A Appendix

A.1 Performance distribution of stream permutations

To determine the most challenging order of tasks in our **Standard Stream**, we measure the performance on all metrics formulated in our evaluation scheme across all the possible permutations of these 5 tasks, i.e., 120. The distribution of each evaluation metric across these permutations is shown in Figure 6.

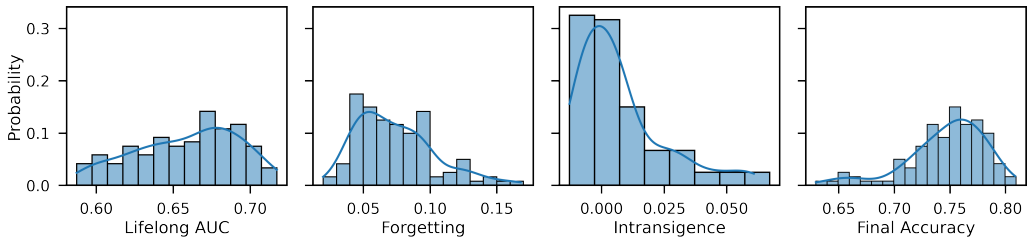

Figure 6: Probability distribution of model performance across all order permutations in our data stream.

A.2 Baseline experiment results on all data streams

We evaluate the performance of all our baseline methods on all the benchmark data streams. Table 2 shows the evaluation measurements on the data streams that focus on specific properties of lifelong learning or language. Table 3 shows the evaluation measurements on the data streams that are designed to capture the representative properties of each evaluation metric.

A.3 Area Under the Lifelong Test Curve

To visualize our online evaluation metric, Area Under the Lifelong Test Curve, we plot the average test accuracy on all tasks throughout the training process. Figure 7 shows the average test curve for both lifelong learning and multi-task learning on the **Standard Stream**. Figure 8 shows the test curve of each task.

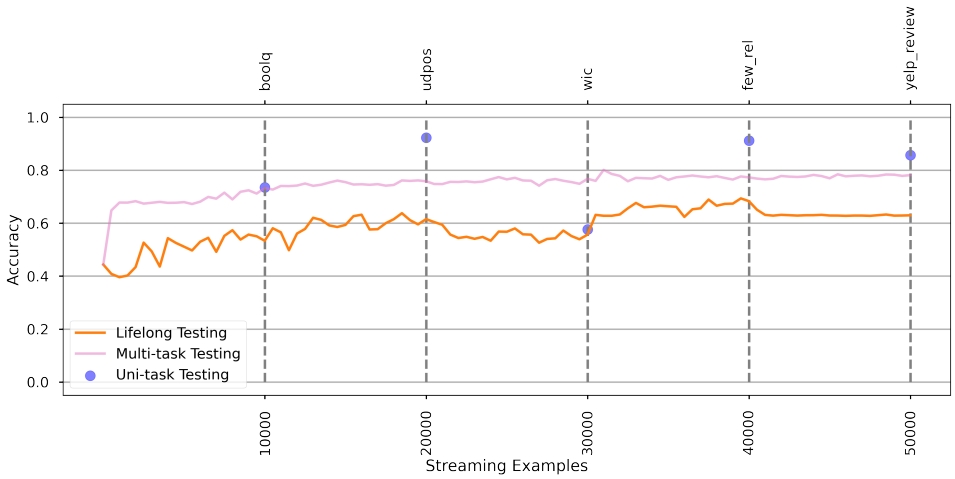

Figure 7: Average test curve on the **Standard Stream** for lifelong learning vs multi-task learning using pre-trained BERT. The final test accuracies of the single-task models are also shown at the task boundaries.

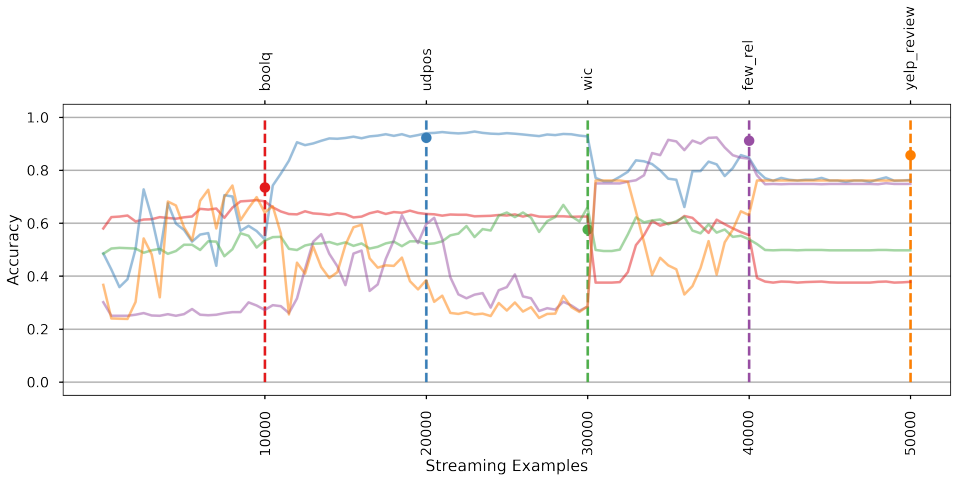

Figure 8: Task-wise test curves on the **Standard Stream** for lifelong learning vs multi-task learning using pre-trained BERT. The final test accuracies of the single-task models are also shown at the task boundaries.

A.4 Hardware and runtime

The train–test runtime of one lifelong learning experiment is approximately 2-3 hours on the **Standard** and other data streams, and 9 hours on the **Long Stream** using a BERT base architecture with batch size of 25 on an Nvidia Titan RTX (24GB GDDR6).

