Table 1: Comparison of the FLEX benchmark with closest prior work. Our benchmark consists of episodes with variable number of shots in the range [1-5] and with class imbalance. "No extra test data" refers to excluding validation data from testing tasks, to avoid unfairly advantaging models that use such data [50]. Our benchmark's number of test episodes is selected to balance statistical accuracy and precision, which suffers in few-episode setups, and compute requirements, which is too costly in many-episode setups (§5).

	CrossFit[75]	LM-BFF[24]	GPT-3[10]	DS[5]	SMLMT[4]	FewGlue[56]	**FLEX (ours)**
Class Transfer	-	-	-	✓	-	-	✓
Domain Transfer	-	-	-	-	✓	-	✓
Task Transfer	✓	-	-	-	✓	-	✓
Pretraining Transfer	-	✓	✓	-	-	✓	✓
Shots per class	{16, 32}	16	variable	{1,5}	{4,8,16,32}	{total 32}[4]	[1–5]
Variable shots	-	-	✓	-	-	-	✓
Unbalanced	-	-	-	-	-	-	✓
Textual labels	✓	✓	✓	-	-	✓	✓
Zero-shot	-	✓	✓	-	-	-	✓
No extra test data	-	-	-	✓	✓	mixed[5]	✓
# test episodes	5	5	1	1000	10	1	90
Reporting	avg	avg, SD	avg	avg, SD	avg, SD	avg, SD	all[6]
# datasets	160	16	37	7	18	8	20

be applied to NLP. However, unifying few-shot NLP work introduces new challenges. For example, the benchmark needs to test all types of transfer studied in separate research threads to measure progress on new techniques that make gains in each of these important generalization settings (§2). Also, given the importance of zero-shot learning and learning from task descriptions [29, 73], the benchmark needs to include zero-shot episodes and textual labels to enable measuring progress for models that do not use conventional supervised training, including methods that leverage the latent knowledge in pretrained LMs [10, 24, 78]. Further, the benchmark must accommodate new, computationally-expensive approaches, without overly reducing the number of evaluation episodes at the expense of statistical accuracy [3, 24, 75].

Need for a robust few-shot model Recent prompt-based models [10] have shown strong results in few-shot learning. These models leverage the power of (often large) pretrained language models and adapt the format of downstream tasks to the underlying pretraining objective (e.g., Masked Language Modeling). This way, given the right natural language prompt (and sometimes verbalizers [55] and additional demonstrative examples), the model can quickly fine-tune on the downstream task [24, 43, 44, 55, 56]. However, adapting task formats to the underlying (masked) language modeling objectives is not straightforward; such models have been shown to be sensitive to varying choices of the prompt/demonstrations, training settings, hyperparameters, and learning algorithms [33, 50, 78], often requiring large held out sets and/or complex methods to overcomes such challenges. Can models eschew complex prompt engineering by unifying pretraining and downstream task formats?

In this paper, we tackle these key issues by introducing FLEX—**F**ew-shot **L**anguage **E**valuation across (**X**) many transfer types—and contributing the following:

- FLEX Principles (§3), a set of requirements and best practices for few-shot NLP evaluation that enables unified, rigorous, valid, and cost-sensitive measurements.
  - Sample Size Design: In support of valid, cost-sensitive measurement, we introduce a novel approach to few-shot sample size design (§5) that optimizes for a benchmark's statistical accuracy and precision while keeping computational costs accessible to a broad range of researchers.
- FLEX benchmark (§4), an implementation of the FLEX Principles. It tests across *four* few-shot transfer settings,[7] and includes a public leaderboard for few-shot NLP that covers 20 datasets across diverse NLP tasks (e.g., NLI, relation classification, entity typing). Table 1 summarizes key differences between FLEX and other few-shot NLP evaluation suites.

[4]The total number of training shots in each episode, not number of shots per class per episode.

[5]Most users use unlabeled examples, though recently, Tam et al. [65] do not.

[6]Average (avg), confidence interval (CI), standard deviation (SD), individual episode metrics

[7]Prior work evaluated at most two settings.