Empowering Cross-lingual Behavioral Testing of NLP Models with Typological Features

Ester Hlavnova **Sebastian Ruder**
Google Research
{ehlavnova,ruder}@google.com

Abstract

A challenge towards developing NLP systems for the world's languages is understanding how they generalize to typological differences relevant for real-world applications. To this end, we propose M2C, a morphologically-aware framework for behavioral testing of NLP models. We use M2C to generate tests that probe models' behavior in light of specific linguistic features in 12 typologically diverse languages. We evaluate state-of-the-art language models on the generated tests. While models excel at most tests in English, we highlight generalization failures to specific typological characteristics such as temporal expressions in Swahili and compounding possessives in Finish. Our findings motivate the development of models that address these blind spots.[1]

1 Introduction

In natural language processing (NLP), there is a need to build systems that serve more of the world's approximately 6,900 languages. As one measure of linguistic diversity, the World Atlas of Language Structures (WALS; Haspelmath et al., 2005) records 192 linguistic features along which languages differ. These range from the order of subject, object, and verb (Dryer, 2013) to the number of basic color categories (Kay and Maffi, 2013). Languages present in existing NLP datasets mostly lie in low-density regions of the space of possible typological features (Ponti et al., 2021). In other words, many linguistic features that are common across the world's languages are not observed in languages that are the focus of NLP research.[2]

It is thus important to investigate to which linguistic features models can generalize and where they face challenges. However, existing datasets do not allow for a fine-grained cross-lingual evaluation and mainly permit comparisons on a language level (Hu et al., 2020). Prior studies focused on syntax and grammar through the lens of acceptability judgements (Ravfogel et al., 2018; Ahmad et al., 2019; Mueller et al., 2020; Papadimitriou et al., 2022). While these enable the evaluation of what a model deems 'natural' in a given language, it is often unclear how such biases relate to real-world applications of NLP technology.

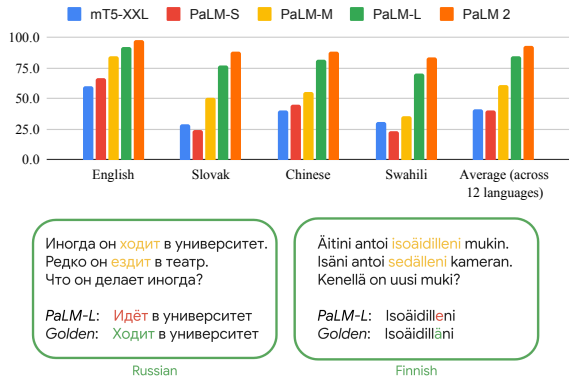

Figure 1: *Top*: Comparison of state-of-the-art models on M2C tests in a selected set of languages. Models perform well on English but poorly on certain tests in other languages. *Bottom*: Even the largest models fail on tests probing language-specific features, e.g., the distinction between habitual and one-time motion verbs in Russian (left) or possessives in Finnish (right); see Appendix B for English glosses and additional examples.

We propose Multilingual Morphological Checklist (M2C) to enable the investigation of a broader set of cross-lingual differences in practical scenarios. Specifically, we create a morphologically-aware behavioral testing framework (Ribeiro et al., 2020) that allows for the specification of tests in a diverse set of languages. Using this framework, we design tests that probe model's behavior in light of specific capabilities and typological features in 12 typologically diverse languages. We focus on a question answering setting as it represents one of

[1]We make all code publicly available at https://github.com/google-research/multi-morph-checklist.

[2]For instance, while tone is present in around 80% of African languages (Adebara and Abdul-Mageed, 2022), few Indo-European languages can be considered tonal.

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 7181–7198

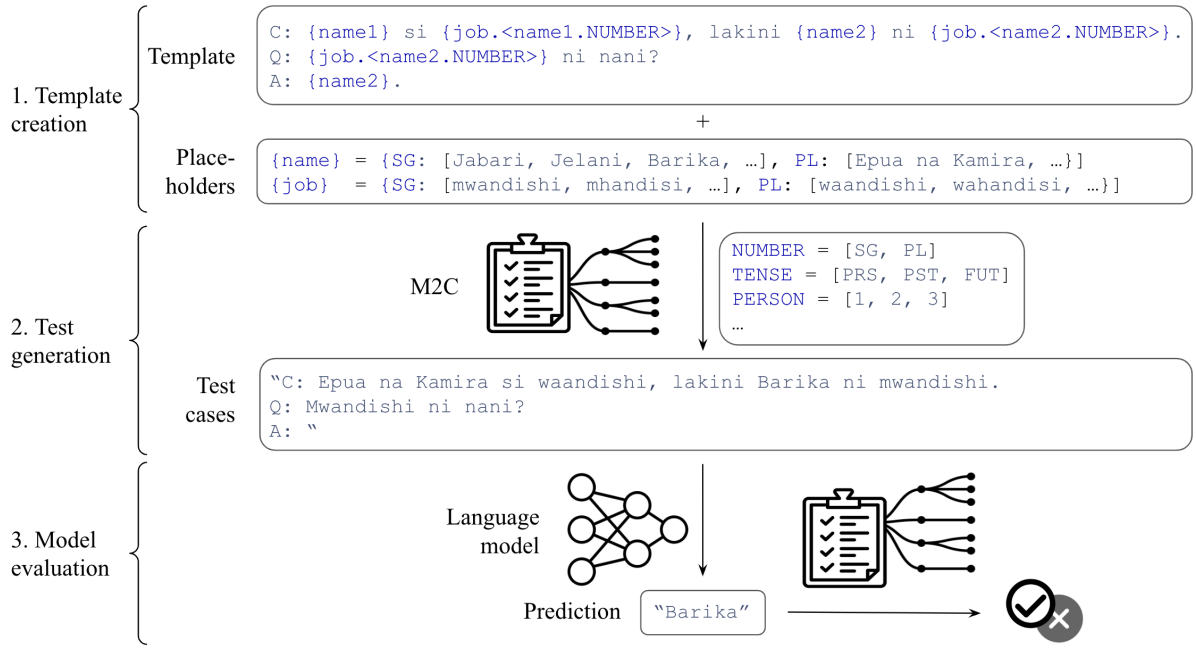

Figure 2: General workflow of using M2C for model evaluation. 1) Templates including context (C), question (Q), and answer (A) and placeholders for morphological features are created. 2) M2C is used to generate test cases. 3) A model is evaluated on the generated tests in a prompting setting and M2C is used to validate the predictions.

the most general and widely useful NLP applications (McCann et al., 2018) and enables zero-shot evaluation of models. We create tests that cover a diverse set of reasoning capabilities involving general linguistic features that are expressed differently across languages—negation, numerals, spatial and temporal expressions, and comparatives—as well as features unique to certain languages such as time in Swahili, measure words in Chinese, compounding possessives in Finnish, and motion verbs in Russian. We evaluate state-of-the-art language models on the generated tests in zero-shot and one-shot settings. Our findings shed light on generalization failures to specific typological features. For instance, all models struggle with time expressions in Swahili and measure words in Chinese. We show the workflow of using M2C, from template creation to model evaluation, in Figure 2.

Our contributions are: (1) We create a new morphologically-aware multilingual behavioral testing framework. (2) We highlight linguistic features that are challenging in different languages. (3) We design tests that probe model capabilities in light of practically relevant typological differences. (4) We evaluate state-of-the-art language models on the generated tests. (5) We shed light on the challenges posed by typological differences in multilingual scenarios.

2 Related Work

Perplexity Perplexity is a standard measure of evaluating language model performance, which has also been used in multilingual settings (Gerz et al., 2018). Besides being difficult to compare across segmentations, perplexity does not provide more fine-grained insights regarding model behavior (Meister and Cotterell, 2021). Acceptability evaluations compare perplexity between minimal pairs of grammatical and ungrammatical sentences (Linzen et al., 2016; Warstadt et al., 2020). Such evaluations have been extended to other languages (Ravfogel et al., 2018; Ahmad et al., 2019; Mueller et al., 2020; Xiang et al., 2021; Papadimitriou et al., 2022), which requires writing extensive language-specific grammars while the relevance of syntax biases in real-world applications remains unclear.