A.5 Impact of replay interval

Here, we investigate the impact of the replay interval. Specifically, we keep P_{write} and N_{replay} the same and vary the $R_{interval}$ in $[500, 1000, 2000, 10000]$; therefore, we end up with the following replay rates: $[20\%, 10\%, 5\%, 1\%]$. Table 4 presents the performance of these different replay rates. Using a two-tailed paired t-test ($\alpha = 0.05$), the replay rates of 5%, 10% and 20% lead to a significant decrease in Forgetting over lifelong learning with p-values of 0.007, 0.016 and 0.005 respectively. Similarly, we find that the replay rates of 1%, 5% and 20% lead to significant improvements in Final Accuracy over lifelong learning with p-values of 0.044, 0.012 and 0.016 respectively. The replay rates of 1% and 10% fail the significance test on these two metrics respectively by a very small margin. However, based on the confidence intervals, we find that simply increasing the replay interval does not result in a consistent improvement of the metrics. We might be able to further improve the performance of experience replay through further hyperparameter optimization; however, due to time and resource constraints, we leave it for future work.

A.6 Analysis of class imbalance

In the following Figures (9–13), we plot the F1 scores, true positive rates and true negative rates on the 'true' class (i.e., where the input statements are true; Figure 3)[3] during lifelong learning for all the tasks in the Standard stream.

A.7 Test and train set analysis

Our test set is formed by randomly selecting 1k examples from the full test set due to the prohibitively expensive computation required to record the evaluation metrics; for example, it takes 2 hours to run an experiment using our sampled test set vs 51 hours when using the full test set on the Standard stream. In this section, we confirm that our test set distribution and results are aligned with those on the full test set on the Standard stream. In Table 7, we verify that the label distribution of our test set matches the distribution of the full test set. In Table 6, we validate that the test results / patterns on the full test set closely resemble those on our test set. Our training set is similarly formed by randomly sampling 10k examples, and upsampling of smaller datasets where necessary. In Table 8, we verify

[3]The only exception is BoolQ where 'false' is the minority class.

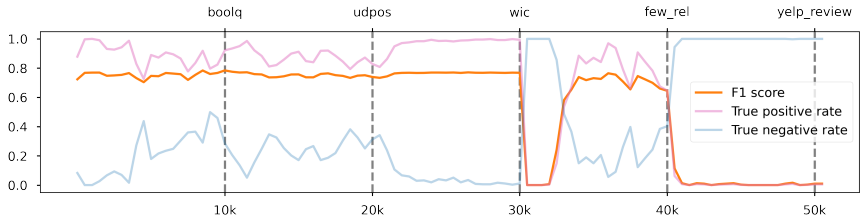

Figure 9: F1 scores, true positive rates and true negative rates during lifelong learning for BoolQ in the Standard stream.

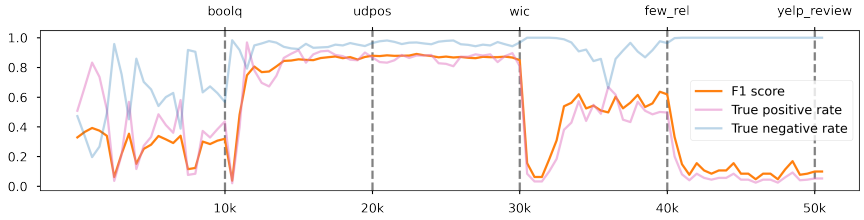

Figure 10: F1 scores, true positive rates and true negative rates during lifelong learning for UDPOS in the Standard stream.

that the label distribution of our train set matches the distribution of the full train set for the Standard stream.

A.8 Impact of cues in the input statements:

To avoid explicit cues in the implicit task identifiers being memorized by the model, we use semantically similar but also syntactically different statements when encoding a task in our framework. However, we run an additional experiment to examine the extent to which the model relies on specific cues in the input statements, such as punctuation marks and the use of certain keywords. Specifically, we increase the number of statement templates used to encode different tasks in the Standard stream. We remove the punctuation marks in some of the statement templates of the BoolQ, UDPOS and FewRel tasks. We also replace keywords such as 'positive' with 'good' and 'negative' with 'bad' in some of the Yelp Review task statements. In Table 10, we present the results on the Standard stream. We find that this does not impact performance nor affect our conclusions. In single-task learning, the Final Accuracy remains almost the same (80.64 vs 80.01). Similarly, in multi-task learning, the AULTC (74.34 vs 73.80) and Final Accuracy (77.72 vs 77.76) stay roughly the same. However, Forgetting (3.95 vs 1.99) and Intransigence (6.17 vs 5.77) decrease, which suggests that more varied statements can be beneficial for these models. In lifelong learning, the AULTC (60.95 vs 61.02) stays the same. The other metrics such as the Final Accuracy stay within the confidence interval of the original measurements in Table 1. This suggests that the model does not rely solely on specific cues in the input implicit statements when attempting to identify tasks and make predictions.

A.9 Gradient overlap

In this section, we plot the gradient overlap during lifelong learning and multi-task learning for all the streams. The y-axis (log-scale) shows the number of parameters that are shared between two consecutive sub-networks, i.e., parameters which received non-zero gradients at time t and $t-1$.

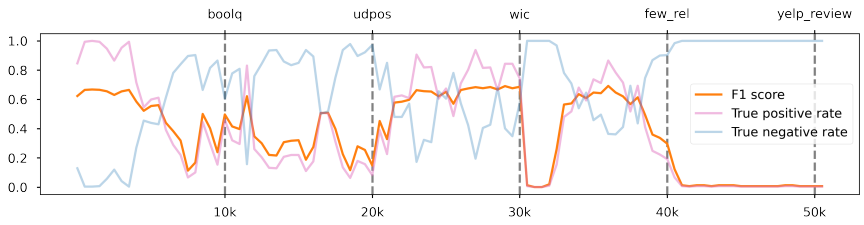

Figure 11: F1 scores, true positive rates and true negative rates during lifelong learning for WIC task in the Standard stream.

