CUGE: A Chinese Language Understanding and Generation Evaluation Benchmark

Yuan Yao[1], **Qingxiu Dong**[2], **Jian Guan**[1], **Boxi Cao**[3], **Zhengyan Zhang**[1], **Chaojun Xiao**[1]
Xiaozhi Wang[1], **Fanchao Qi**[1], **Junwei Bao**[4], **Jinran Nie**[5], **Zheni Zeng**[1], **Yuxian Gu**[1]
Kun Zhou[6], **Xuancheng Huang**[1], **Wenhao Li**[1], **Shuhuai Ren**[2], **Jinliang Lu**[7], **Chengqiang Xu**[1]
Huadong Wang[1], **Guoyang Zeng**[1], **Zile Zhou**[8], **Jiajun Zhang**[7], **Juanzi Li**[1], **Minlie Huang**[1]
Rui Yan[8], **Xiaodong He**[4], **Xiaojun Wan**[9], **Xin Zhao**[6], **Xu Sun**[2], **Yang Liu**[1]
Zhiyuan Liu[1*], **Xianpei Han**[3*], **Erhong Yang**[5*], **Zhifang Sui**[2*], **Maosong Sun**[1*]
[1]Department of Computer Science and Technology, Tsinghua University
[2]MOE Key Lab of Computational Linguistics, School of EECS, Peking University
[3]Institute of Software, Chinese Academy of Sciences [4]JD AI Research, Beijing, China
[5]School of Information Science, Beijing Language and Culture University
[6]School of Information, Renmin University of China
[7]National Laboratory of Pattern Recognition, Institute of Automation, CAS
[8]Gaoling School of Artificial Intelligence, Renmin University of China
[9]Wangxuan Institute of Computer Technology, Peking University
Beijing Academy of Artificial Intelligence

Abstract

Realizing general-purpose language intelligence has been a longstanding goal for natural language processing, where standard evaluation benchmarks play a fundamental and guiding role. We argue that for general-purpose language intelligence evaluation, the benchmark itself needs to be comprehensive and systematic. To this end, we propose **CUGE**, a **C**hinese Language **U**nderstanding and **G**eneration **E**valuation benchmark with the following features: (1) Hierarchical benchmark framework, where datasets are principally selected and organized with a language capability-task-dataset hierarchy. (2) Multi-level scoring strategy, where different levels of model performance are provided based on the hierarchical framework. To facilitate CUGE, we provide a public leaderboard that can be customized to support flexible model judging criteria. Evaluation results on representative pre-trained language models indicate ample room for improvement towards general-purpose language intelligence. CUGE is publicly available at `cuge.baai.ac.cn`.

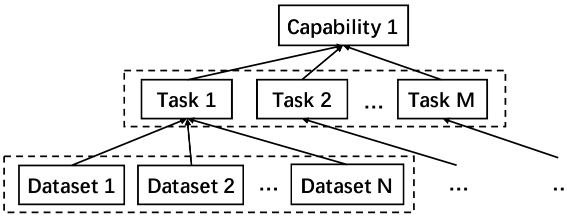


Figure 1: CUGE selects and organizes datasets in a language capability-task-dataset hierarchical framework, based on which multi-level model scores are provided.

1 Introduction

Human language intelligence is general across different language capabilities and tasks. Realizing such general-purpose language intelligence has been a longstanding goal for natural language processing (NLP), where standard evaluation benchmarks play a fundamental and guiding role. To this end, rather than focusing on model performance on specific datasets, researchers have proposed several standard evaluation benchmarks that summarize model performance on a collection of diverse datasets (Wang et al., 2018, 2019; Xu et al., 2020; Liu et al., 2020). Notably, GLUE (Wang et al., 2018) and SuperGLUE (Wang et al., 2019) benchmarks have drawn increasing attention from NLP communities recently and have greatly promoted the development of NLP techniques, especially in the era of pre-trained language models (Devlin et al., 2019; Brown et al., 2020; Raffel et al., 2020).

[*] Corresponding Authors: Z. Liu (liuzy@tsinghua.edu.cn), X. Han (xianpei@iscas.ac.cn), E. Yang (yerhong@blcu.edu.cn), Z. Sui (szf@pku.edu.cn), M. Sun (sms@tsinghua.edu.cn)

1. Update Note (April 14th, 2022): We add two new datasets, including grammatical error correction dataset YACLC from Beijing Language and Culture University, and reading comprehension dataset GCRC from Shanxi University, and also improve the description consistency of all datasets.

Despite their popularity, there are still important limitations of existing benchmarks. (1) *Flat benchmark framework.* Most existing evaluation benchmarks are dataset-oriented, where commonly used datasets are selected and loosely organized in a flat structure without the guidance of language capabilities to be evaluated. This lack of comprehensiveness and systematicness undermine the reliability of a benchmark as the indicator of general-purpose language intelligence. (2) *Oversimplified scoring strategy.* Most benchmarks summarize the model performance via a simple average of metrics across different datasets, without considering the characteristics of different metrics and datasets. Moreover, since the metrics on all datasets are directly averaged into a single overall value, the performance on language capabilities and tasks cannot be reflected.

