MOTIVATION.

[question1] For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.
[answer1] 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
[question2] Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?
[answer2]
[question3] Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.
[answer3]
[question4] Any other comments?
[answer4] No.

COMPOSITION.

[question5] What do the instances that comprise the dataset represent (for example, documents, photos, people, countries)? Are there multiple types of instances (for example, movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.
[answer5] 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.
[question6] How many instances are there in total (of each type, if appropriate)?
[answer6]
[question7] Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (for example,

geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (for example, to cover a more diverse range of instances, because instances were withheld or unavailable).
[answer7]
[question8] What data does each instance consist of? "Raw" data (for example, unprocessed text or images) or features? In either case, please provide a description.
[answer8]
[question9] Is there a label or target associated with each instance? If so, please provide a description.
[answer9]
[question10] Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (for example, because it was unavailable). This does not include intentionally removed information, but might include, for example, redacted text.
[answer10]
[question11] Are relationships between individual instances made explicit (for example, users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.
[answer11]
[question12] Are there recommended data splits (for example, training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.
[answer12]
[question13] Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.
[answer13]
[question14] Is the dataset self-contained, or does it link to or otherwise rely on external resources (for example, websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (that is, including the external resources as they existed at the time the dataset was created); c) are there any restrictions (for example, licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.
[answer14]
[question15] Does the dataset contain data that might be considered confidential (for example, data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.

[answer15]
[question16] Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.
[answer16]
If the dataset does not relate to people, you may skip the remaining questions in this section.
[SkipA] NO
[question17] Does the dataset identify any sub-populations (for example, by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.
[answer17]
[question18] Is it possible to identify individuals (that is, one or more natural persons), either directly or indirectly (that is, in combination with other data) from the dataset? If so, please describe how.
[answer18]
[question19] Does the dataset contain data that might be considered sensitive in any way (for example, data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.
[answer19]
[question20] Any other comments?
[answer20] No.

COLLECTION PROCESS.

[question21] How was the data associated with each instance acquired? Was the data directly observable (for example, raw text, movie ratings), reported by subjects (for example, survey responses), or indirectly inferred/derived from other data (for example, part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.
[answer21] 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.

[question22] What mechanisms or procedures were used to collect the data (for example, hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?
[answer22]

[question23] If the dataset is a sample from a larger set, what was the sampling strategy (for example, deterministic, probabilistic with specific sampling probabilities)?

[answer23]

[question24] Who was involved in the data collection process (for example, students, crowdworkers, contractors) and how were they compensated (for example, how much were crowdworkers paid)?

[answer24]

[question25] Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (for example, recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.
[answer25]

[question26] Were any ethical review processes conducted (for example, by an institutional review board)? If so, please provide a description of these review

processes, including the outcomes, as well as a link or other access point to any supporting documentation.
[answer26]
If the dataset does not relate to people, you may skip the remaining questions in this section.
[SkipB] NO
[question27] Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (for example, websites)?
[answer27]
[question28] Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.
[answer28]
[question29] Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.
[answer29]
[question30] If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).
[answer30]
[question31] Has an analysis of the potential impact of the dataset and its use on data subjects (for example, a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.
[answer31]
[question32] Any other comments?
[answer32] No.

PREPROCESSING/CLEANING/LABELING.

[question33] Was any preprocessing/cleaning/labeling of the data done (for example, discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remaining questions in this section.
[answer33]

[question34] Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (for example, to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.
[answer34]
[question35] Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.
[answer35]
[question36] Any other comments?
[answer36] No.

USES.

[question37] Has the dataset been used for any tasks already? If so, please provide a description.
[answer37]
[question38] Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.
[answer38]
[question39] What (other) tasks could the dataset be used for?
[answer39] 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.
[question40] Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (for example, stereotyping, quality of service issues) or other risks or harms (for example, legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?
[answer40]
[question41] Are there tasks for which the dataset should not be used? If so, please provide a description.
[answer41]
[question42] Any other comments?
[answer42] No.

DISTRIBUTION.

[question43] Will the dataset be distributed to third parties outside of the entity (for example, company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.

[answer43]

[question44] How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?

[answer44] Our lifelong learning library has been released on Github. For broader distribution, we will also release the suite of data streams directly through the Datasets library from Hugging Face.

[question45] When will the dataset be distributed?

[answer45]

[question46] Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/ or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

[answer46] under the MIT license.

[question47] Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

[answer47]

[question48] Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

[answer48]

[question49] Any other comments?

[answer49] No.

MAINTENANCE.

[question50] Who will be supporting/hosting/maintaining the dataset?

[answer50]

[question51] How can the owner/curator/manager of the dataset be contacted (for example, email address)?

[answer51]

[question52] Is there an erratum? If so, please provide a link or other access point.

[answer52] 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.

[question53] Will the dataset be updated (for example, to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (for example, mailing list, GitHub)?

[answer53] All
future releases and updates will be distributed through the Github repository.

[question54] If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (for example, were the individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.

[answer54]

[question55] Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers.

[answer55] We welcome contributions and feature requests from the research community

[question56] If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description.

[answer56]

[question57] Any other comments?

[answer57] No.