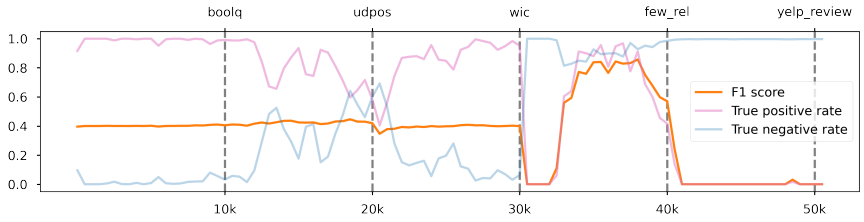

Figure 12: F1 scores, true positive rates and true negative rates during lifelong learning for FewRel in the Standard stream.

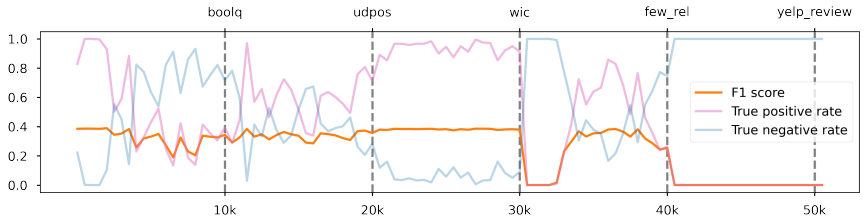

Figure 13: F1 scores, true positive rates and true negative rates during lifelong learning for Yelp Reviews in the Standard stream.

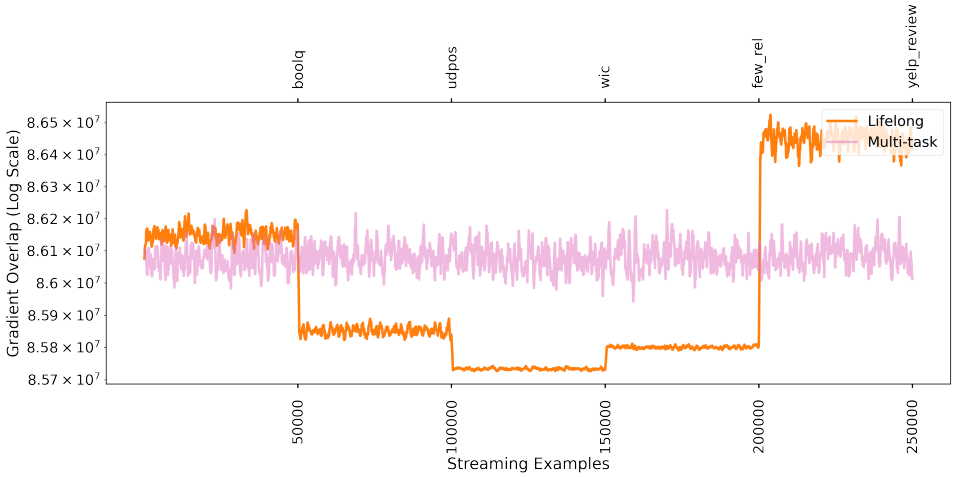

Figure 14: Gradient overlap on the Large Stream.

Table 2: Baseline results on data streams focusing on different data settings.

Stream	Method	AULTC	Forgetting	Intransigence	Final Accuracy
Large	Lifelong	67.96	7.50	0.70	77.58
	Replay (1%)	69.93	3.77	0.72	81.05
	Replay (5%)	70.51	3.80	1.18	81.37
	Replay (10%)	70.23	3.24	2.83	81.60
	Replay (20%)	69.91	3.14	1.56	81.29
	Single-task	–	–	–	84.19
	Multi-task	77.89	5.08	7.10	79.68
Larger	Lifelong	67.04	11.14	-0.73	74.46
	Replay (1%)	70.79	4.05	-0.96	81.09
	Replay (5%)	71.19	4.41	-1.60	81.18
	Replay (10%)	69.51	5.41	1.37	79.56
	Replay (20%)	69.65	6.17	0.41	79.17
	Single-task	–	–	–	82.41
	Multi-task	80.38	3.86	2.14	81.57
Long	Lifelong	71.19	8.14	-1.31	81.55
	Replay (1%)	72.11	5.13	-0.73	84.40
	Replay (5%)	75.46	2.59	-0.38	87.17
	Replay (10%)	74.46	2.95	-0.51	86.69
	Replay (20%)	74.85	1.59	-0.57	88.03
	Single-task	–	–	–	86.51
	Multi-task	83.71	1.91	1.78	86.52
Linguistic Hierarchy	Lifelong	67.62	9.67	1.65	73.10
	Replay (1%)	67.86	4.45	3.83	78.90
	Replay (5%)	71.33	3.07	0.94	80.12
	Replay (10%)	72.11	3.01	2.16	80.19
	Replay (20%)	71.59	3.23	0.93	80.41
	Single-task	–	–	–	79.44
	Multi-task	74.12	5.44	6.26	76.76
Multilingual A	Lifelong	94.00	1.35	-0.50	95.38
	Replay (1%)	93.85	1.26	-0.77	95.44
	Replay (5%)	94.11	0.50	-1.01	96.30
	Replay (10%)	94.36	0.44	-0.69	96.36
	Replay (20%)	94.52	1.05	-0.99	95.78
	Single-task	–	–	–	95.35
	Multi-task	95.22	0.32	-0.87	96.64
Multilingual B	Lifelong	84.11	13.21	0.02	83.89
	Replay (1%)	85.73	9.01	-0.99	87.81
	Replay (5%)	87.71	3.64	-0.76	92.93
	Replay (10%)	88.84	4.16	-0.60	92.57
	Replay (20%)	89.52	1.31	0.07	95.25
	Single-task	–	–	–	95.60
	Multi-task	93.85	0.92	0.60	95.88
Multidomain A	Lifelong	83.37	6.56	1.51	86.71
	Replay (1%)	83.05	8.10	1.44	84.96
	Replay (5%)	85.44	3.10	0.33	90.61
	Replay (10%)	84.41	1.57	0.50	92.03
	Replay (20%)	84.35	1.56	1.72	91.58
	Single-task	–	–	–	92.28
	Multi-task	91.86	1.33	-0.67	93.26
Multidomain B	Lifelong	66.43	0.00	0.00	66.59
	Replay (1%)	72.77	2.55	-8.24	77.21
	Replay (5%)	73.26	1.08	-8.67	78.92
	Replay (10%)	74.44	2.04	-9.65	78.54
	Replay (20%)	73.97	1.65	-9.91	79.18
	Single-task	–	–	–	66.59
	Multi-task	74.02	1.70	-9.55	78.87

