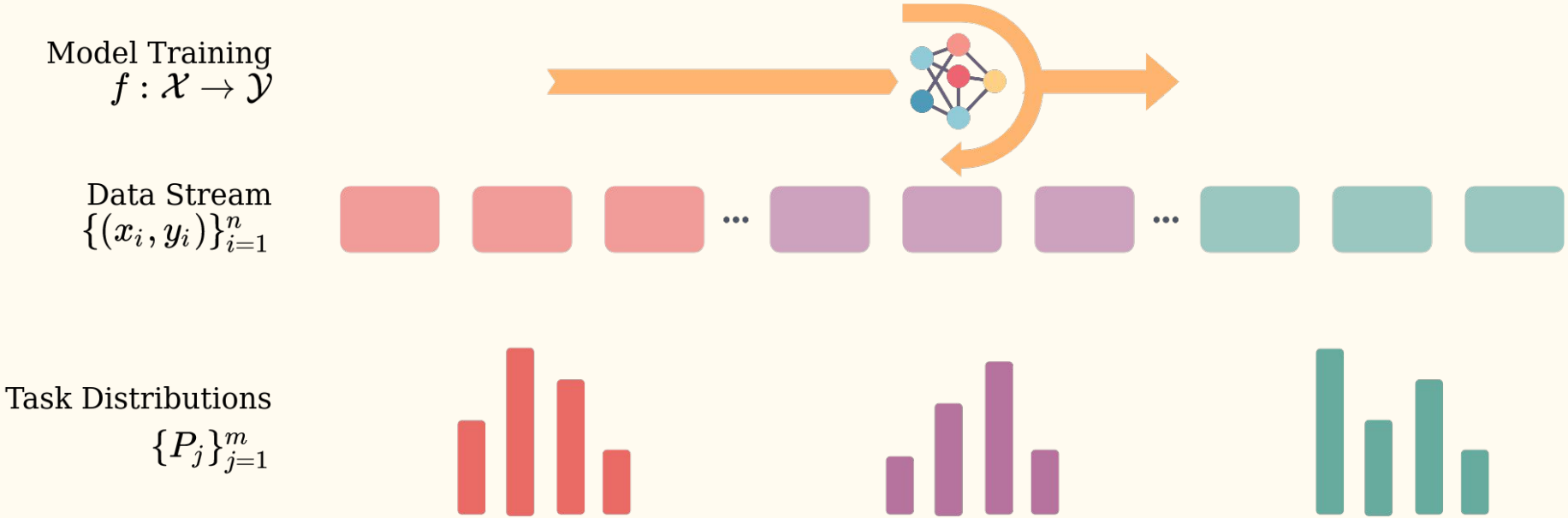
Data Stream
$\{(x_i, y_i)\}_{i=1}^{n}$

Single Task
Distribution P

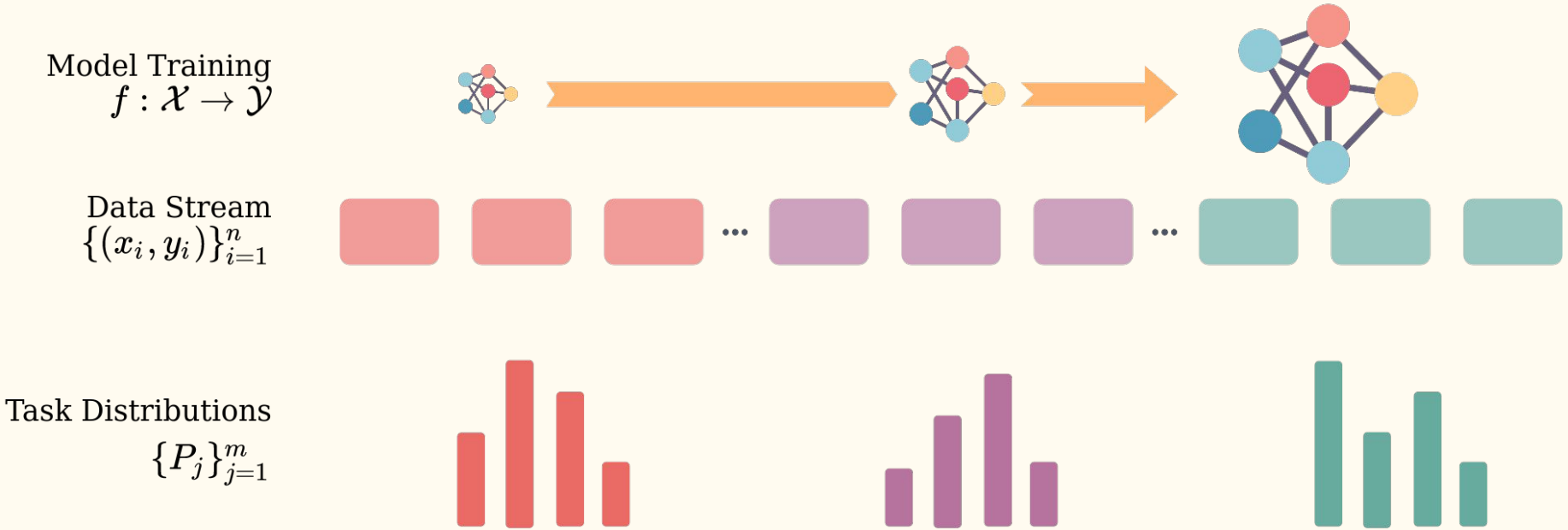
Desideratum#1: Task Plurality



Desideratum#2: Task Generality



Desideratum#3: Online Stream



Desideratum#4: Space Complexity

Limitations of existing work