Evaluation of large models Most benchmarks designed for evaluating large models focus on assessing their performance on a collection of complex tasks (Wang et al., 2019; Hu et al., 2020; Hendrycks et al., 2021; Gehrmann et al., 2021; Srivastava et al., 2022). However, such benchmarks are unable to highlight more fine-grained model limitations (Ethayarajh and Jurafsky, 2020) and are outpaced by the development of new models.

Behavioral testing Behavioral testing sheds light on model capabilities via the design of simple targeted tasks. Early work such as bAbI (Weston et al., 2016) focused on toy tasks requiring simple reasoning capabilities while oLMpics (Talmor et al., 2020) consisted of 8 short classification tasks for masked language models. Recently, LMentry (Efrat et al., 2022) provides simple tests assessing fundamental generation capabilities. A common test bed is natural language inference (Naik et al., 2018; McCoy et al., 2019) where analyses of reasoning types have been extended to other languages (K et al., 2021; Joshi et al., 2020; Hartmann et al., 2021) but require existing data.

The CheckList framework (Ribeiro et al., 2020) enables the generation of behavioral tests for NLP models but its templates are English-centric. English Checklist tests have been extended to other languages via translation (Ruder et al., 2021; K et al., 2022). Such approaches, however, struggle with comprehensively covering linguistic features specific to a language and are not able to easily represent morphological variation. Relatedly, Jiang et al. (2020) create templates that integrate morphology for simple knowledge retrieval queries while Kassner et al. (2021) automatically translate knowledge retrieval queries into other languages. Compared to their approach, our framework allows for integrating morphology into a broader range of tests and is more scalable and flexible.

3 CheckList

CheckList (Ribeiro et al., 2020) relies on templates to generate a large amount of samples in order to evaluate models' behavior regarding different tasks and capabilities in a controlled manner. A template consists of a string with placeholders such as `{first_name}` delimited by curly brackets, *e.g.*, "`{first_name} is {adj}`". The user provides a set of values for each placeholder, for instance, `{first_name}` = {Michael, John, ... } and `{adj}` = {busy, friendly, ... }, which are used to populate the templates with their Cartesian product. The generated samples can then be applied to systematically test a model's performance in a specific setting.

Multilingual tests CheckList has been designed for English and provides mainly English-specific functionality. For example, it matches indefinite articles with nouns based on their starting letter, *i.e.*, the placeholder `{a:job}` generates "a lawyer" and "an engineer". As a consequence, CheckList is not capable of effectively generating tests in languages with richer morphology, which require maintaining agreement between multiple parts of the template—a feature that is beyond the scope of CheckList.

While multilingual tests can be generated by translating English tests (Ruder et al., 2021; K et al., 2022), optionally including template extraction and human verification, such generated templates struggle with handling rich morphology. In addition, in order to systematically probe linguistic features specific to a language, it is crucial to be able to efficiently generate in-language tests from scratch.

4 M2C Framework

We propose the M2C (Multilingual Morphological Checklist) framework in order to enable the generation of tests in a broad set of languages, including languages with rich morphology. A user provides a template as a string, a list of values for each placeholder, and an optional configuration dictionary in case of duplicate placeholders. The placeholder values can either be passed without inflections (for example, names in English) as a list of strings, or as a list of dictionaries with their corresponding inflected values. Each key of the dictionary is a feature combination (*e.g.*, `MASC.PL`) and the value is the corresponding string (*e.g.* "apples"). As such, each entity can have multiple inflections, for instance, in English "apple" and "apples". We show the general M2C workflow in Figure 2.

Morphological categories Our library follows the UniMorph Schema representation (Sylak-Glassman, 2016), which decomposes morphology into 23 dimensions and over 212 features. For example, Gender is one dimension, which contains features such as Feminine (FEM), Masculine (MASC), and Neuter (NEUT).

The ability to indicate these dimensions using a clear codification allows us to describe both the value attributes given to placeholders and their dependence on one another. As an example, in order to differentiate between "Juliette est grande" and "Julien est grand" in French, it is necessary to ensure gender agreement between noun and adjective by including the Gender attribute in the template. To cover such functionality, we introduce a syntax describing the morphological dependence between placeholders: `{X.<Y.D>}` signifies that `X` should have the same feature for dimension `D` as `Y`. In the above example, this is realized by "`{first_name} est {adj.<first_name.GENDER>}`".

Language-specific dimensions While initially relying on the UniMorph schema, we found cases where the existing dimensions are not sufficient to describe morphology of placeholders within the templates, which is especially necessary for dealing with exceptions. For instance, the trifold article distinction in Italian masculine gender—*il treno*, *l'hotel*, *lo studente*—depends on whether the noun starts with a consonant, vowel or *h*, or a specific consonant combination[3] respectively. In order to lexically encode such exceptions, we provide the ability to add dimensions, in this case STARTSWITH, which includes features VOW, CONS, and CONS2. While the goal of M2C is not to be exhaustive, it should enable encoding a sufficient number of dimensions to allow the user to write templates for diverse use cases.[4]

Advanced templating system To cover the variety of morphological phenomena, we designed a templating system with a rich syntax. When describing dependence rules, features can be added sequentially and are commutative, *e.g.*, `<first_name.GENDER.NUMBER>` is equivalent to `<first_name.NUMBER.GENDER>` where NUMBER = {singular, plural}. Often, only two or three output values are necessary, which directly depend on a placeholder's feature. We allow a simple expression to be passed directly in the template to make this rule explicit:

```
{val_1:placeholder.feature_1 | ... | val_n:placeholder.feature_n},
```

e.g., `{is:first_name.SG|are:first_name.PL}`, which produces "is" for a singular `{first_name}` and "are" for a plural one. Finally, we allow multiple placeholders with the same type, *e.g.*, `{first_name1}` and `{first_name2}`, to be populated by values of a common type, *i.e.*, `first_name`. In the case of multiple placeholders, we can provide a configuration for each placeholder type that specifies boolean repetition and order fields to, for instance, avoid having examples like "John and John" (repetition) or "John and Mary" and "Mary and John" (order).

Manual enumeration of features and their corresponding values is a barrier to scaling. To circumvent this, we integrate UnimorphInflect (Anastasopoulos and Neubig, 2019), which uses models trained on Unimorph data using the Unimorph Schema to generate inflections in 55 languages. As Unimorph models are imperfect—test accuracies range from 90%+ in many languages to 23% in Arabic—we envision a workflow where inflections are generated at scale using UnimorphInflect and then manually inspected by annotators for correctness. We expect the increase in productivity, and thus reduction in cost, to be significant by leveraging semi-automated as opposed to manual generation for languages with good performance.[5]

Answer validation Most prior benchmarks for behavioral testing of language models have focused on classification tasks (Talmor et al., 2020; Ribeiro et al., 2020). As M2C aims to support the evaluation of generative models using arbitrary templates, we implement functionality to match a range of outputs for each template, based on morphology, string matching and regex.[6]

Summary Overall, the M2C framework enables the systematic and controlled generation of high-quality tests at scale in a broad set of languages. As such, it occupies a middle ground between libraries such as SimpleNLG (Gatt and Reiter, 2009) that generate high-quality data but require encoding each language-specific rule, and template expansion via generative language models (Honovich et al., 2022), which are highly scalable but less reliable and underperform on languages with limited data (Hu et al., 2020). M2C enables modular design by allowing the addition of user-specified dimensions and features for specific templates and languages without requiring to encode all possible rules of a language. Furthermore, an advanced templating syntax and the semi-automatic generation of inflections may improve user productivity.

5 Capabilities and Typological Features

Languages We generate tests targeting capabilities and typological features in 12 typologically diverse languages: English (EN), Spanish (ES), Italian (IT), French (FR), German (DE), Swedish (SV), Finnish (FI), Slovak (SK), Russian (RU), Swahili (SW), Mandarin Chinese (ZH), and Arabic (AR).

Recent models have excelled at a wide range of tasks in English requiring a diverse set of reasoning

[3] *gn, pn, ps, x, y, z, s* followed by another consonant or *i* followed by a vowel.

[4] UniMorph defines a generic dimension 'Language Specific features' with attributes LGSPEC1, .., LGSPECN, which does not provide the clarity and flexibility of our setup.

[5] In order to ensure high-quality tests for the experiments in §6, we manually enumerate all relevant inflections.

[6] For each of the templates in §6, we curate possible outputs and implement regex and functions capturing them.