Table 3: Baseline results on the metric-specific data streams.

Stream	Method	AULTC	Forgetting	Intransigence	Final Accuracy
AULTC	Lifelong	67.52	5.73	-0.25	77.26
	Replay (1%)	68.70	11.70	-0.45	72.27
	Replay (5%)	68.25	3.66	1.75	79.65
	Replay (10%)	70.03	1.82	2.44	81.95
	Replay (20%)	69.59	2.32	1.39	81.45
	Single-task	–	–	–	80.06
	Multi-task	74.77	4.13	4.25	78.14
Forgetting	Lifelong	67.23	4.86	0.91	78.49
	Replay (1%)	66.97	6.29	-0.28	76.25
	Replay (5%)	71.90	1.25	0.13	81.72
	Replay (10%)	71.56	2.12	-0.03	81.43
	Replay (20%)	71.70	1.76	1.99	81.57
	Single-task	–	–	–	80.06
	Multi-task	74.77	4.13	3.26	78.14
Intransigence	Lifelong	60.01	13.87	6.08	64.24
	Replay (1%)	61.37	4.34	4.01	74.69
	Replay (5%)	65.15	2.51	3.89	77.92
	Replay (10%)	67.83	4.57	2.93	77.88
	Replay (20%)	66.57	2.93	6.47	76.60
	Single-task	–	–	–	80.06
	Multi-task	74.77	4.13	4.25	78.14
Final Accuracy	Lifelong	64.69	6.08	3.47	75.95
	Replay (1%)	64.53	6.24	-0.07	76.33
	Replay (5%)	64.29	4.11	0.97	77.93
	Replay (10%)	65.48	3.72	0.80	78.46
	Replay (20%)	63.61	2.53	1.17	80.50
	Single-task	–	–	–	80.06
	Multi-task	74.77	4.13	5.67	78.14

Table 4: Comparison of different experience replay rates on the **Standard stream**.

	AULTC	Forgetting	Intransigence	Final Accuracy
Lifelong	60.95 ± 1.64	12.78 ± 3.30	1.52 ± 1.70	68.26 ± 3.87
Replay 1%	62.62 ± 1.11	7.13 ± 3.08	0.86 ± 1.17	74.70 ± 3.54
Replay 5%	63.28 ± 2.14	3.34 ± 1.01	1.12 ± 0.76	78.59 ± 1.26
Replay 10%	62.81 ± 2.14	4.35 ± 2.13	1.82 ± 2.95	76.21 ± 5.08
Replay 20%	63.99 ± 2.70	4.91 ± 4.17	1.30 ± 1.46	76.55 ± 4.47
Single task	–	–	–	80.64 ± 1.37
Multi-task	74.34 ± 0.30	3.95 ± 0.74	6.17 ± 1.37	77.72 ± 0.89

Table 5: Difference (Δ) in performance between lifelong learning and multi-task/single-task learning as we investigate the impact of different data stream sizes.

	Multi-task - Lifelong				Single-task - Lifelong
Stream	**ΔAULTC**	**ΔForgetting**	**ΔIntransigence**	**ΔFinal Accuracy**	**ΔFinal Accuracy**
Standard	13.39	-8.83	4.65	9.46	12.38
Large	9.93	-2.42	6.40	2.10	6.61
Larger	13.34	-7.28	2.87	7.11	7.95

Table 6: Lifelong learning metrics on the full test set of the Standard Stream.

	AULTC	Forgetting	Intransigence	Final Accuracy
Lifelong	59.12	14.04	1.55	68.80
Single task	–	–	–	79.32
Multi-task	74.24	2.09	4.46	77.83

Table 7: The label distribution of the full test set and our test set on the Standard stream.

Task	Testset	False	True	Total Size
BoolQ	Our testset	37.50%	62.50%	1,000
	Full testset	37.82%	62.17%	3,270
UDPOS	Our testset	75.00%	25.00%	1,000
	Full testset	75.07%	24.92%	1,035
WiC	Our testset	50.00%	50.00%	638
	Full testset	50.00%	50.00%	638
FewRel	Our testset	75.00%	25.00%	1,000
	Full testset	75.00%	25.00%	11,200
Yelp Review	Our testset	76.10%	23.90%	1,000
	Full testset	75.00%	25.00%	50,000

Table 8: The label distribution of the full train set and our train set for the Standard stream.