- Lack of Task Plurality
- Lack of Task Generality

Task Plurality

- ⛔ CALM [1]
- ⛔ Lifelong Few Rel [2]
- ⛔ Lifelong Question Relations [2]
- ⛔ Lifelong Extractive QA [3]
- ⭕ Lifelong Text Classification [3]

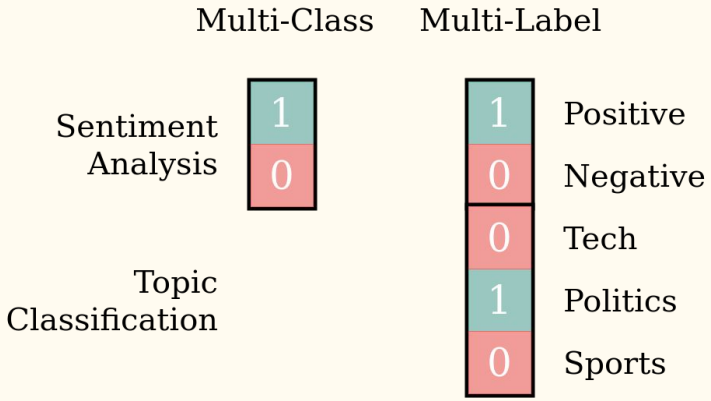
Task Generality

- Lifelong Text Classification (LTC)