Test	Template	Generated test
Negation	C: {job2.NOM.<name2.NUMBER.GENDER>} {name2} و {job1.NOM.<name1.NUMBER.GENDER>} {name1}. Q: من {ليس:name2.SG|ليسا:name2.DU} {job1.NOM.<name2.NUMBER>.MASC}؟ A: {name2}.	C: أحمد مهندس وعمر كاتب. Q: من ليس مهندس؟ A: عمر.
Numerals	C: На столе {number1.<fruit1.GENDER>} {fruit1.NOM.<number1.NUMBER>} и {number2.<fruit2.GENDER>} {fruit2.NOM.<number2.NUMBER>}. {name} {съел:name.MASC|съела:name.FEM} {number3.<fruit1.GENDER>} {fruit1.<ACC:number3.SG|NOM>.<number3.NUMBER>}. Q: Сколько {fruit1.NOM.GTPL} на столе? A: {$diff(number1,number3)}.	C: На столе три ягоды клубники и пять ананасов. Анна съела две ягоды клубники. Q: Сколько ягод клубники на столе? A: Одна.
Spatial	C: {ART1.DEF.<obj1.NUMBER.STARTSWITH.GENDER>.TO_CAPITALIZE} {obj1} e {ART2.DEF.<obj2.NUMBER.STARTSWITH.GENDER>} {obj2} sono {prep.<place.STARTSWITH.GENDER>} {place}. {name} mette {ART2.DEF.<obj2.NUMBER.STARTSWITH.GENDER>} {obj2} sul pavimento. Q: {Dov'è:obj1.SG|Dove sono:obj1.PL} {ART3.DEF.<obj1.NUMBER.STARTSWITH.GENDER>} {obj1}? A: {prep.<place.STARTSWITH.GENDER>.TO_CAPITALIZE} {place}.	C: Il libro e le penne sono accanto al tavolo. Leonardo mette le penne sul pavimento. Q: Dov'è il libro? A: Accanto al tavolo.
Temporal	C: {name1} na {name2} ni {job1.PL} lakini {name1} atabadilisha kazi na atakuwa {job2.SG}. Q: {name1.TO_CAPITALIZE} atakuwa nani? A: {job2.SG.TO_CAPITALIZE}.	C: Jabari na Jelani ni waandishi lakini Jabari atabadilisha kazi na atakuwa mwalimu Q: Jabari atakuwa nani? A: Mwalimu.
Comparative	C: 如果{obj1}{comp1.GT}一点，{name}会{act}它。 如果{obj2}{comp2.GT}一点，{name}会{act}它。 Q: 如果它不那么{comp1.LT}，{name}会{act}什么? A: {obj1}	C: 如果公寓小一点，佳丽会买它。 如果电脑便宜一点，佳丽会买它。 Q: 如果它不那么大，佳丽会买什么? A: 公寓。

Table 1: Templates including context (C), question (Q), and answer (A) with generated test examples for linguistic features in Arabic, Russian, Italian, Swahili, and Mandarin Chinese. Placeholders are defined within curly brackets with their morphological dependence.

and understanding capabilities (Wang et al., 2019; Hendrycks et al., 2021). As most languages are morphologically richer than English, they encode the linguistic features representing such capabilities in more complex ways. The features we investigate are relevant in a variety of real-world applications including sentiment analysis (Wiegand et al., 2010), question answering (Dua et al., 2019), grounding (Kordjamshidi et al., 2020), reasoning with temporal change (Lazaridou et al., 2021) and quantitative attributes (Elazar et al., 2019).

We investigate capabilities and linguistic features present in all our investigated languages as well as linguistic features unique to certain languages. For each feature, we highlight differences in its cross-lingual instantiation and challenges for natural language understanding and generation. We create templates using the M2C framework to test a model's understanding of each capability and feature. We show a subset in Table 1.

5.1 Language-agnostic features

Negation In Indo-European languages, negation is often expressed via a separate particle such as *not* (English), *inte* (Swedish), etc. In contrast, in Swahili, for instance, negation morphemes are fused with the verb root and thus harder to identify. For other negation terms such as *kein* (German) models need to produce the correct agreement when generating text. In addition to gender and number agreement with the subject, Arabic negation takes up to five forms in singular, three forms in dual, and five forms in plural, *e.g.*, ليس (SG.MASC) and ليست (SG.FEM).

Numerals Models must be able to recognize and reason with numbers in their spelled-out and numerical forms across different writing and numeral systems, *e.g.*, *seventeen* (English) and 17 (Western Arabic numerals) and سبعة عشر and ١٧ (Eastern Arabic numerals). For generation in Russian and Slovak, models must inflect the noun depending on the quantity of the object. Slovak, for instance, has separate inflections for quantities of one, two/three-/four, and five and more, which also vary based on the object's animacy.

Spatial expressions In Russian, prepositions are associated with different cases, for example the instrumental case for за (*behind*) and the prepositional case for *on*. Such case agreement needs to be taken into account when generating text in Russian. Finnish, in addition to prepositions, follows a system of postpositions, which relate the location of one thing to another and require objects to be inflected in either partitive or genitive case.

Temporal expressions Some languages with rich morphology such as Finnish and Swahili encode temporal expressions in less complex ways than their inflection-sparser counterparts. In Swahili, verbal structure follows a simple compounding schema of subject marker + tense marker

	Prompt: Svara på frågan.
Spatial	Kontext: Pennan är under stolen och telefonen är på fönstret. Fråga: Var är telefonen? Svar: På fönstret **Kontext: Boken är under soffan och pennan är på hyllan.** **Fråga: Var är pennan?** **Svar:**

Table 2: Zero-shot and few-shot prompt example in Swedish spatial template. The zero-shot prompt only includes the information in bold while the one-shot prompt also includes the additional exemplar.

+ verb, e.g. *a-na-soma* (*he reads*) or *u-ta-soma* (*you will read*).

Comparatives Commonly, comparatives are expressed by a suffix or using a quantifier, *e.g.*, *more/less*. Spanish and French follow the latter approach by placing *más/menos* and *plus/moins* before the adjective with only a few standard exceptions. On the other hand, in Finnish, for example, the formation of comparatives follows a complex system of rules for compounding that includes categories depending on the endings of adjectives and a suffix *mpi*.

5.2 Language-specific features

Time in Swahili In many languages, the day is divided into two periods: a.m. and p.m., with the daily cycle starting at midnight (0:00) and running through noon (12:00). In Swahili, time is based on sunset and sunrise, defined to be 6 pm and 6 am respectively in standard time. For example, 11.30 am in standard time is 5.30 in the morning in Swahili time. Understanding different time systems is key not only for in-language reasoning but also for cross-lingual applications.

Possessives in Finnish Compounding in Finnish along with its system of 15 cases is one of the most challenging aspects of the language. One relevant feature are the possessive suffixes, which attach to the stem of nouns, *e.g.*, *koulu* (*school*) becomes *kouluni* (*my school*) and *koulumme* (*our school*). Possession is expressed via a suffix *-lla*, which compounds with other suffixes, *e.g.*, *siskollani* (*my sister has*), which must be correctly inflected by models in order to achieve the intended meaning.

Particles in Mandarin Chinese Another language specific-feature are measure words in Mandarin Chinese, which include over 150 cases and are used for different types of objects depending on their characteristics, *e.g.*, "本" for books, "双" for pairs, or "辆" for vehicles.

Motion verbs in Russian In most Slavic languages, motion verbs are a challenging concept as they behave differently than other verb categories. While most verbs have two forms (imperfective and perfective), motion verbs have three forms: one perfective form and two imperfective forms. Of the imperfective forms, the definite form indicates unidirectional or current one-time motion while the indefinite form represents multi-directional or habitual motion.

6 Experiments

Experimental setting We evaluate models on the generated tests in a question answering setting as can be seen in Figure 2. Each test consists of a context, a question, and an answer that needs to be predicted by the model. For each template, we generate 2,000 test examples on which the model is evaluated. A model's performance on a template is its accuracy of predicting a valid answer for a test averaged across all tests of the template.