Task	Trainset	False	True	Total Size
BoolQ	Our trainset	37.51%	62.49%	10,000
	Full trainset	35.53%	58.74%	9,427
UDPOS	Our trainset	74.97%	25.03%	10,000
	Full trainset	75.00%	25.00%	3,176
WiC	Our trainset	50.03%	49.97%	638
	Full trainset	50.00%	50.00%	5,428
FewRel	Our trainset	75.36%	24.64%	10,000
	Full trainset	75.00%	25.00%	44,800
Yelp Review	Our trainset	76.40%	23.60%	10,000
	Full trainset	75.00%	25.00%	650,000

Table 9: Single-task accuracy for the tasks in the Standard stream.

Task	BoolQ	UDPOS	FewRel	WiC	Yelp Review
Accuracy	73.52%	92.32%	91.21%	57.61%	85.66%

Table 10: Lifelong learning metrics when including more diverse statement templates in the Standard stream.

	AULTC	Forgetting	Intransigence	Final Accuracy
Lifelong	61.02	15.20	2.48	65.20
Replay 10%	64.87	2.72	1.37	78.92
Single task	–	–	–	80.01
Multi-task	73.80	1.99	5.77	77.76

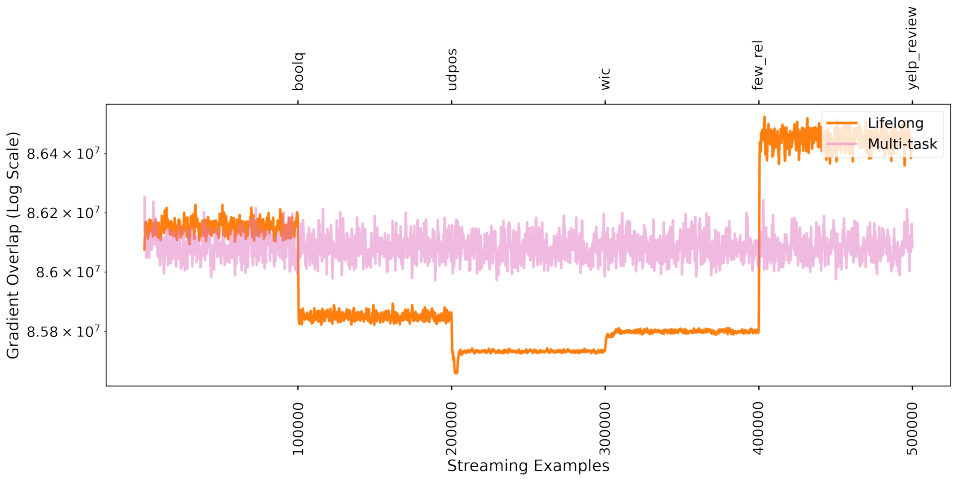

Figure 15: Gradient overlap on the Larger Stream.

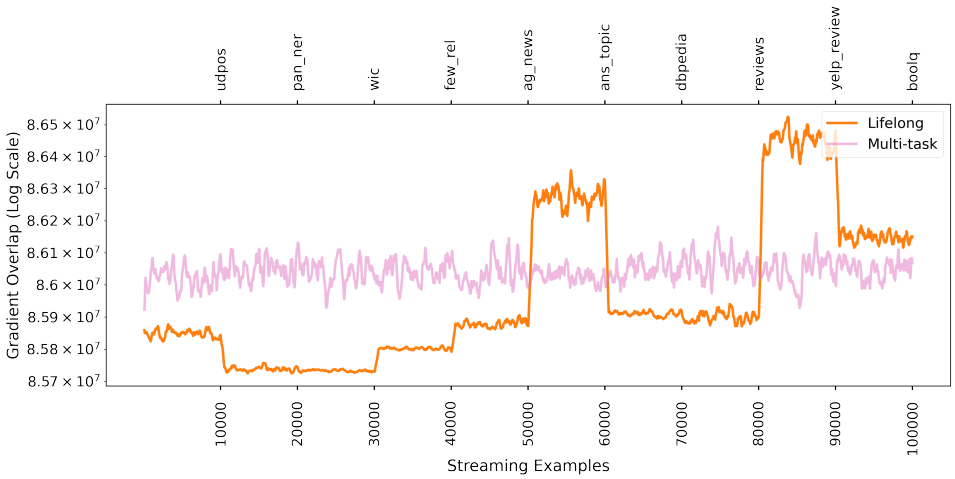

Figure 16: Gradient overlap in the Long Stream

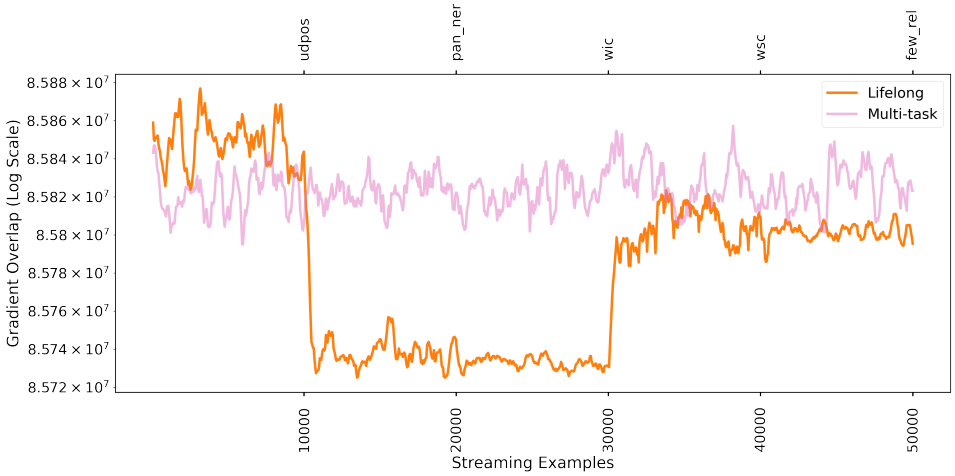

Figure 17: Gradient overlap on the Linguistic Stream.