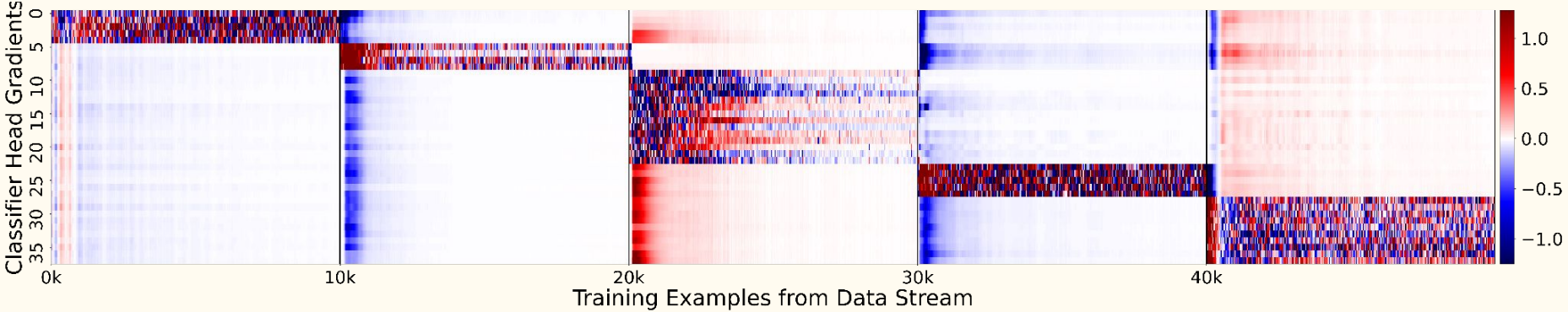
Misleading gradients in the LTC dataset

Lifelong learning over
many classification tasks
is a multi-label problem.

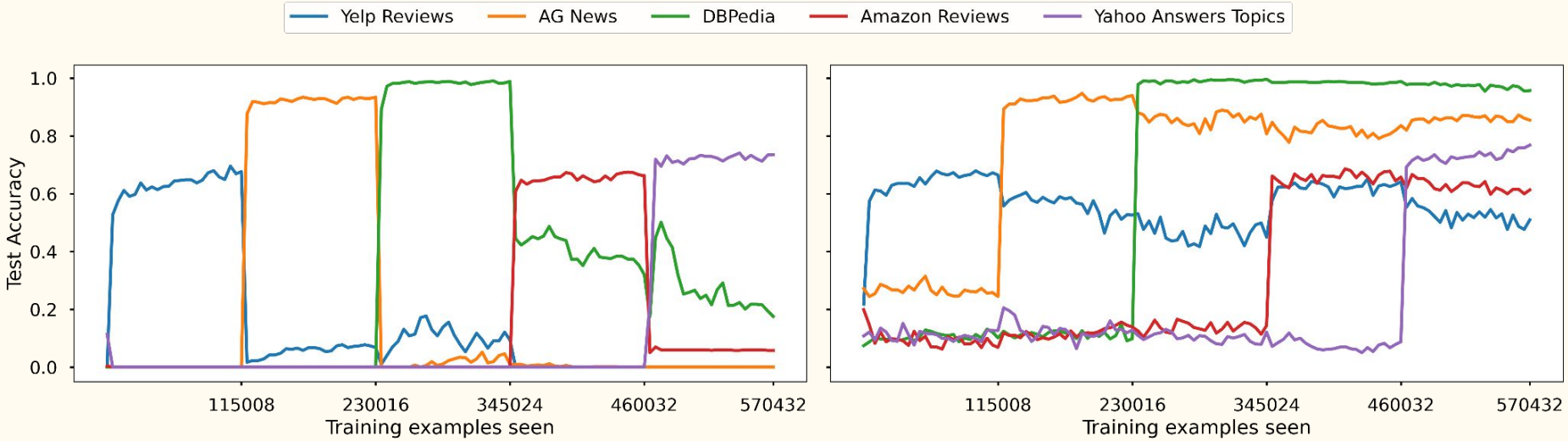
Framing it as a
multi-class problem
results in misleading gradients.



