Report on Mandarin and Cantonese PIQA

Anonymous EMNLP submission

1 Introduction

In this short paper, we document the manual construction of a PIQA dataset in Mandarin Chinese and Cantonese, and report the performance of seven Qwen2.5-Base models on the dataset. Six native speakers created a total of 409 goal-solution pairs in Mandarin and 225 pairs for Cantonese. Examples of the pairs are shown in Table 1 and Table 2.

2 Data Construction

Entries in our dataset are either elicited from a source (Elicited) or created by the annotator (Created).

Elicited pairs Entries are based on some modification of online resources or guidebooks. Online resources include:

1. **Online encyclopedia in Mandarin and Cantonese:** Wikipedia[1], wikiHow[2] and Fandom[3].

2. **Guidebook:** guidebooks on outdoor survival, furniture assembly, and cooking.

For instance, if a source indicates that to cook a dish, the pork must be blanched in boiling water, then the goal would be: "to blanch the pork for the dish", the correct solution: "one should use boiling water", and the incorrect solution: "one should use cold water".

Created pairs Entries of this type are created from real-life experiences, observations and the interests of our annotators. For instance, one entry would be, "When doing push-ups the face of the person" (goal), "faces down" (correct solution), or "faces up" (incorrect solution).

The Chinese PIQA subset was created by four native speakers of Mandarin Chinese, while the Cantonese subset was developed by two native speakers of Cantonese. Each annotator produced at least 100 valid goal-solution pairs, either elicited or created.

To ensure data validity, peer review was conducted within each subset: all goal-solution pairs created by annotator X were reviewed by another annotator Y, and problematic pairs were discussed and then either corrected or removed.

3 The Resulting Dataset

Similar to the original PIQA dataset, our dataset consists of question and solution pairs. Both the correct and incorrect solutions minimally differ in terms of technical details (e.g., materials and required quantity). Questions and solutions in our dataset vary in length, ensuring the necessary difficulty for LLMs.

We constructed question and solution pairs from one general category and four domain-specific categories for both Chinese and Cantonese subsets:

1. **General** includes solutions for general life tips, instructions, textbooks, and related to physical properties.

2. **Activity** includes guides and knowledge for sports activities, like outdoor activities and strength training.

3. **Food** involves recipe details and physical underpinnings, as well as general properties about foods.

4. **Geography** involves location distribution, characteristic landscapes, and transportation.

[1]For Cantonese, see https://zh-yue.wikipedia.org

[2]For Mandarin Chinese and Cantonese, see https://zh.wikihow.com

[3]For Cantonese, see https://www.fandom.com

lang	category	goal	sol_0	sol_1	count
cmn	activity	可以使用若干根木棍，一块布和一些绳子制作一顶简单的圆锥形帐篷。(You can use several wooden sticks, a piece of cloth and some rope to make a simple conical tent.)	在搭建帐篷时，我们最好将木棍切成大致相同的长度。(When setting up the tent, we'd better cut the wooden sticks to roughly the same length.)	在搭建帐篷时，我们最好将木棍切成完全不同的长度。(When setting up the tent, we'd better cut the wooden sticks to completely different lengths.)	54
cmn	art	口琴演奏时，演奏人 (When playing the harmonica, the performer)	可以同时讲话 (can speak at the same time)	不能同时讲话 (cannot speak at the same time)	43
cmn	food	在炖一碗猪蹄莲藕汤时，汤里的莲藕和猪蹄，更容易沉底的是 (When stewing a bowl of pig's trotters and lotus root soup, between the lotus root and pig's trotters in the soup, the one more likely to sink to the bottom is)	莲藕 (lotus root)	猪蹄 (pig's trotters)	55
cmn	general	一件小棉袄比一只大手套更 (A small cotton jacket is more _____ than a large glove)	大 (big)	小 (small)	222
cmn	geography	汽车面向正南方向倒车时往右打方向盘，则汽车 (When a car is facing south and reversing, if one turns the steering wheel to the right, the car)	向东行 (moves eastward)	向西行 (moves westward)	35

Table 1: Examples randomly sampled from our Mandarin PIQA (lang = cmn). Solution 0 is correct for all except the second and third question.

5. **Art** focuses on culturally specific architectures, gardening, artifacts, and their related knowledge.