To better benchmark general-purpose language intelligence, we propose **CUGE**, a **C**hinese Language **U**nderstanding and **G**eneration **E**valuation benchmark with the following features: (1) *Hierarchical benchmark framework.* As shown in Figure 1, CUGE selects and organizes datasets in a language capability-task-dataset hierarchical framework, covering 7 important language capabilities, 18 mainstream NLP tasks and 21 representative datasets. The framework is carefully designed according to the human language examination syllabus and the current NLP research status. With the guidance of the framework, we expect the dataset selection and organization more principled and better reflect general-purpose language evaluation needs. (2) *Multi-level scoring strategy.* Based on the hierarchical framework, in addition to the overall score, CUGE provides model performance evaluation on different levels, including performance on datasets, tasks and language capabilities, more systematically investigating and showing model language intelligence. Moreover, CUGE normalizes the score on each dataset according to the performance of a standard baseline model, largely mitigating the influence of different metrics and datasets.

To facilitate GUGE, we release an online evaluation platform and a public leaderboard. The leaderboard can be easily customized according to model capabilities and tasks in interest to support flexible model judging criteria. Experiments results on representative pre-trained language models show that, although strong performance can be achieved by existing NLP techniques in some language capabilities and tasks, there is still ample room for improvement towards general-purpose language intelligence. The online evaluation platform and leaderboard are publicly available at cuge.baai.ac.cn.

2 Design Principles

The design principles of CUGE are highly correlated to the target of evaluating and innovating general-purpose language intelligence. In this section, we introduce the design principles of the core components of CUGE, including the hierarchical framework, multi-level scoring strategy, dataset evaluation system and the evaluation platform.

Hierarchical Benchmark Framework For general-purpose language intelligence evaluation, the benchmark itself needs to be comprehensive and systematic. To this end, different from existing flat dataset-oriented benchmarks built from the bottom up, CUGE is hierarchically constructed in a top-down approach. Specifically, on the top level, according to the human language examination syllabus (i.e., Chinese syllabus for college entrance examination[1]) and the current NLP research status, we summarize general-purpose language intelligence into 7 important language capabilities: (1) language understanding: word-sentence level, (2) language understanding: discourse level, (3) information acquisition and question answering, (4) language generation, (5) conversational interaction, (6) multilingualism, and (7) mathematical reasoning. Then we identify representative NLP tasks for each capability, serving as the task level. Finally, we select representative datasets for each task, resulting in the dataset level. Compared with previous works, the hierarchical benchmark framework of CUGE more systematically organizes existing evaluation resources and more comprehensively reflects general-purpose language evaluation needs.

Multi-level Scoring Strategy In addition to a single overall value, we would like to also provide fine-grained, multi-level evaluation of models. Therefore, based on the hierarchical benchmark framework of CUGE, we propose a multi-level scoring strategy for the evaluation results. In addition, it is also desirable to eliminate the influence of different metrics and datasets during score calculation. We thus normalize the scores on datasets

[1] www.neea.edu.cn

Corpus	|Train|	|Dev|	|Test|	Task	Metrics	Domain
Language Understanding: Word-Sentence Level						
PKU-SEG	40.3k	10.5k	9.9k	word segmentation	F1 score	news
WordSeg-Weibo	20.1k	2.1k	8.6k	word segmentation	F1 score	social media
PKU-SEGPOS	31.7k	5.2k	4.8k	word segmentation and POS	F1 score	news
CCPM	21.8k	2.7k	2.7k	classical poetry matching	F1 score	Chinese poetry
CMeEE	15.0k	5.0k	3.0k	named entity recognition	F1 score	medical
FinRE	7.5k	1.5k	3.7k	relation extraction	F1 score	financial
YACLC	8.0k	1.0k	1.0k	grammatical error correction	F1 score	essays
Language Understanding: Discourse Level						
SPR	12.7k	1.6k	4.8k	humor detection	F1 score	TV show
ClozeT	0.6k	0.3k	0.3k	story cloze test	accuracy	literature
C^3	11.9k	3.8k	3.8k	reading comprehension	accuracy	mixed-genre
GCRC	7.0k	0.8k	0.8k	reading comprehension	accuracy, F1 score	mixed-genre
Information Acquisition and Question Answering Capability						
WantWords	78.0k	19.0k	19.0k	reverse dictionary	accuracy@1	mixed-genre
KBQA	24.0k	-	0.6k	open-domain question answering	EM	mixed-genre
Sogou-Log	8,052k	500k	1.0k	document retrieval	nDCG@k	mixed-genre
Language Generation Capability						
LCSTS	2,401k	9.0k	0.7k	text summarization	Rouge	news
CEPSUM	434k	5.0k	5.0k	text summarization	Rouge	e-commerce
E-Reviews	115k	1.0k	3.0k	data-to-text generation	BLEU, Distinct	e-commerce
Conversational Interaction Capability						
KdConv	62.9k	9.0k	9.1k	knowledge-driven conversation	BLEU, Distinct	film, music, travel
Multilingual Capability						
WMT20-EnZh	21,000k	4.0k	4.0k	machine translation	BLEU	mixed-genre
NCLS-EnZh	365k	3.0k	3.0k	cross-lingual text summarization	Rouge	mixed-genre
Mathematical Reasoning Capability						
Math23k	21.0k	1.0k	1.0k	mathematical computation	accuracy	math word problem

Table 1: Descriptions and statistics of language capabilities, tasks and datasets.

based on representative standard baseline models.