Figure 18: Gradient overlap on the MultilingualA Stream.

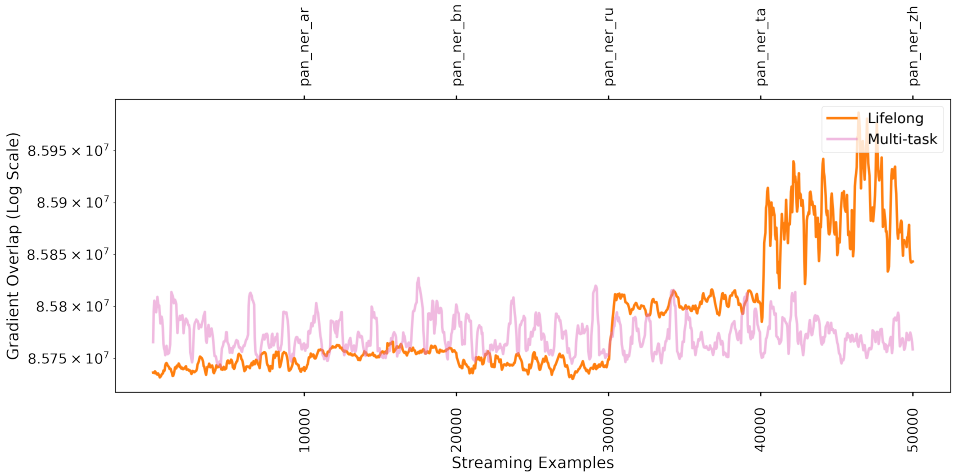

Figure 19: Gradient overlap on the MultilingualB Stream.

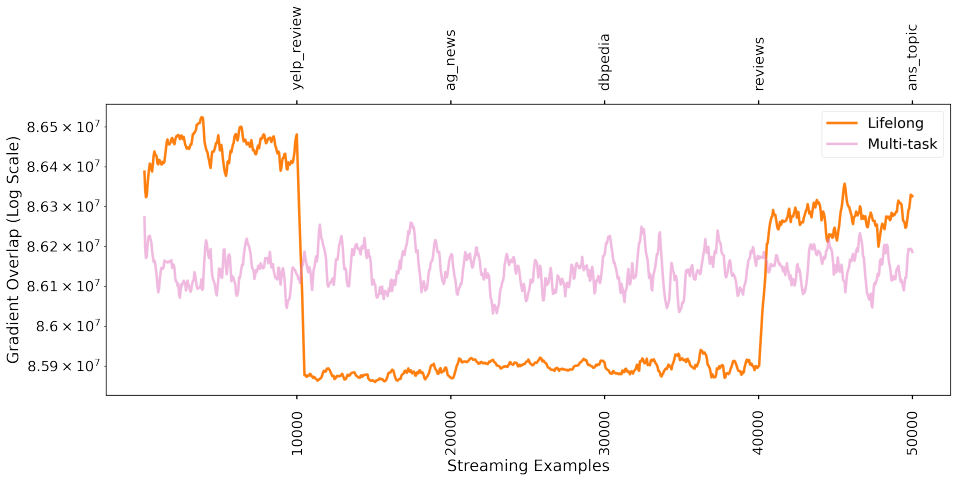

Figure 20: Gradient overlap on the MultidomainA Stream.

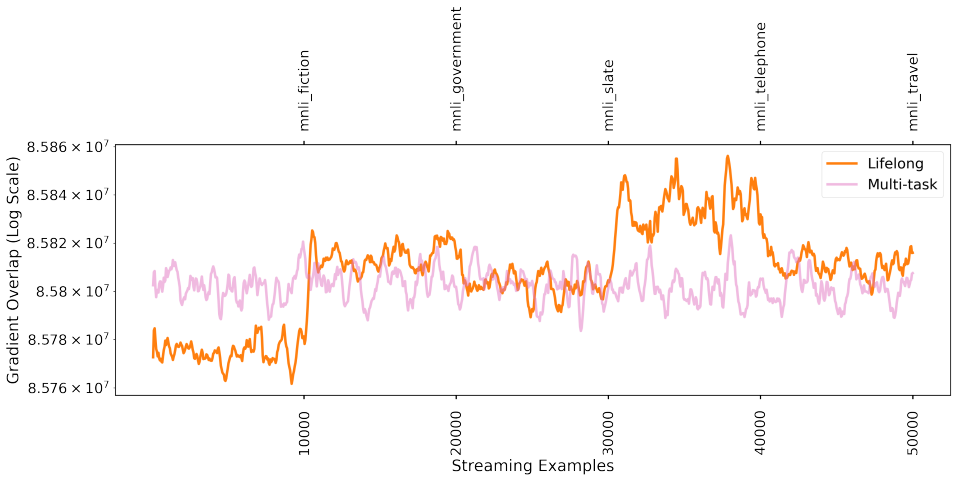

Figure 21: Gradient overlap on the MultidomainB Stream.