We evaluate models in both zero-shot and one-shot settings for each capability and language. In the one-shot setting, a test randomly generated using the same template is used as the exemplar. This simplifies the task in two ways: i) it provides the model with a clear format for generating the answer and may enable the model to infer the answer's relationship to the rest of the template. While we conduct one-shot experiments to show the impact of additional instructions, zero-shot evaluation is the only setting that fully tests the model's understanding and generative capabilities independent of confounders such as the exemplar choice (Zhao et al., 2021), in line with prior work on behavioral testing (Ribeiro et al., 2020; Efrat et al., 2022). We provide an example of both settings in Table 2.

Models We evaluate five state-of-the-art pretrained language models of different sizes: an LM-adapted version (Vu et al., 2022) of mT5-XXL (13B parameters; Xue et al., 2021); PaLM-S (8B parameters), PaLM-M (62B parameters), and PaLM-L (540B parameters; Chowdhery et al., 2022); and PaLM 2 (Google et al., 2023). All models have

	EN	ES	IT	FR	DE	SV	FI	SK	RU	ZH	SW	AR	Avg.
mT5-XXL	59.6	32.0	43.9	41.4	50.4	39.3	44.8	28.5	39.1	40.0	30.6	52.1	41.8
PaLM-S	66.5	38.9	36.6	47.9	47.1	53.3	39.8	23.9	33.9	44.7	23.4	29.4	40.4
PaLM-M	84.5	70.9	60.1	78.2	71.8	66.2	53.5	50.6	54.0	55.1	35.1	48.8	60.7
PaLM-L	92.5	89.5	89.2	92.0	86.7	90.7	87.4	76.8	80.5	82.0	70.6	78.1	84.7
PaLM 2	**98.1**	**98.2**	**93.6**	**98.3**	**95.0**	**97.0**	**88.7**	**88.5**	**93.1**	**88.3**	**83.9**	**91.2**	**92.8**

Table 3: Average accuracy (in %) of different models on the generated tests in a zero-shot setting.

	Test type	Model	EN	ES	IT	FR	DE	SV	FI	SK	RU	ZH	SW	AR	Avg.
Negation	In context	mT5-XXL	80.7	72.8	85.5	80.2	63.1	55.8	84.4	31.8	45.3	30	33.7	43.1	56.9
		PaLM 2	**99.9**	**100**	**98.4**	**100**	**100**	**100**	**100**	**100**	**100**	**90.1**	**100**	**92.3**	**98.3**
	In question	mT5-XXL	19.1	30.1	23.4	25.1	36.1	20.6	19.7	16.7	9.6	5.2	3.7	58.2	22.6
		PaLM 2	**100**	**100**	**98.9**	**100**	**99.8**	**99.3**	**100**	**100**	**100**	**76.6**	**99.6**	**95.1**	**97.2**
Numerals	Addition	mT5-XXL	0.4	0.2	2.3	2	1.7	1.6	0	0	0	0	0.1	42.6	4.6
		PaLM 2	**96.1**	**100**	**68.7**	**99.7**	**96.5**	**100**	**100**	**99.9**	**96.9**	**66.5**	**94.5**	**79.3**	**91.1**
	Subtraction	T5-XXL	33.4	21.5	24.2	22.2	33	31.3	26.8	19.8	12.9	23	5.9	32.1	23.0
		PaLM 2	**95**	**92.4**	**90**	**93.6**	**93.6**	**89.1**	**87.5**	**88.4**	**93.6**	**81.2**	**68.7**	**87.4**	**87.8**
Spatial	Prepositions	mT5-XXL	98.8	28	51.4	40.2	78.3	59.6	27.6	51.3	49.5	99.9	52.8	74.4	55.7
		PaLM 2	**100**	**100**	**94.8**	**100**	**100**	**100**	**100**	**100**	**99.9**	**100**	**100**	**98.7**	**99.4**
	Position	mT5-XXL	90.9	15	74.5	61.1	95.2	35.1	60.3	29	50	**100**	49	65.3	57.7
		PaLM 2	**100**	**100**	**99.9**	**100**	**100**	**100**	**99**	**100**	**99.9**	**100**	**46.7**	**91.0**	**94.2**
Temporal	Past	mT5-XXL	86.3	27.8	44.4	62.1	50.4	77.5	78.7	61.7	93.1	81.1	35.2	68.9	61.9
		PaLM 2	**99.3**	**100**	**89.8**	**100**	**86.8**	**100**	**100**	**83.5**	**96.9**	**96.7**	**62.9**	**96.2**	**92.1**
	Future	mT5-XXL	85.7	79.8	48.4	56.9	55.3	55	62.2	38.3	93.5	52.7	39	58.7	58.2
		PaLM 2	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**95**	**99.1**	**100**	**100**	**99.8**	**99.4**
Comparative	Standard	mT5-XXL	58.1	44	37.3	48.7	45.3	28.3	60	31.3	17.3	7.7	51.7	45.3	37.9
		PaLM 2	**100**	**97.7**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**99.3**	**100**	**100.0**	**99.7**
	Conditional	mT5-XXL	42.4	1.1	47.8	15.8	45.5	28.1	**28.7**	4.7	19.8	0	35.2	32.1	23.5
		PaLM 2	**90.6**	**92.1**	**95.1**	**89.4**	**73.3**	**81.7**	0	**18.1**	**44.2**	**72.7**	**66.3**	**72.1**	**64.1**

Table 4: Accuracy (in %) of mT5-XXL and PaLM 2 on the generated tests in a zero-shot setting.

been trained on large amounts of web text but have not been otherwise fine-tuned for instruction-following or few-shot learning.

Generation Predictions are generated using greedy decoding with a temperature of 0 and a maximum of 20 decoding steps.

7 Results

7.1 Performance across Languages

We show the average results across tests covering language-agnostic features across languages and models in Table 3. We present the detailed results across test types for mT5-XXL and PaLM 2 in Table 4 and for PaLM-S, PaLM-M, and PaLM-L in Appendix A. We show results on language-specific features for all models in Table 5.

M2C tests are challenging, particularly for smaller models and for certain languages. mT5-XXL and PaLM-S achieve comparatively poor performance on average across languages. While performance is highest for English, across the other languages both models only pass at most 50% of tests—and less than a third for Slovak (SK), Swahili (SW), and Arabic (AR) for PaLM-S. These results highlight that the tests generated with M2C are challenging for the majority of state-of-the-art models and demonstrate that a clear gap between performance on English and performance in other languages remains for most models.

Competence with language-agnostic features emerges at scale. We observe a 20 point improvement in average performance from PaLM-S to PaLM-M to PaLM-L, highlighting that model robustness to linguistic features improves with scale. The strongest model, PaLM 2, reaches almost perfect performance on English and on the Indo-European languages. Compared to PaLM-L,

PaLM 2 achieves the largest improvements on Slovak, Russian, Swahili, and Arabic. On Finnish, Slovak, Chinese, and Swahili average performance of PaLm 2 is still below 90%, however, indicating that there is headroom left in terms of competence with regard to language-agnostic features for even the strongest current models.

7.2 Performance across Linguistic Features

Language-agnostic features The most challenging test types for mT5-XXL and PaLM 2 in Table 4 are numerals and comparatives. mT5 performs poorly on addition and only slightly better on subtraction while PaLM 2 achieves around 90% performance on most languages. On comparatives, both models have more difficulty in the conditional case. While PaLM 2 passes negation tests with almost perfect accuracy across different languages, mT5 displays reduced performance, particularly when the question is negated and for non-Indo-European languages. This highlights that robust reasoning with negation only emerges at scale. On spatial and temporal tests, mT5 achieves reasonable performance in most languages, while PaLM 2 achieves perfect performance in most cases and only underperforms in Swahili.

Language-specific features We show the results on the language-specific feature tests in Table 5. All models have acquired a reasonable ability to distinguish between different forms of motion verbs in Russian. Small and medium-sized models generally fail to reason with compounding possessives in Finnish and time expressions in Swahili while all models are unable to perfectly employ the correct measure words in Chinese, despite it being a high-resource language. Similarly, even PaLM 2 is unable to correctly reason with time expressions in Swahili. We show examples of errors in model predictions for each test type together with English glosses in Appendix B.

7.3 Evaluating Morphological Correctness

The generated tests focus on evaluating a model's understanding capabilities with regard to specific capabilities and linguistic features. As the linguistic features are often expressed via morphology, we additionally calculate the fraction of errors due to morphology in the models' output for the tests with morphological variation in the answer. This enables us to assess a model's ability to generate morphologically correct forms. For instance, in Slovak, a model must generate the correct accents and suffixes, *e.g.*, it is an error if the model predicts the *Trináste* (13*th*) instead of *Trinást'* (13). We automatically identify and manually curate these errors for PaLM-L and report the proportion of morphology-related errors for a subset of tests and languages in Table 6. We show examples of errors in model predictions that are due to morphology in Appendix C.