3D Dataset Evaluation System The quality of datasets is crucial to a benchmark. However, there is little systematic investigation on this problem in existing benchmarks. To this end, we propose a three-dimensional evaluation system for CUGE's dataset selection and evaluation. Inspired by the assessment of reliability (Cronbach, 1946) and validity (Alloway et al., 2008) in educational psychology, we propose *reliability*, *difficulty*, and *validity* as three main dimensions of our dataset evaluation system. We expect the dataset evaluation system will contribute to a more scientific and reliable dataset selection process of CUGE in the future.

Specifically, (1) *reliability* refers to the consistency and accountability of the scores given by the datasets and the corresponding metrics. (2) *Difficulty* reflects the hardness and discriminative capability of the datasets. (3) *Validity* represents the relevance of the evaluation dataset and evaluation goals. The evaluation methods and metrics for dataset quality assessment are still open research problems, and CUGE is actively exploring this direction for building high-quality benchmarks.

Previous works have shown that evaluation benchmarks need to be continually adjusted according to the current status of NLP research (Wang et al., 2018, 2019). We expect the 3D dataset evaluation system can also serve as a principled criterion for the continual adjustment of datasets in CUGE in the future.

Flexible Online Platform To facilitate CUGE, we build an online evaluation platform that features customizable leaderboards and interactive forums. In addition to the leaderboard for each dataset, participants can also customize leaderboards consisting of multiple datasets according to their evaluation objectives. To encourage academic integrity, participants are required to check the honor code before submissions, and interactive forums are provided for submission discussion.

In addition to the evaluation of the existing datasets in CUGE, the platform also supports the release and evaluation of new datasets. Each dataset submitted and released on CUGE will have a dataset-specific leaderboard on the platform. The CUGE benchmark will adjust its dataset collection according to the results of existing and newly released datasets on the platform each year.

3 Benchmark Framework

CUGE selects and organizes datasets in a language capability-task-dataset hierarchical framework, covering 7 language capabilities, 18 mainstream NLP tasks and 21 representative datasets.

3.1 Language Understanding: Word-Sentence Level

The language understanding: word-sentence level evaluates the model's capability to understand a given text and perform word- and sentence-level syntactic and semantic tasks.

3.1.1 Word Segmentation

The task is to identify the sequence of words in a sentence and mark the boundaries between words.

PKU-SEG The dataset (Emerson, 2005) is based on the PKU dataset released by The Second International Chinese Word Segmentation Bakeoff in 2005. This dataset is annotated from the news corpus of the People's Daily. We further add the data annotated from the corpus of the People's Daily in January and December 2000. Finally, the data is re-integrated and divided.

WordSeg-Weibo The dataset (Qiu et al., 2016) comes from the NLPCC 2016 evaluation task. The dataset is collected from Sina Weibo website[2]. Different from the traditional single word segmentation evaluation method, this dataset introduces a new multi-granularity word segmentation evaluation criterion. Besides the training data, we also provide the background data, from which the training and test data are drawn.

3.1.2 Word Segmentation and POS Tagging

The task seeks to identify the boundaries between words and assign a pre-defined part-of-speech tag to each word in a given sentence.

PKU-SEGPOS The dataset is the part-of-speech (POS) tagging dataset collected from the corpus of People's Daily. The corpus from January 2000 and December 1-15, 2020, the corpus from December 16-23, 2000, and the corpus from December 24-31, 2000 compose the training set, the validation set, and the test set, respectively.

3.1.3 Classical Poetry Matching

The task is that given a modern description of Chinese classical poetry, the model is supposed to select one from four candidate poems that semantically matches the given description most.

CCPM The dataset (Li et al., 2021) is a multiple-choice dataset for Chinese classical poetry matching. This dataset comes from the Chinese classical poems and their corresponding modern Chinese translations provided on the website.

3.1.4 Named Entity Recognition

The task seeks to locate and classify named entities in unstructured text into pre-defined categories.

CMeEE The dataset (Hongying et al., 2020) is based on the CHIP2020 evaluation. Given a predefined schema, the task is to identify and extract entities from the given sentence and classify them into nine categories: disease, clinical manifestations, drugs, medical equipment, medical procedures, body, medical examinations, microorganisms, and department.

3.1.5 Entity Relation Extraction

The task is to extract semantic relations between entity pairs in given sentences.