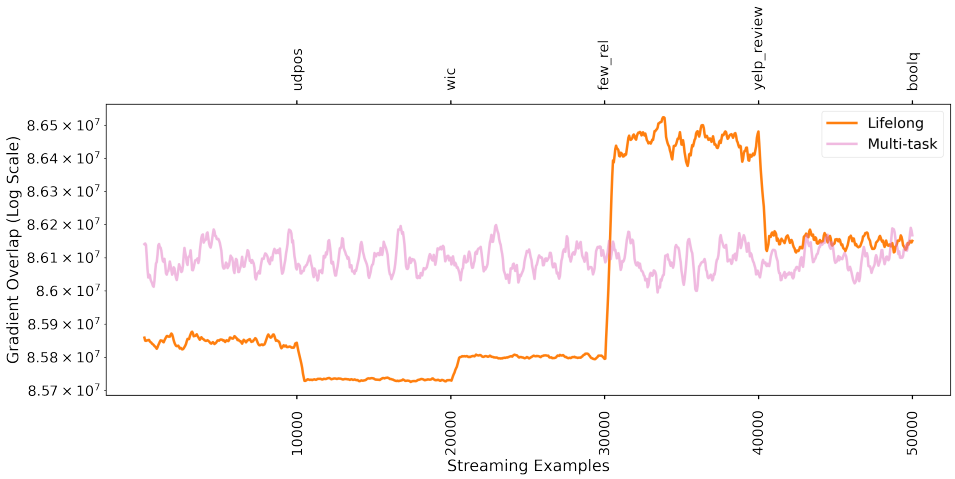

Figure 22: Gradient overlap on the AUC Stream.

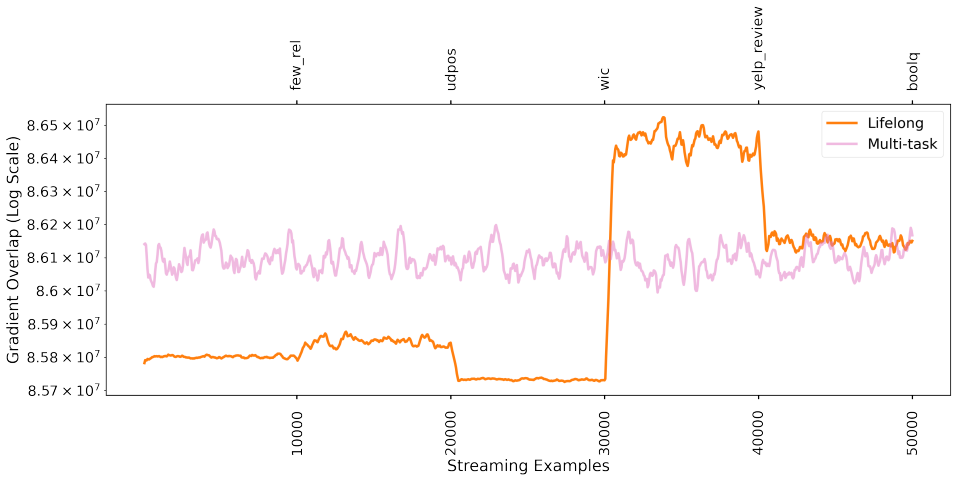

Figure 23: Gradient overlap on the Forgetting Stream.

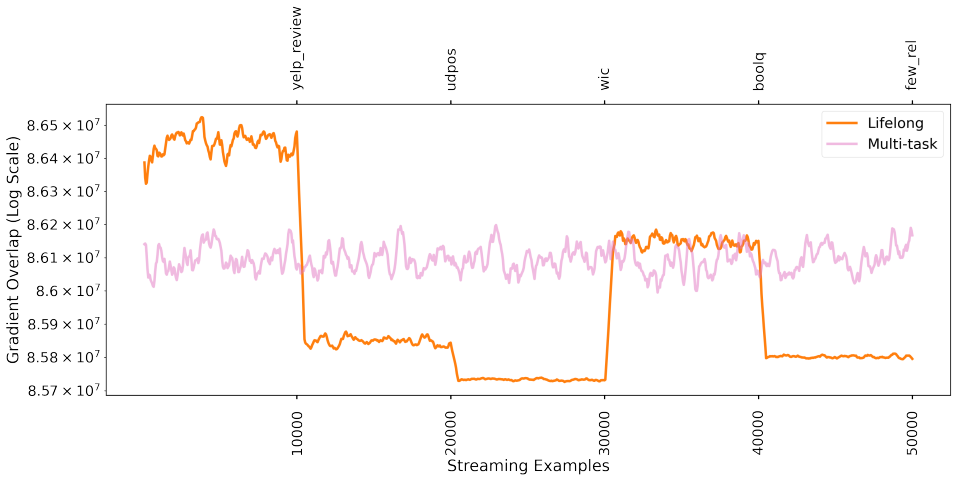

Figure 24: Gradient overlap on the Intransigence Stream.

Figure 25: Gradient overlap on the Final Accuracy Stream.

Table 11: List of explored datasets.

Dataset	Task
Universal Dependencies English-LinES (UDPOS) [61]	Part-of-speech tagging
WikiANN English (PANNER) [36]	Named entity recognition
Few-shot Relation Extraction (FewRel) [22]	Relation extraction
Word-in-Context (WiC) [38]	Word-in-Context Classification
Winograd Schema Challenge (WSC) [31]	Co-reference resolution
DBpedia [29]	Topic classification
AG News [64]	Topic classification
Yahoo Answers Topics [64]	Topic classification
Amazon Reviews [25]	Sentiment Analysis
Yelp Reviews [64]	Sentiment Analysis
Boolean Questions (BoolQ) [11]	Multi-choice question answering
Choice of Plausible Alternatives (COPA) [45]	Multi-choice question answering
Reading Comprehension with Commonsense Reasoning (ReCoRD) [63]	Extractive question answering
Multi-Sentence Reading Comprehension (MultiRC) [26]	Extractive question answering
Commitment Bank (CB) [13]	Natural language inference
Recognizing Tail Entailment (RTE) [12, 19, 20, 3]	Natural language inference

Table 12: List of selected datasets.