For certain tests with morphological variation in the answer, a non-negligible fraction of errors are due to producing incorrect morphological forms. For negation in Slovak, around half of PaLM-L's errors are due to morphology such as an incorrect use of diacritics or suffixes, highlighting a weakness of subword-based models. For numerical reasoning, models frequently produce incorrectly inflected numerals. Similarly, models generate outputs with an incorrect case or number for tests related to spatial and temporal expressions and comparatives.

	FI	RU	ZH	SW	Avg.
mT5-XXL	1.2	62.6	38.8	0	25.7
PaLM-S	3.6	68.1	5.1	0	19.2
PaLM-M	12.4	86.9	61.4	0	40.2
PaLM-L	63.4	90	71.6	13.6	59.7
PaLM 2	98.7	99.4	77.5	69	86.2

Table 5: Accuracy (in %) on tests testing language-specific features: time (Swahili), possessives (Finish), particles (Chinese), motion verbs (Russian).

	Languages	FI	SK	RU
Negation	In context	31.6	45.7	27.6
	In question	10	51.8	3.2
Numerals	Addition	8	16.2	4.2
	Subtraction	12.4	30	11.8
Spatial	Prepositions	7.8	8.2	0
	Position	0	0	0.1
Temporal	Past	0	21.8	39.8
	Future	0	8.3	0
Comparative	Standard	0	0	0
	Conditional	4.5	3.2	25.6

Table 6: Percentage of morphological errors (in %) by PaLM-L on the generated tests with zero-shot setting. Example erroneous predictions corresponding to highlighted cells are in Appendix C.

7.4 One-shot Evaluation

We show one-shot results for all models in Appendix D. The one-shot setting generally improves results as it allows the model to infer the format of the answer and potentially its relationship to the rest of the template. Improvements are larger for smaller models, which benefit more from information about the template. Nevertheless, even in this setting models are unable to achieve perfect accuracy across all languages. Reasoning with numerals and comparatives are still challenging for most models while improvements on numerals are also relatively smaller than on other test types. Models struggle particularly in Swahili across different test types. Overall, these results demonstrate that even in one-shot settings, large language models are not able to systematically generalize to certain typological features in multilingual settings.

8 Conclusion

In this paper, we have introduced M2C, a multilingual morphological framework for targeted behavioral evaluation of language-specific capabilities. As world languages present different challenges, M2C aims to provide flexibility in defining a suitable templating system with its individual dimensions and features. We have conducted experiments on state-of-the-art large language models, highlighted typological features that models struggle with, and quantified errors occurring due to morphology. We hope M2C inspires further research focused on tackling typological and morphological challenges with large language models.

Acknowledgements

We thank Jialu Liu, Jiaming Shen, and Jonas Pfeiffer for helpful feedback on a draft of this paper.

Broader Impact Statement

Accessibility Our new behavioral testing framework enables the generation of tests that incorporate morphology, which makes the systematic and fine-grained evaluation of NLP models more accessible across a diverse set of languages. For many such languages, it was previously not feasible to gain a fine-grained understanding of a model's capabilities.

Risks Risks are limited and mainly relate to obtaining a biased view of a capability due to the use of limited templates.

Limitations The creation of templates still requires native speaker expertise and an understanding of a language's grammar. Morphological inflection models are imperfect so morphological forms may need to be enumerated to ensure high-quality tests. We leave model-in-the-loop template creation and improving morphological inflection models for future work. While we design representative templates with thousands of permutations for each capability, a larger set of templates and arguments may be necessary to ensure a comprehensive coverage.

References

Ife Adebara and Muhammad Abdul-Mageed. 2022. Towards afrocentric NLP for African languages: Where we are and where we can go. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3814–3841, Dublin, Ireland. Association for Computational Linguistics.

Wasi Ahmad, Zhisong Zhang, Xuezhe Ma, Eduard Hovy, Kai-Wei Chang, and Nanyun Peng. 2019. On difficulties of cross-lingual transfer with order differences: A case study on dependency parsing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2440–2452, Minneapolis, Minnesota. Association for Computational Linguistics.

Antonios Anastasopoulos and Graham Neubig. 2019. Pushing the limits of low-resource morphological inflection.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Matthew S. Dryer. 2013. Order of subject, object and verb. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2368–2378, Minneapolis, Minnesota. Association for Computational Linguistics.

Avia Efrat, Or Honovich, and Omer Levy. 2022. LMentry: A Language Model Benchmark of Elementary Language Tasks. *arXiv preprint:2211.02069*.

Yanai Elazar, Abhijit Mahabal, Deepak Ramachandran, Tania Bedrax-Weiss, and Dan Roth. 2019. How large are lions? inducing distributions over quantitative attributes. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3973–3983, Florence, Italy. Association for Computational Linguistics.

Kawin Ethayarajh and Dan Jurafsky. 2020. Utility is in the eye of the user: A critique of NLP leaderboards. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4846–4853, Online. Association for Computational Linguistics.

Albert Gatt and Ehud Reiter. 2009. Simplenlg: A realisation engine for practical applications. *Proceedings of the 12th European Workshop on Natural Language Generation, ENLG 2009*, pages 90–93.

Sebastian Gehrmann, Tosin Adewumi, Karmanya Aggarwal, Pawan Sasanka Ammanamanchi, Anuoluwapo Aremu, Antoine Bosselut, Khyathi Raghavi Chandu, Miruna-Adriana Clinciu, Dipanjan Das, Kaustubh Dhole, Wanyu Du, Esin Durmus, Ondřej Dušek, Chris Chinenye Emezue, Varun Gangal, Cristina Garbacea, Tatsunori Hashimoto, Yufang Hou, Yacine Jernite, Harsh Jhamtani, Yangfeng Ji, Shailza Jolly, Mihir Kale, Dhruv Kumar, Faisal Ladhak, Aman Madaan, Mounica Maddela, Khyati Mahajan, Saad Mahamood, Bodhisattwa Prasad Majumder, Pedro Henrique Martins, Angelina McMillan-Major, Simon Mille, Emiel van Miltenburg, Moin Nadeem, Shashi Narayan, Vitaly Nikolaev, Andre Niyongabo Rubungo, Salomey Osei, Ankur Parikh, Laura Perez-Beltrachini, Niranjan Ramesh Rao, Vikas Raunak, Juan Diego Rodriguez, Sashank Santhanam, João Sedoc, Thibault Sellam, Samira Shaikh, Anastasia Shimorina, Marco Antonio Sobrevilla Cabezudo, Hendrik Strobelt, Nishant Subramani, Wei Xu, Diyi Yang, Akhila Yerukola, and Jiawei Zhou. 2021. The GEM benchmark: Natural language generation, its evaluation and metrics. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 96–120, Online. Association for Computational Linguistics.

Daniela Gerz, Ivan Vulić, Edoardo Ponti, Jason Naradowsky, Roi Reichart, and Anna Korhonen. 2018. Language modeling for morphologically rich languages: Character-aware modeling for word-level prediction. *Transactions of the Association for Computational Linguistics*, 6:451–465.

Rohan Anil Google, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernandez Abrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vlad Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, Guy Gur-Ari, Steven Hand, Hadi Hashemi, Le Hou, Joshua Howland, Andrea Hu, Jeffrey Hui, Jeremy Hurwitz, Michael Isard, Abe Ittycheriah, Matthew Jagielski, Wenhao Jia, Kathleen Kenealy, Maxim Krikun, Sneha Kudugunta, Chang Lan, Katherine Lee, Benjamin Lee, Eric Li, Music Li, Wei Li, YaGuang Li, Jian Li, Hyeontaek Lim, Hanzhao Lin, Zhongtao Liu, Frederick Liu, Marcello Maggioni, Aroma Mahendru, Joshua Maynez, Vedant Misra, Maysam Moussalem, Zachary Nado, John Nham, Eric Ni, Andrew Nystrom, Alicia Parrish, Marie Pellat, Martin Polacek, Alex Polozov, Reiner Pope, Siyuan Qiao, Emily Reif, Bryan Richter, Parker Riley, Alex Castro Ros, Aurko Roy, Brennan Saeta, Rajkumar Samuel, Renee Shelby, Ambrose Slone, Daniel Smilkov, David R. So, Daniel Sohn, Simon Tokumine, Dasha Valter, Vijay Vasudevan, Kiran Vodrahalli, Xuezhi Wang, Pidong Wang, Zirui Wang, Tao Wang, John Wieting, Yuhuai Wu, Kelvin Xu, Yunhan Xu, Linting Xue, Pengcheng Yin, Jiahui Yu, Qiao Zhang, Steven Zheng, Ce Zheng, Weikang Zhou, Denny Zhou, Slav Petrov, and Yonghui Wu. 2023. Palm 2 technical report.