FinRE The dataset (Li et al., 2019) is a manually labeled financial news relation extraction dataset. Given a sentence and its head and tail entities, the model needs to predict the relation between the head and tail entities. This dataset is annotated from the Sina Finance News corpus, in which the named entity is a commercial company, and the relation contains 44 financial relation categories and an NA category, including special relation categories in the financial and financial fields such as ownership, shareholding, competition, acquisition, transaction, cooperation, and shareholding.

3.1.6 Grammatical Error Correction

The task aims to judge the acceptability and correct the grammatical errors of the given text.

YACLC The dataset (Wang et al., 2021) is collected from a language learning website [3], where

[2] www.weibo.com

[3] lang-8.com

Chinese as foreign language learners share their essays. Given a sentence produced by language learners, models are required to (1) evaluate the grammatical correctness, (2) correct the grammatical errors with minimal edits, and (3) make the sentence more fluent with minimal edits.

3.2 Language Understanding: Discourse Level

The language understanding: discourse level evaluates the model's capability to understand a given text and perform discourse-level syntactic and semantic tasks.

3.2.1 Humor Detection

The task is to recognize, classify and generate humor based on computer technology, which has important theoretical and application value.

SPR The dataset[4] selects "I Love My Home" as data source. According to the changes in the scene and plot, the sitcom is divided into several dialogues. In a conversation, there are different characters to communicate, resulting in continuous utterances. Utterances in the same dialogue appear in order and have a contextual relationship. Compared with single-sentence humor, the humor in dialogue may come from the context, rather than the content of the utterance itself. Models are required to judge whether the utterance is humorous based on the context and content, and to identify the punchline in sitcoms.

3.2.2 Story Cloze Test

The task is to select the missing sentence from multiple candidates to fill into a story and form a reasonable logical plot.

ClozeT The dataset comes from children's stories crawled from the Web. When constructing options, crowd-sourced annotators are asked to extract a sentence from the story that can be inferred based on the context and common sense as the correct option, and rewrite it to a sentence contrary to common sense as the wrong option.

3.2.3 Reading Comprehension

This task is to answer questions about given unstructured texts.

C^3 The dataset (Sun et al., 2019) is a multiple-choice Chinese machine reading comprehension dataset, which is collected from test questions for Chinese as a second language learners. Given a Chinese paragraph/dialogue and a question, the dataset provides a number of answer candidates. According to the given content, models are required to choose the correct answers from the candidates.

GCRC The dataset (Tan et al., 2021) is a multiple-choice Chinese machine reading comprehension dataset, which is collected from Chinese college entrance examination. Given an article and a question, models are required to select the correct answers from candidates. The dataset features explainable evaluation for reading comprehension systems. In addition to answering the questions, models are also asked to (1) select the supporting evidence sentences from the article to answer the question, (2) determine the reason for rejecting each wrong answer, and (3) predict the skills required to answer the question.

3.3 Information Acquisition and Question Answering Capability

Information acquisition and question answering capability requires retrieving the answer for a query from given non-structured documents or structured knowledge bases.

3.3.1 Reverse Dictionary

The reverse dictionary task requires taking the description of a target word as input and outputting the target word.

WantWords The dataset (Zheng et al., 2020) is constructed based on dictionary definitions, which are collected from Modern Chinese Dictionary (6th Edition). Models are required to select the word described by the text query from the vocabulary.

3.3.2 Open-domain Question Answering

Open-domain question answering is the task of answering a question based on a knowledge source.

KBQA The dataset[5] is based on the open-domain question answering shared task 7 in NLPCC 2018. Given a natural language question, models are required to produce answers based on the background knowledge base in open domain.

[4] cips-cl.org/static/CCL2020/humorcomputation.html

[5] tcci.ccf.org.cn/conference/2018/taskdata.php

3.3.3 Document Retrieval

The task requires retrieving the relevant document for a given query.

Sogou-Log The dataset (Luo et al., 2017) contains 35 million search sessions with 96k distinct queries collected from a Chinese commercial search engine[6]. The query log consists of corresponding queries, displayed documents, user clicks and dwell times. Each query has 12 documents on average, which tend to be high-quality since the results are collected from a mainstream search engine. The queries in the test sets are sampled from those that appear more than 1k times in all query logs, i.e., head queries. Besides, the dataset uses the performance on the tail queries to evaluate the model robustness.

3.4 Language Generation Capability

The Language generation capability requires generating readable natural language texts conditioned on given inputs.

3.4.1 Text Summarization

The task requires generating a short text to include the important information of a given long text.

LCSTS The dataset (Hu et al., 2015) is a text summarization dataset collected from Sina Weibo. This corpus contains more than 2M Chinese texts paired with short summaries, which are typically written by the corresponding authors. LCSTS also provides 10.7k summaries that are manually annotated with relevance scores to the corresponding texts. We only take those examples whose relevance scores are larger than 2 as the test set.