Dataset	Task	License
UDPOS	Part-of-speech tagging	CC BY-NC-SA 4.0
PANNER	Named entity recognition	None
FewRel	Relation extraction	MIT
WiC	Word-in-Context Classification	CC BY-NC-SA 4.0
AG News	Topic classification	Custom
DBpedia	Topic classification	CC BY-SA 3.0
Yahoo Answers Topics	Topic classification	Custom
Amazon Reviews	Sentiment Analysis	Custom
Yelp Reviews	Sentiment Analysis	Custom
BoolQ	Question answering	CC BY-SA 3.0

B Datasheet

Motivation We present an experimental framework along with a suite of benchmarks for lifelong learning using pre-trained language models. Not only is there a scarcity of lifelong learning benchmarks in the domain of NLP, but also none of the available benchmarks frame the lifelong learning problem in the most general form, i.e., having multiple tasks without explicit task identifiers. To this end, we propose the Degree-of-Belief framework which can incorporate multiple tasks without giving away explicit task identifiers. In this framework, the model states its belief in the truth of a statement given a context, and its past knowledge. Using this experimental setup, we design a suite of benchmark data streams consisting of multiple tasks, domains and languages that can be used to investigate, evaluate and experiment with lifelong learning models.

Composition To design the suite of benchmarks, we need a collection of datasets that can be used to form the data streams. We first investigated 16 datasets encompassing 10 different tasks, shown in Table 11, to find the ones that the model can learn reasonably well using 10k training examples. The selection criteria was imposed because of two reasons: i) to avoid conflating the challenge of learning the task with the challenge of lifelong learning itself; ii) to keep the experiment runtime on our benchmarks reasonably low. Based on the criteria, we end up with 10 datasets presented in Table 12. Not all of these 10 datasets have open licenses. Therefore, we develop and release a Lifelong Learning Library [4] to download, transform and organize these datasets into data streams based on our experimental framework for general lifelong learning. The library can also be used to extend the framework to new tasks and design custom lifelong data streams to facilitate additional experiments as needed. To guarantee availability of the datasets over time, we internally use the Datasets library from Hugging Face which provides reliable access to the largest archive of NLP datasets. We also link to the official homepage of each dataset in Table 12 for archival purposes.

[4] https://amanhussain.com/lifelong-learning/

Preprocessing To generate the different data streams, our library downloads the datasets and then transforms them into a format suitable for our experimental framework. The conversion steps for each of the datasets are described below:

1. **BoolQ**: The 'passage' is the context, the 'question' is the statement, and the 'label' is the truth label.
2. **UDPOS**: The context is the text to be tagged with the part-of-speech token labels. The statement consists of the sequence of correct part-of-speech tags. For each statement, we form three false statements by corrupting the part-of-speech tags randomly with a probability of 0.5.
3. **PANNER**: It follows the same transformation steps used in UDPOS, except for using named-entity tags instead of part-of-speech tags.
4. **WiC**: 'sentence1' and 'sentence2' are concatenated. The statement is constructed using the candidate 'word' w in one of these two templates randomly: 'w is the polysemous word' or 'w is used with the same sense'.
5. **FewRel**: The context is formed by the sentence(s) that feature(s) a head and a tail entity. The true statement is formed as: head entity – relation name – tail entity, where relation name is the correct relation label for these head and tail entities. For one true statement, we form three false statements by replacing the relation name with any of the incorrect relation labels randomly.
6. **Amazon Reviews**: The scores of 1 and 2 stars are converted to the 'negative' label. Similarly, the scores of 4 and 5 stars are converted to the 'positive' label, while the score of 3 stars is converted to the 'neutral' label. The context is formed by concatenating the 'review title' and 'review body'. The statement is formed by using one of these two templates randomly: 'It is a s review' or 'The sentiment is s', where s is one of these sentiment labels: ['negative', 'neutral', 'positive'].
7. **Yelp Reviews**: It follows the same transformation steps used in Amazon Reviews.
8. **AG News**: The context is the news article headline. The statement is formed by using either one of these templates randomly: 'The topic of the news headline is y' or 'The headline belongs to the y topic', where y is the correct topic.
9. **DBpedia**: It follows transformation steps similar to those used in AG News.
10. **Yahoo answers topics**: It follows transformation steps similar to those used in AG News.

For all the benchmark data streams provided by our library, the training set of each task consists of 10k examples, which is achieved via upsampling of smaller datasets and downsampling larger ones. Our library recommends the use of a continuous evaluation scheme (Area Under the Lifelong Test Curve) to measure test accuracy on all tasks throughout the lifelong learning process. Thus, the test set of each task is constrained to atmost 1k examples to keep the runtime of each experiment low and controlled.

Uses Our library is meant to be used for evaluating novel lifelong learning methods and/or investigating different properties of lifelong learning. It is designed such that it can be easily extended to adapt new tasks into our experimental framework and design new data streams for additional experiments. We will maintain a leaderboard of the proposed lifelong learning methods and data streams. To the best of our knowledge, the datasets do not contain any personally identifiable information or offensive content. However, we will maintain an erratum board to acknowledge and correct any biases, mistakes, etc. that might have accidentally been introduced.

Distribution and Maintenance Our lifelong learning library has been released on Github under the MIT license. We welcome contributions and feature requests from the research community. All future releases and updates will be distributed through the Github repository. For broader distribution, we will also release the suite of data streams directly through the Datasets library from Hugging Face.