Mareike Hartmann, Miryam de Lhoneux, Daniel Hershcovich, Yova Kementchedjhieva, Lukas Nielsen, Chen Qiu, and Anders Søgaard. 2021. A multilingual benchmark for probing negation-awareness with minimal pairs. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 244–257, Online. Association for Computational Linguistics.

Martin Haspelmath, Matthew S Dryer, David Gil, and Bernard Comrie. 2005. *The world atlas of language structures*. OUP Oxford.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring Massive Multitask Language Understanding. In *Proceedings of ICLR 2021*.

Or Honovich, Thomas Scialom, Omer Levy, and Timo Schick. 2022. Unnatural Instructions: Tuning Language Models with (Almost) No Human Labor. *arXiv preprint arXiv:2212.09689*.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson.

2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *International Conference on Machine Learning*, pages 4411–4421. PMLR.

Zhengbao Jiang, Antonios Anastasopoulos, Jun Araki, Haibo Ding, and Graham Neubig. 2020. X-FACTR: Multilingual factual knowledge retrieval from pretrained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5943–5959, Online. Association for Computational Linguistics.

Pratik Joshi, Somak Aditya, Aalok Sathe, and Monojit Choudhury. 2020. TaxiNLI: Taking a ride up the NLU hill. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 41–55, Online. Association for Computational Linguistics.

Karthikeyan K, Shaily Bhatt, Pankaj Singh, Somak Aditya, Sandipan Dandapat, Sunayana Sitaram, and Monojit Choudhury. 2022. Multilingual CheckList: Generation and evaluation. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 282–295, Online only. Association for Computational Linguistics.

Karthikeyan K, Aalok Sathe, Somak Aditya, and Monojit Choudhury. 2021. Analyzing the effects of reasoning types on cross-lingual transfer performance. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 86–95, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Nora Kassner, Philipp Dufter, and Hinrich Schütze. 2021. Multilingual LAMA: Investigating knowledge in multilingual pretrained language models. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3250–3258, Online. Association for Computational Linguistics.

Paul Kay and Luisa Maffi. 2013. Number of basic colour categories. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Parisa Kordjamshidi, James Pustejovsky, and Marie-Francine Moens. 2020. Representation, learning and reasoning on spatial language for downstream NLP tasks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts*, pages 28–33, Online. Association for Computational Linguistics.

Angeliki Lazaridou, Adhiguna Kuncoro, and Elena Gribovskaya. 2021. Mind the Gap : Assessing Temporal Generalization in Neural Language Models. In *Proceedings of NeurIPS 2021*.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. 2018. The Natural Language Decathlon : Multitask Learning as Question Answering. *arXiv preprint arXiv:1806.08730*.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Clara Meister and Ryan Cotterell. 2021. Language model evaluation beyond perplexity. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5328–5339, Online. Association for Computational Linguistics.

Aaron Mueller, Garrett Nicolai, Panayiota Petrou-Zeniou, Natalia Talmina, and Tal Linzen. 2020. Cross-linguistic syntactic evaluation of word prediction models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5523–5539, Online. Association for Computational Linguistics.

Aakanksha Naik, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, and Graham Neubig. 2018. Stress test evaluation for natural language inference. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2340–2353, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Isabel Papadimitriou, Kezia Lopez, and Dan Jurafsky. 2022. Multilingual BERT has an accent: Evaluating English influences on fluency in multilingual models. *arXiv preprint arXiv:2210.05619*.

Edoardo Maria Ponti, Rahul Aralikatte, Disha Shrivastava, Siva Reddy, and Anders Søgaard. 2021. Minimax and neyman–Pearson meta-learning for outlier languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1245–1260, Online. Association for Computational Linguistics.

Shauli Ravfogel, Yoav Goldberg, and Francis Tyers. 2018. Can LSTM learn to capture agreement? the case of Basque. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 98–107, Brussels, Belgium. Association for Computational Linguistics.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 4902–4912, Online. Association for Computational Linguistics.

Sebastian Ruder, Noah Constant, Jan Botha, Aditya Siddhant, Orhan Firat, Jinlan Fu, Pengfei Liu, Junjie Hu, Dan Garrette, Graham Neubig, and Melvin Johnson. 2021. XTREME-R: Towards more challenging and nuanced multilingual evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10215–10245, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

John Sylak-Glassman. 2016. The Composition and Use of the Universal Morphological Feature Schema (UniMorph Schema).

Alon Talmor, Yanai Elazar, Yoav Goldberg, and Jonathan Berant. 2020. oLMpics-on what language model pre-training captures. *Transactions of the Association for Computational Linguistics*, 8:743–758.

Tu Vu, Aditya Barua, Brian Lester, Daniel Cer, Mohit Iyyer, and Noah Constant. 2022. Overcoming catastrophic forgetting in zero-shot cross-lingual generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9279–9300, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Alex Wang, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2019. SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. In *Proceedings of NeurIPS 2019*.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Jason Weston, Antoine Bordes, Sumit Chopra, Alexander M Rush, Bart Van Merriënboer, Armand Joulin, and Tomas Mikolov. 2016. Towards ai-complete question answering: A set of prerequisite toy tasks. In *Proceedings of ICLR 2016*.

Michael Wiegand, Alexandra Balahur, Benjamin Roth, Dietrich Klakow, and Andrés Montoyo. 2010. A survey on the role of negation in sentiment analysis. In *Proceedings of the Workshop on Negation and Speculation in Natural Language Processing*, pages 60–68, Uppsala, Sweden. University of Antwerp.

Beilei Xiang, Changbing Yang, Yu Li, Alex Warstadt, and Katharina Kann. 2021. CLiMP: A benchmark for Chinese language model evaluation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2784–2790, Online. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *International Conference on Machine Learning*, pages 12697–12706. PMLR.

A Zero-shot Results

We show zero-shot results for PaLM-S, PaLM-M, and PaLM-L across different tests and languages in Table 7.

B Examples of Errors on Language-specific Feature Tests

We show examples of errors on language-specific feature tests with PaLM-L together with English glosses in Table 8.

C Examples of Morphological Errors

We show example errors in predictions of PaLM-L that are due to morphology in Table 9.

D One-shot Results

We show one-shot results for all models in Table 10. We show summary statistics of the average relative change in performance of the one-shot setting compared to the zero-shot setting for each language and model in Table 11.