CEPSUM The dataset (Yuan et al., 2020) is a collection of product summaries on a mainstream Chinese e-commerce platform JD. The products come from two categories including home appliances and bags. The summaries are generated by thousands of experts, and the auditing groups of the e-commerce platform verify the quality.

3.4.2 Data-to-Text Generation

The task requires generating natural language texts from structured data.

E-Reviews The dataset (Shao et al., 2019) is collected from a Chinese e-commerce platform. Given a table that contains a set of attribute-value pairs to describe a commodity, models are asked to produce natural language advertising texts.

3.5 Conversational Interaction Capability

The task requires generating fluent and reasonable responses to users' posts.

3.5.1 Knowledge-driven Conversation Generation

KdConv The dataset (Zhou et al., 2020) is collected for research on knowledge-driven multi-turn conversation. KdConv consists of 4.5k dialogues that cover three domains including film, music, and travel, where each domain contains 1.5k dialogues. Therefore, the dataset can be used to explore knowledge transfer among these domains. Besides, these dialogues cover diversified topics and are manually annotated with related knowledge facts.

3.6 Multilingual Capability

Multilingual capability requires handling inputs and outputs in multiple languages.

3.6.1 Machine Translation

The task requires translating a natural language text to another specified language and maintaining the semantics.

WMT20-EnZh The dataset (Barrault et al., 2020) is based on the shared task of the workshop on machine translation in 2020, where the corpus is collected from news websites. We focus on machine translation between Chinese and English. The test set for English to Chinese is produced by translating at the paragraph level.

3.6.2 Cross-lingual Text Summarization

Given a document from the source language, the task requires generating a short text that summarizes the key information in the target language.

NCLS-EnZh The dataset (Zhu et al., 2019) is constructed based on round-trip translation, where the corpus comes from CNN, DailyMail and Sina Weibo websites. The reference texts in the test set are further corrected by multiple human annotators. Here we also focus on the English-to-Chinese summarization.

3.7 Mathematical Reasoning Capability

The capability requires inferring the answer to a math word problem.

[6] www.Sogou.com

	NLU-WSL	NLU-DL	IA&QA	NLG	CI	ML	MR	CUGE Index
mT5-Small	87.70 (100)	41.50 (100)	29.20 (100)	33.10 (100)	8.76 (100)	9.10 (100)	18.40 (100)	100
mT5-Large	89.90 (103)	56.30 (136)	31.65 (108)	34.40 (104)	9.76 (111)	11.10 (122)	34.30 (186)	117
mT5-XXL	90.60 (103)	86.40 (208)	35.90 (123)	34.80 (105)	12.68 (145)	24.00 (264)	61.60 (335)	152
CPM-2	91.60 (104)	86.10 (207)	35.90 (123)	35.90 (108)	13.12 (150)	26.20 (288)	69.40 (377)	157

Table 2: Performance of representative pre-trained language models on lite version of CUGE (%). In addition to the raw performance, we also report the performance normalized by mT5-small (in parenthese), based on which the overall CUGE Index is calculated. NLU-WSL: language understanding: word-sentence level, NLU-DL: language understanding: discourse level, IA&QA: information acquisition and question answering, NLG: language generation, CI: conversational interaction, ML: multilingualism, and MR: mathematical reasoning.

3.7.1 Mathematical Computation

Math23k The dataset (Wang et al., 2017) is a collection of math word problems from online education websites. The problems can be solved based on knowledge of linear algebra with only one variable. In addition to the answer, each problem in the dataset is also annotated with an equation for solving the problem.

Capability	Task	Dataset
NLU-WSL	Classical Poetry Matching	CCPM
NLU-DL	Reading Comprehension	C^3
IA&QA	Document Retrieval	Sogou-Log
NLG	Text Summarization	LCSTS
CI	Conversation Generation	KdConv
ML	Machine Translation	WMT20-EnZh
MR	Mathematical Computation	Math23K

Table 3: Lite version of CUGE. Each capability is instantiated with the most representative task and dataset. NLU-WSL: language understanding: word-sentence level, NLU-DL: language understanding: discourse level, IA&QA: information acquisition and question answering, NLG: language generation, CI: conversational interaction, ML: multilingualism, and MR: mathematical reasoning.

4 Model Scoring Strategy

In terms of model scoring, we adopt a multi-level scoring and model-based normalization strategy.

Multi-level Scoring Based on the hierarchical benchmark framework, we aggregate model scores from the bottom up. Based on the scores on the dataset level (e.g., F1 score and accuracy), the score of the corresponding task can be obtained from the normalized average of dataset scores. Similarly, the capability scores are the average of task scores, and the overall CUGE score is the average of capability scores. Through this strategy, we put forward a multi-level, fine-grained evaluation of the whole capability-task-dataset framework.