Misleading Gradients => Catastrophic Forgetting



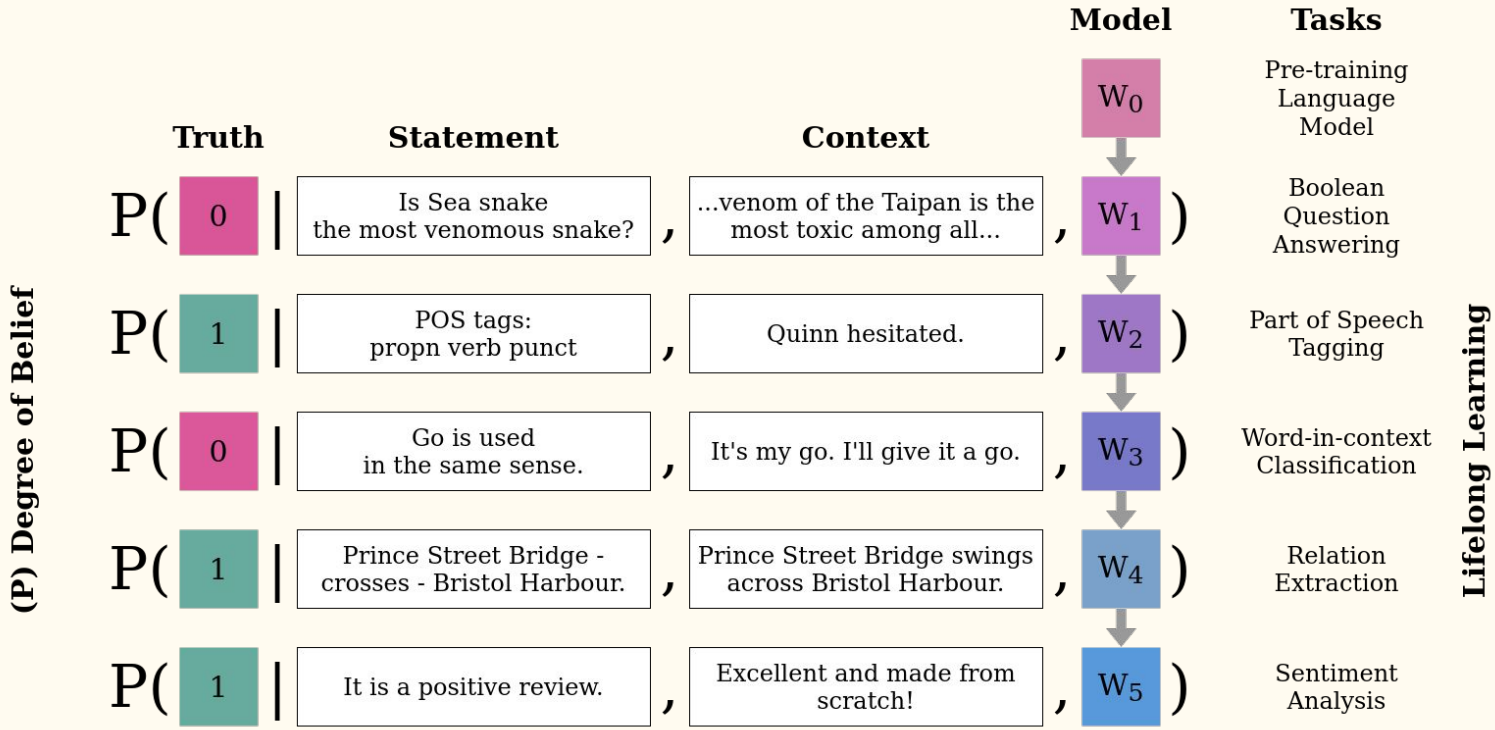
LTC dataset leaks explicit task identifiers



If the task identifiers are explicitly known,
we can train multiple classification heads, one for each task, to avoid catastrophic interference.

Multiple tasks
without explicitly
or indirectly leaking
task identifiers?

Our Solution



Degree-of-Belief framework for General Lifelong Learning

Evaluation Benchmarks

Baselines

Single-task learning

Multi-task learning

Experience replay

Metrics

Forgetting

Intransigence

Final Avg. Accuracy

Area Under the
Lifelong Test Curve

Data streams

Standard stream : 5 tasks

Long stream : 10 tasks

Large stream: 50k examples

Larger stream: 100k

Multidomain streams

Multilingual streams

Linguistic Hierarchy stream

Key Takeaways

- Memory-based techniques cannot outperform simple Experience replay
- Pre-trained transformers can identify tasks without explicit task ids
- They can learn to use a subset of its parameter space for each task

References

[1] Evaluating Online Continual Learning with CALM, G. Kruszewski et al. (2021)

[2] Sentence Embedding Alignment for Lifelong Relation Extraction, H. Wang et al. (2010)

[3] Episodic Memory in Lifelong Language Learning, C.d.M. d'Autume et al. (2019)