4 Experimental Setup and Results

We tested the performance of Qwen2.5-Base series on our dataset, with results shown in Table 3.[4] We observe increasing accuracy from 60.41% (0.5B) to 75.08% (72B), suggesting that our dataset can discriminate models of different sizes and is challenging even for models of very large size.

[4]All models are quantized into 4-bits for evaluation.

lang	category	goal	sol_0	sol_1	count
yue	activity	低脂嘅食物係唔係比高脂食物健康啲？(Are low-fat foods healthier than high-fat foods?)	唔係，低脂食品可能含有大量嘅鹽、新增劑或代糖；代糖可能會影響人體代謝，增加內臟嘅脂肪，反而唔利於減肥 (No, low-fat foods may contain large amounts of salt, additives or artificial sweeteners; artificial sweeteners may affect human metabolism, increase visceral fat, and actually be detrimental to weight loss)	唔係，低脂食品可能含有大量嘅鹽、糖或新增劑；新增劑可能會縮短食品賞味期限，增加內臟嘅脂肪，反而唔利於減肥 (No, low-fat foods may contain large amounts of salt, sugar or additives; additives may shorten the shelf life of food, increase visceral fat, and actually be detrimental to weight loss)	17
yue	art	演員透過象徵性嘅姿態同埋動作，交代出劇中人物嘅性格。小生係男性角色，佢嘅臺步係 (Actors convey the character's personality through symbolic gestures and movements. The xiaosheng is a male role, and his stage walk is)	丁字步，要表現氣宇軒昂 (T-shaped step, to show dignity and bearing)	撇步，要表現輕盈 (diagonal step, to show lightness)	13
yue	food	粵菜依然會用乾貨同醃過嘅嘢食，一啲廚師會結合用新鮮材料同乾貨。乾貨通常會 (Cantonese cuisine still uses dried goods and pickled foods, some chefs combine fresh ingredients with dried goods. Dried goods are usually)	先浸喺水度然後先煮，或者會畀人煮一段好長嘅時間。(first soaked in water then cooked, or cooked for a very long time.)	先煮然後先浸喺水度，或者會畀人煮一段好長嘅時間。(first cooked then soaked in water, or cooked for a very long time.)	63
yue	general	保溫杯瓶膽入面嘅聲響可以分辨出瓶膽好唔好，聲越大，質素越 (The sound inside the thermos bottle's inner chamber can distinguish whether the chamber is good or not. The louder the sound, the quality is)	好 (better)	差 (worse)	89
yue	geography	馬鞍山綫曾為港鐵系統的一條路綫，於 2004 年 12 月 21 日通車，來往烏溪沙及大圍。後來佢 (The Ma On Shan Line was once a line in the MTR system, opened on December 21, 2004, running between Wu Kai Sha and Tai Wai. Later it)	因屯馬綫全綫通車而併入屯馬綫 (was merged into the Tuen Ma Line due to the full opening of the Tuen Ma Line)	因大圍至啟德段通車而併入屯馬綫 (was merged into the Tuen Ma Line due to the opening of the Tai Wai to Kai Tak section)	43

Table 2: Examples randomly sampled from our Cantonese PIQA (lang = yue). Solution 0 is correct for all these examples.

			Accuracy (%)						
Language	Category	Count	0.5B	1.5B	3B	7B	14B	32B	72B
cmn	general	222	59.01	62.16	63.96	66.22	65.77	68.02	70.72
	activity	54	68.52	72.22	75.93	81.48	81.48	85.19	83.33
	food	55	58.18	56.36	65.45	63.64	65.45	72.73	69.09
	geography	35	48.57	51.43	48.57	57.14	60.0	68.57	68.57
	art	43	65.12	62.79	60.47	72.09	74.42	86.05	76.74
	(total)	409	59.9	61.86	64.06	67.73	68.22	72.86	72.62
yue	general	89	65.17	71.91	71.91	70.79	73.03	74.16	73.03
	activity	17	58.82	70.59	88.24	88.24	94.12	82.35	100.0
	food	63	57.14	60.32	68.25	74.6	79.37	79.37	77.78
	geography	43	55.81	65.12	62.79	60.47	72.09	90.7	83.72
	art	13	76.92	69.23	69.23	84.62	76.92	84.62	92.31
	(total)	225	61.33	67.11	70.22	72.0	76.44	80.0	79.56
(all)	—	634	60.41	63.72	66.25	69.24	71.14	75.39	75.08

Table 3: Accuracy of Qwen2.5 Models (0.5B to 72B Parameters) on our dataset.