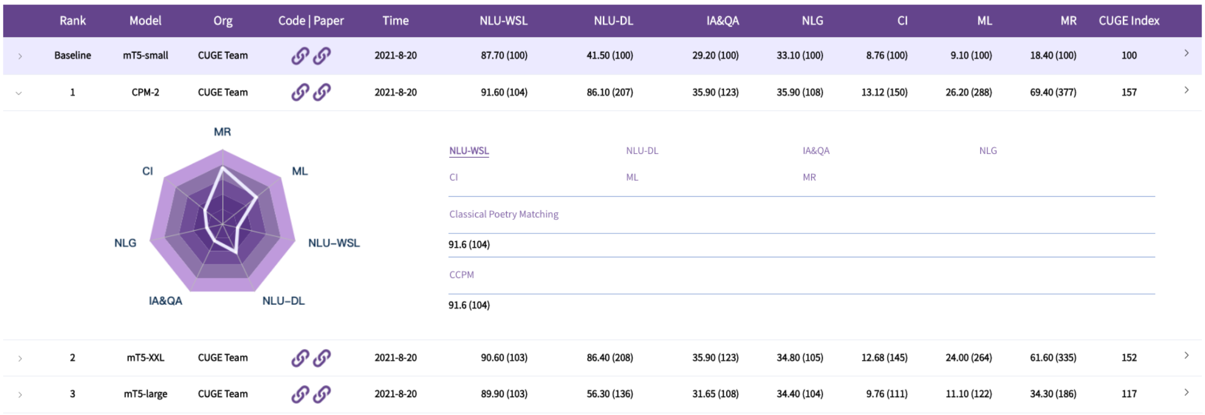
Score Normalization Existing NLP benchmarks typically calculate overall scores using the average of model scores on datasets. However, this procedure neglects characteristics of different datasets and metrics. For instance, the numerical results of BLEU (Papineni et al., 2002) are typically small as compared with F1 scores. As a result, with score averaging, the overall score can be dominated by metrics with larger scale.

To address the issue, we normalize the scores based on scores of representative standard baseline models, i.e., mT5-Small (Xue et al., 2020), so as to eliminate the influence of disturbing factors such as different metrics. Specifically, the normalized score on dataset is given by p/b, where p and b are dataset performance of model under evaluation and the standard baseline model respectively. Normalization on standard model performance essentially gives a ratio to different metrics and datasets, therefore making the overall score more reasonable.

5 Using CUGE

To facilitate CUGE, we build a public online evaluation platform. Participants can view the benchmark framework and leaderboard, download datasets, and participate in the evaluation by submitting prediction files. Specifically, CUGE platform enjoys the following notable features:

Customizable Leaderboard CUGE characterizes datasets with multi-dimensional labels, such as language capability and task. Users can customize the leaderboard by selecting the labels, supporting flexible model judging criteria. CUGE also recommends a standard lite leaderboard, as shown in Table 3. Specifically, for each language capability, we first select the most representative task and the most representative dataset of this task, and then combine the datasets to build the lite leaderboard, which enables convenient and rapid evaluation on GUGE platform, as shown in Figure 2.



	Rank	Model	Org	Code | Paper	Time	NLU-WSL	NLU-DL	IA&QA	NLG	CI	ML	MR	CUGE Index	
›	Baseline	mT5-small	CUGE Team	🔗 🔗	2021-8-20	87.70 (100)	41.50 (100)	29.20 (100)	33.10 (100)	8.76 (100)	9.10 (100)	18.40 (100)	100	›
⌄	1	CPM-2	CUGE Team	🔗 🔗	2021-8-20	91.60 (104)	86.10 (207)	35.90 (123)	35.90 (108)	13.12 (150)	26.20 (288)	69.40 (377)	157	›
›	2	mT5-XXL	CUGE Team	🔗 🔗	2021-8-20	90.60 (103)	86.40 (208)	35.90 (123)	34.80 (105)	12.68 (145)	24.00 (264)	61.60 (335)	152	›
›	3	mT5-large	CUGE Team	🔗 🔗	2021-8-20	89.90 (103)	56.30 (136)	31.65 (108)	34.40 (104)	9.76 (111)	11.10 (122)	34.30 (186)	117	›

NLU-WSL NLU-DL IA&QA NLG
CI ML MR

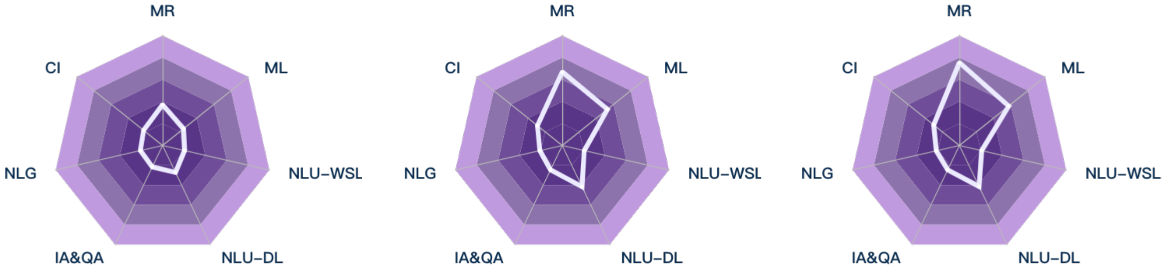
Classical Poetry Matching

91.6 (104)

CCPM

91.6 (104)

Figure 2: Lite leaderboard of CUGE platform.