	Test type	Model	EN	ES	IT	FR	DE	SV	FI	SK	RU	ZH	SW	AR	Avg.
Negation	In context	PaLM-S	47.6	31.6	33.7	40.6	42.2	31.1	29.8	29.2	20.1	47.7	37.6	30.3	35.1
		PaLM-M	97.6	61.3	75.1	89.1	71	89.2	71.4	83.5	44.6	50.4	45.4	40.1	68.2
		PaLM-L	99.9	99	99.3	99.2	99.7	100	99.7	95.2	90.5	90.2	97.6	91.2	96.8
	In question	PaLM-S	56.5	30.4	44.7	29.2	37.2	38.8	39.9	30.9	46	49.8	9.4	26.9	36.6
		PaLM-M	85.6	43.1	58.2	61.8	57.9	60.9	36.3	51.9	45.3	34.4	62.8	36.4	52.9
		PaLM-L	99.7	90.1	94.2	95.2	99.6	99.3	99.3	90	92.6	65	73.3	74.2	89.4
Numerals	Addition	PaLM-S	66.7	43.5	33.5	36.9	43.3	48.3	3.5	1.9	1.5	1	0	4.8	23.7
		PaLM-M	77.2	55.8	17.2	64.3	56.8	10.4	15.9	9.3	22.7	36.1	2.8	12.5	31.8
		PaLM-L	96.5	92.6	94.1	97.9	98.2	87	61.7	58.8	74.6	82.5	47.3	59.7	79.2
	Subtraction	PaLM-S	47.8	8.3	18.1	27.1	32	4	17.4	4	4.8	20	0	11.4	16.2
		PaLM-M	44.5	38.2	36.4	65.4	46.2	35.4	27.5	17.6	34.1	59	0.1	31.9	36.4
		PaLM-L	93.6	92.4	77.9	93	62.2	86	87.9	77.2	90.4	86.6	33	53.1	77.8
Spatial	Prepositions	PaLM-S	84.9	73.6	52	75.9	86.2	88.6	41.5	31.5	30.1	87.4	14.1	31.9	58.1
		PaLM-M	99	99	50	94.5	93.3	77	67.1	72.2	87.9	96.2	88.7	65.3	82.5
		PaLM-L	99.9	95.8	96.9	99.7	100	99.5	93.3	88.2	99	100	99.4	96	97.3
	Position	PaLM-S	54.6	49.4	38.6	44.8	59	58.5	44.9	14.9	11.8	49.2	20.9	29.1	39.6
		PaLM-M	65.9	73.3	42	56	62.5	57	46.7	31.4	36.7	62.4	18.9	42.8	49.6
		PaLM-L	61.1	58.7	74.6	55.1	73.7	74.6	88.1	66.9	41	89.2	23.7	74.7	65.1
Temporal	Past	PaLM-S	81.9	14.1	21.5	36.5	31.1	95	74.5	26	90.5	80.5	38.4	41.2	52.6
		PaLM-M	99.6	94.5	92.9	95.8	94.4	92.1	98	93.8	94.7	96.3	40	67.8	88.3
		PaLM-L	100	98.5	84.4	99.8	99.9	95.3	100	95.9	95.6	99.8	93.6	94.3	96.4
	Future	PaLM-S	92	36.3	15	77.5	32.6	95.5	60.2	20.6	89.1	86.8	58.1	51.2	59.6
		PaLM-M	99.9	94.2	93.6	94.8	91.5	95.4	98.6	59.9	72.2	88.4	30.4	69.3	82.4
		PaLM-L	100	98.4	98.9	96.2	100	99.2	99.8	81.2	91.4	95.6	98.4	92.4	96.0
Comparative	Standard	PaLM-S	75.3	57.3	69.3	96.7	73.7	71.7	79.7	62	43.3	25	55.7	38.1	62.3
		PaLM-M	92.3	69	80.7	83	82.7	77	58	71.7	69.7	26.3	61.3	73.1	70.4
		PaLM-L	100	87.3	100	98.7	100	100	99.3	100	98.7	86	99.3	91.6	96.7
	Conditional	PaLM-S	57.2	44.2	39.2	13.4	33.9	1.3	6.4	17.6	1.7	0	0.1	29.5	20.4
		PaLM-M	82.9	80.6	55.2	77	62.1	67.7	15.1	14.2	32.3	1.6	0.6	49.2	44.9
		PaLM-L	73.9	82.4	71.8	85.3	33.6	65.8	44.8	14.3	31	25	40.4	53.6	51.8

Table 7: Accuracy (in %) of PaLM-S, PaLM-M, and PaLM-L on generated tests in a zero-shot setting.

Language	Test and prediction	English gloss
Russian	C: Иногда он ходит в университет. Редко он ездит в театр. Q: Что он делает иногда? A: Ходит в университет. P: Идёт в университет.	C: Sometimes he goes (by foot) to the university. Rarely does he go (by transportation) to the theatre. Q: What does he do sometimes? A: Goes to the university (multiple times). P: Going to the university (one time).
Finnish	C: Äitini antoi isoäidilleni mukin. Isäni antoi sedälleni kameran. Q: Kenellä on uusi muki? A: Isoäidilläni. P: Isoäidilleni.	C: My mother gave my grandmother a mug. My father gave my uncle a camera. Q: Who has a new mug? A: My grandmother has. P: To my grandmother.
Chinese	C: 桌子旁边放着六样东西，都是狗。 Q: 多少狗在桌子旁边? A: 六只。 P: 六个。	C: Next to the table are six things, all are dogs. Q: How many dogs are next to the table? A: Six (measure word for animals). P: Six (generic measure word).
Swahili	C: Sadiki anakula saa nne usiku na anaendesha masaa matatu baadaye. Q: Anaendesha saa ngapi? A: Saa saba usiku P: Saa moja usiku.	C: Sadiki eats at 10 PM and then drives three hours after. Q: What time does he run? A: At 1 AM. P: At 7 PM.

Table 8: Examples of errors in PaLM-L predictions and English glosses for language-specific feature tests. Each example includes a context (C), question (Q), answer (A), and the model prediction (P). Tests probe motion verbs in Russian, possessives in Finnish, measure words in Chinese, and time expressions in Swahili.

Test type, Language	Test and prediction	English gloss
Negation Slovak	C: Pavol a Oskar nie sú vedci, ale Bohuš a Miroslav sú. Q: Kto sú vedci? A: Bohuš a Miroslav P: Bohús a Miroslav.	C: Pavol and Oskar are not scientists, but Bohuš and Miroslav are. Q: What are scientists? A: Bohuš and Miroslav. P: Bohús and Miroslav.
Numerals Russian	C: На столе три груши и девять арбузов. Елена съела одну грушу. Q: Сколько груш на столе? A: Две P: Два	C: There are three pears and nine watermelons on the table. Elena ate one pear. Q: How many pears are on the table? A: Two. (Feminine Nominative) P: Two. (Masculine Nominative)
Spatial Finnish	C: Mukit ovat ikkunan päällä ja tietokoneet tuolin alla. Q: Missä ovat tietokoneet? A: Tuolin alla. P: Tuolien alla.	C: The mugs are on the window and the computers are under the chair. Q: Where are the computers? A: Under the chair. (Genitive Singular) P: Under the chairs. (Genitive Plural)
Temporal Slovak	C: Peter a Katarína boli vedcami, ale Katarína zmenila zamestnanie a teraz je kuchárka. Q: Čím je Katarína? A: Kuchárkou. P: Kuchárka.	C: Peter and Katarína were scientists, but Katarína changed jobs and is now a cook. Q: Who is Katarína? A: Cook. (Instrumental) P: Cook. (Nominative)
Comparative Finnish	C: Jos vene olisi uudempi, Ylvä käyttäisi sitä. Jos pyörä olisi pienempi, Ylvä käyttäisi sitä. Q: Mitä Ylvä käyttäisi jos se olisi vähemmän vanha? A: Venettä. P: Vene.	C: If the boat was newer, Ylvä would use it. If the bike was smaller, Ylvä would use it. Q: What would Ylvä use if it was less old? A: Boat. (Partitive) P: Boat. (Nominative)

Table 9: Examples of morphological errors in PaLM-L predictions and English glosses for generated tests. Examples correspond to highlighted cells in Table 6. Each example includes a context (C), question (Q), answer (A), and the model prediction (P).