(a) mT5-Large (b) mT5-XXL (c) CPM-2

Figure 3: Capability performance visualization for representative pre-trained language models on CUGE.

Academic Integrity Encouragement As a standard evaluation benchmark, CUGE highly values academic integrity of evaluation participants. Therefore, participants are required to check the honor code before submissions, which forbids usage of human labels on the CUGE test set in any form. Participants are also encouraged to release technique reports and codes on the leaderboard.

Interactiveness In the future, CUGE will provide official forums for users to encourage discussion on: (1) Submissions on leaderboard, which helps better understanding of submissions and supports supervision for academic integrity; (2) Datasets in CUGE, which helps the adjustment of CUGE datasets.

6 Experiments

In this section, we present experimental results of representative baseline models on CUGE.

6.1 Experimental Setup

In our experiments, we select mT5-Small (Xue et al., 2020), a representative pre-trained language model with 300M parameters as our standard baseline model to normalize the performance of models under evaluation. mT5 adopts an encoder-decoder structure, and is capable of both language understanding and generation, making it a good standard baseline model for CUGE. Based on the normalization of mT5-Small, we evaluate mT5-Large (1.2B), mT5-XXL (13B), and CPM-2 (11B) (Zhang et al., 2021) on the lite version of CUGE.

6.2 Results

We report the experimental results in Table 2 and visualize the normalized capability performance in Figure 3, from which we have the following observations: (1) mT5-XXL significantly outperforms mT5-Large and mT5-small, showing that increased number of parameters can lead to better performance in different language capabilities. (2) With smaller model size, CPM-2 outperforms mT5-XXL, which shows that improved pre-training procedure can produce stronger language capabilities of pre-trained language models. (3) However, the performance improvement over different language capabilities is highly imbalanced. For example, the improvement of language generation capability is substantially smaller than multilingual capability. Such overall evaluation and investigation of language capabilities cannot be achieved in existing benchmarks. The results show the necessity of constructing a comprehensive and systematic evaluation benchmark to better evaluate and guide general-purpose language intelligence research.

7 Conclusion and Future Work

In this work, we present CUGE, a Chinese language understanding and generation evaluation benchmark. CUGE features capability-task-dataset hierarchical framework and multi-level scoring strategy. To facilitate CUGE, we build a public online evaluation platform that supports customizable leaderboards. Experimental results on representative pre-trained language models indicate ample room for improvement towards general-purpose language intelligence.

Note that CUGE and its platform are still in progress. The current version of CUGE mainly adopts existing datasets for model evaluation, and the platform only implements part of the ultimate design. In the future, we plan to (1) continually build high-quality datasets tailored for the evaluation objective of CUGE, (2) conduct more detailed dataset quality evaluation, and (3) fully implement the design of CUGE platform.

8 Acknowledgements

We would like to thank Beijing Academy of Artificial Intelligence (BAAI) for providing support.

References

Tracy P Alloway, Susan E Gathercole, Hannah Kirkwood, and Julian Elliott. 2008. Evaluating the validity of the automated working memory assessment. *Educational Psychology*, 28(7):725–734.

Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, et al. 2020. Findings of WMT20. In *Proceedings of WMT20*, pages 1–55, Online. Association for Computational Linguistics.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Lee J Cronbach. 1946. A case study of the splithalf reliability coefficient. *Journal of educational psychology*, 37(8):473.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL*, pages 4171–4186.

Thomas Emerson. 2005. The second international Chinese word segmentation bakeoff. In *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*.

Zan Hongying, Li Wenxin, Zhang Kunli, Ye Yajuan, Chang Baobao, and Sui Zhifang. 2020. Building a pediatric medical corpus: Word segmentation and named entity annotation. In *Workshop on Chinese Lexical Semantics*, pages 652–664.

Baotian Hu, Qingcai Chen, and Fangze Zhu. 2015. LCSTS: A large scale chinese short text summarization dataset. In *Proceedings of EMNLP*, pages 1967–1972.

Wenhao Li, Fanchao Qi, Maosong Sun, Xiaoyuan Yi, and Jiarui Zhang. 2021. CCPM: A chinese classical poetry matching dataset. *arXiv preprint arXiv:2106.01979*.

Ziran Li, Ning Ding, Zhiyuan Liu, Haitao Zheng, and Ying Shen. 2019. Chinese relation extraction with multi-grained information and external linguistic knowledge. In *Proceedings of ACL*, pages 4377–4386.

Dayiheng Liu, Yu Yan, Yeyun Gong, Weizhen Qi, Hang Zhang, Jian Jiao, Weizhu Chen, Jie Fu, Linjun Shou, Ming Gong, et al. 2020. GLGE: A new general language generation evaluation benchmark. *arXiv preprint arXiv:2011.11928*.