	Test type	Model	EN	ES	IT	FR	DE	SV	FI	SK	RU	ZH	SW	AR	Avg.	0-shot Δ
Negation	In context	mT5-XXL	99.6	97.3	98	97.7	92.1	98.6	96.6	97.5	98.3	73.1	97.8	63.4	91.9	35.0
		PaLM-S	92.2	88.2	91	69.5	85.8	87.7	87.4	83.8	73.6	81.3	92.4	45.6	81.5	46.4
		PaLM-M	99.8	99.9	99.4	99.9	99.2	99.5	99.1	99.1	96.4	96.9	88	61.6	94.9	26.7
		PaLM-L	99.7	100	99.9	100	100	99.8	100	99.6	100	99.7	99.9	95.1	99.5	2.9
		PaLM 2	99.6	100	99.9	100	100	99.4	99.9	98.4	100	100	99.9	98.1	99.6	1.3
	In question	mT5-XXL	75.2	78.4	74.4	76.9	74.2	79	74	71.6	77.6	51.7	75.2	60.3	72.1	53.2
		PaLM-S	39.9	44.8	33.5	23.1	35.8	38.5	35.5	37.2	44.4	40.8	38.3	33.8	37.1	0.5
		PaLM-M	78.2	92.6	93.8	95	92.6	96	94.2	90.9	73.6	75.1	81.3	61.8	85.4	32.5
		PaLM-L	97.7	99.8	99.9	99.4	99.4	99.3	99.9	99.7	99.7	97.8	98.5	93.6	98.8	10.4
		PaLM 2	96.2	100	98.5	99.8	99.9	90.6	86.9	99.4	99.9	98.2	99.5	96.9	97.2	0.0
Numerals	Addition	mT5-XXL	8.4	7.1	0.8	5.5	2.1	8.5	7.3	0.6	10.4	12.4	1.9	58	10.4	5.0
		PaLM-S	20.1	13.3	13.3	10.7	21	22.4	7.1	4.6	9.2	10.6	5.9	9.2	12.3	-11.5
		PaLM-M	95.7	71.8	69.7	89.3	90.7	61.2	50.3	47.5	81.7	87.2	10.8	18.9	64.6	32.8
		PaLM-L	99.3	100	96.7	100	99.5	96.9	80.7	79.2	83.8	97.2	72.1	71.3	88.9	11.2
		PaLM 2	100	100	100	100	100	100	99.9	98.3	99.8	100	89.6	95.1	98.4	7.3
	Subtraction	mT5-XXL	29.6	27	8.2	26.7	22.7	25.5	20.7	0.6	9.1	19.4	1.3	42.1	18.5	-4.5
		PaLM-S	25	23.4	18.6	23.9	25.4	21.7	13.8	16.3	16.7	16.1	10.8	14.5	18.9	2.6
		PaLM-M	56.7	64.1	61.3	58.8	49.7	29.8	34.7	26.7	40.6	38.8	11.7	38.2	42.6	6.2
		PaLM-L	92.5	94.3	94.8	97	79.2	96.2	95.4	85.8	86.4	89.9	39.2	64.2	83.9	7.5
		PaLM 2	99.9	99.8	100	99.9	96.1	100	98.6	99.8	99	88.1	60.9	98.9	94.6	6.9
Spatial	Prepositions	mT5-XXL	91.1	90.6	66.8	93.4	22.7	84.7	73.8	2.6	41	85.8	9.1	81.2	59.2	3.5
		PaLM-S	79.9	50	52.8	54.6	67.8	48.5	52	45.8	53.7	97	42.4	41.2	57.1	-1.0
		PaLM-M	99.6	97	97.5	99.2	94.6	92.5	83.7	93.8	93.7	97.9	77.2	78.1	92.1	9.6
		PaLM-L	100	100	100	100	100	100	99.2	100	100	100	99.3	97.7	99.7	2.6
		PaLM 2	100	100	100	98.4	100	100	100	100	100	100	100	100	99.9	0.5
	Position	mT5-XXL	98.1	100	95.8	99.9	98.5	97.4	99.9	100	100	96.1	72.6	73.1	93.9	36.3
		PaLM-S	85.5	67.2	66.4	59.1	68.7	66.7	88	75	38.3	99	53.6	36.7	67.0	27.4
		PaLM-M	99.9	93.5	99.6	100	99.6	91	98.6	98.7	93.1	99.6	97	81.6	96.0	46.4
		PaLM-L	100	99.9	99.9	99.9	100	99.9	99.8	100	99.5	99.9	85.9	81.2	96.9	31.4
		PaLM 2	100	100	100	99.9	100	100	100	100	100	100	99.9	99.9	100	5.7
Temporal	Past	mT5-XXL	90.4	99.1	90.9	96.4	93.8	87.7	87	97.7	91.5	90.9	86.5	75.7	90.7	28.8
		PaLM-S	95.1	75.1	84.2	94.4	44.6	84	57.7	32.6	96.8	78	77.7	54.9	72.9	20.3
		PaLM-M	94.5	91.8	61.7	75	79.7	53.2	41.5	60.5	34.3	84.5	84.4	74.8	69.7	-18.7
		PaLM-L	99.8	98.6	97.2	97	99.6	99.8	99.9	96.1	99	99.9	100	99.4	98.8	2.7
		PaLM 2	100	99.9	100	100	100	100	100	100	99.9	100	100	99.4	99.9	7.9
	Future	mT5-XXL	90.7	96.4	98	92.7	93.4	91	84.5	91.3	89.5	86.5	89.6	61.2	88.6	30.4
		PaLM-S	98.6	54.9	57.4	90.3	45	71.1	44.7	13.5	98.7	72.8	79.2	59.3	65.5	5.9
		PaLM-M	92.4	93.5	93.2	69.8	90.2	58.4	55.5	61.4	47.8	62.1	42.2	78.2	70.4	-12.0
		PaLM-L	100	97.4	98.8	92.5	99.3	100	100	93.2	99.2	99.7	100	98.4	98.0	2.5
		PaLM 2	100	100	100	100	100	100	100	100	100	100	100	100	100	0.6
Comparative	Standard	mT5-XXL	86	90.3	92.3	81.7	82.7	83.7	92.3	85	89.3	85.3	81	71.2	85.0	47.1
		PaLM-S	90.3	97.7	90	87.3	94	77	76	89.3	89.3	56.3	98	47.2	82.7	20.4
		PaLM-M	88	90.3	97.7	84	91	93.3	80.7	83	91.7	88.7	97	87.1	89.4	19.0
		PaLM-L	100	100	100	99.7	100	100	100	99.7	100	99.7	100	94.9	99.5	3.0
		PaLM 2	100	100	100	100	100	99.3	100	100	100	100	100	100	99.9	0.2
	Conditional	mT5-XXL	47.4	72.2	73.9	64.5	61.7	74.4	72.7	62.6	42.4	69.1	73.5	74.5	66.7	43.2
		PaLM-S	86.2	74.9	60.5	70.1	38.5	70.6	88.6	43.8	36.7	78.5	77.2	55.3	65.1	44.7
		PaLM-M	72.5	82.9	76	67.3	76.3	60.1	61.3	52.6	39.9	73.8	78.6	71.3	67.7	22.8
		PaLM-L	75.7	70.4	83.2	70.8	53.1	47.2	48.9	35.6	30.6	63	50.3	82.8	57.8	8.0
		PaLM 2	93.7	96.1	97.2	93.9	83.5	82.3	86.9	22.5	47.8	93.1	77.4	85.1	78.7	14.6

Table 10: Accuracy (in %) of mT5-XXL, PaLM-S, PaLM-M, PaLM-L, and PaLM 2 on generated tests in a one-shot setting. The right-most column shows the relative change compared to the zero-shot setting for each model.

	EN	ES	IT	FR	DE	SV	FI	SK	RU	ZH	SW	AR	Avg
mT5-XXL	20.3%	136.8%	59.2%	77.5%	27.8%	85.9%	58.1%	114.2%	66.0%	67.7%	92.1%	22.1%	69.0%
PaLM-S	7.3%	51.7%	55.3%	21.8%	11.8%	10.4%	38.5%	85.2%	64.5%	40.9%	145.6%	35.1%	47.3%
PaLM-M	3.9%	23.8%	41.3%	7.2%	20.2%	11.0%	30.9%	41.3%	28.2%	46.0%	90.4%	33.4%	31.5%
PaLM-L	4.3%	7.3%	8.8%	3.9%	7.3%	3.6%	5.7%	15.8%	11.6%	15.5%	19.7%	12.5%	9.7%
PaLM 2	0.9%	1.4%	6.4%	0.9%	3.1%	0.2%	9.7%	3.8%	1.7%	10.9%	10.6%	6.7%	4.7%

Table 11: Average relative improvement of the one-shot vs the zero-shot setting for all models across all languages.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
In the Broader Impact Statement.

☑ A2. Did you discuss any potential risks of your work?
In the Broader Impact Statement.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Section 1.

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Section 4. A new behavioral testing library.

☑ B1. Did you cite the creators of artifacts you used?
Section 4.

☒ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
The library will be released under a Creative Commons license.

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
We describe how the library should be used in Section 4 and Figure 1.

☒ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
No created data contains personally identifying information.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Sections 4 and 5.

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Yes, in Section 6.

C ☑ Did you run computational experiments?

Section 6.

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Section 6.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 6.

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Yes.

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Yes.

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

All templates were created by the authors of the paper.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.