Cheng Luo, Yukun Zheng, Yiqun Liu, Xiaochuan Wang, Jingfang Xu, Min Zhang, and Shaoping Ma. 2017. SogouT-16: a new web corpus to embrace IR research. In *Proceedings of SIGIR*, pages 1233–1236.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of ACL*, pages 311–318.

Xipeng Qiu, Peng Qian, and Zhan Shi. 2016. Overview of the NLPCC-ICCPOL 2016 shared task: Chinese word segmentation for micro-blog texts. pages 901–906.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *JMLR*, 21:1–67.

Zhihong Shao, Minlie Huang, Jiangtao Wen, Wenfei Xu, and Xiaoyan Zhu. 2019. Long and diverse text generation with planning-based hierarchical variational model. In *Proceedings of EMNLP-IJCNLP*, pages 3257–3268.

Kai Sun, Dian Yu, Dong Yu, and Claire Cardie. 2019. Investigating prior knowledge for challenging chinese machine reading comprehension.

Hongye Tan, Xiaoyue Wang, Yu Ji, Ru Li, Xiaoli Li, Zhiwei Hu, Yunxiao Zhao, and Xiaoqi Han. 2021. GCRC: A new challenging MRC dataset from Gaokao Chinese for explainable evaluation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1319–1330.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2019. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. *arXiv preprint arXiv:1905.00537*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355.

Yan Wang, Xiaojiang Liu, and Shuming Shi. 2017. Deep neural solver for math word problems. In *Proceedings of EMNLP*, pages 845–854.

Yingying Wang, Cunliang Kong, Liner Yang, Yijun Wang, Xiaorong Lu, Renfen Hu, Shan He, Zhenghao Liu, Yun Chen, Erhong Yang, et al. 2021. YACLC: A chinese learner corpus with multidimensional annotation. *arXiv preprint arXiv:2112.15043*.

Liang Xu, Hai Hu, Xuanwei Zhang, Lu Li, Chenjie Cao, Yudong Li, Yechen Xu, Kai Sun, Dian Yu, Cong Yu, et al. 2020. CLUE: A chinese language understanding evaluation benchmark. In *Proceedings of COLING*, pages 4762–4772.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mT5: A massively multilingual pre-trained text-to-text transformer.

Peng Yuan, Haoran Li, Song Xu, Youzheng Wu, Xiaodong He, and Bowen Zhou. 2020. On the faithfulness for E-commerce product summarization. In *Proceedings of COLING*, pages 5712–5717.

Zhengyan Zhang, Yuxian Gu, Xu Han, Shengqi Chen, Chaojun Xiao, Zhenbo Sun, Yuan Yao, Fanchao Qi, Jian Guan, Pei Ke, et al. 2021. CPM-2: Large-scale cost-effective pre-trained language models. *arXiv preprint arXiv:2106.10715*.

Lei Zheng, Fanchao Qi, Zhiyuan Liu, Yasheng Wang, Qun Liu, and Maosong Sun. 2020. Multi-channel reverse dictionary model. In *Proceedings of AAAI*, volume 34, pages 312–319.

Hao Zhou, Chujie Zheng, Kaili Huang, Minlie Huang, and Xiaoyan Zhu. 2020. KdConv: A chinese multi-domain dialogue dataset towards multi-turn knowledge-driven conversation. In *Proceedings of ACL*, pages 7098–7108.

Junnan Zhu, Qian Wang, Yining Wang, Yu Zhou, Jiajun Zhang, Shaonan Wang, and Chengqing Zong. 2019. NCLS: Neural cross-lingual summarization. In *Proceedings of EMNLP-IJCNLP*, pages 3054–3064.

A Contributions

Yuan Yao, Qingxiu Dong, Jian Guan and Boxi Cao built the benchmark and wrote the report. Yuan Yao wrote Section 1, 4, 5, 6 and 7; Qingxiu wrote Section 2; Jian Guan wrote Section 3.3-3.7; BoxiCao wrote Section 3.1-3.2. Yuan Yao and Zhiyuan Liu proofread the report.

Yuan Yao, Chengqiang Xu, Huadong Wang and Guoyang Zeng led the platform design and implementation.

Yuan Yao, Qingxiu Dong, Jian Guan, Boxi Cao, Xiaozhi Wang, Fanchao Qi, Junwei Bao, Jinran Nie, Zheni Zeng, Xuancheng Huang, Kun Zhou, Wenhao Li, Shuhuai Ren, Jinliang Lu and Zile Zhou reviewed, selected and organized the datasets.

Zhengyan Zhang, Chaojun Xiao and Yuxian Gu conducted the experiments.

Zhiyuan Liu, Xianpei Han, Erhong Yang, Zhifang Sui and Maosong Sun designed and led the research.

Jiajun Zhang, Juanzi Li, Minlie Huang, Rui Yan, Xiaodong He, Xiaojun Wan, Xin Zhao, Xu Sun and Yang Liu participated in the discussions of the research and provided valuable suggestions.