

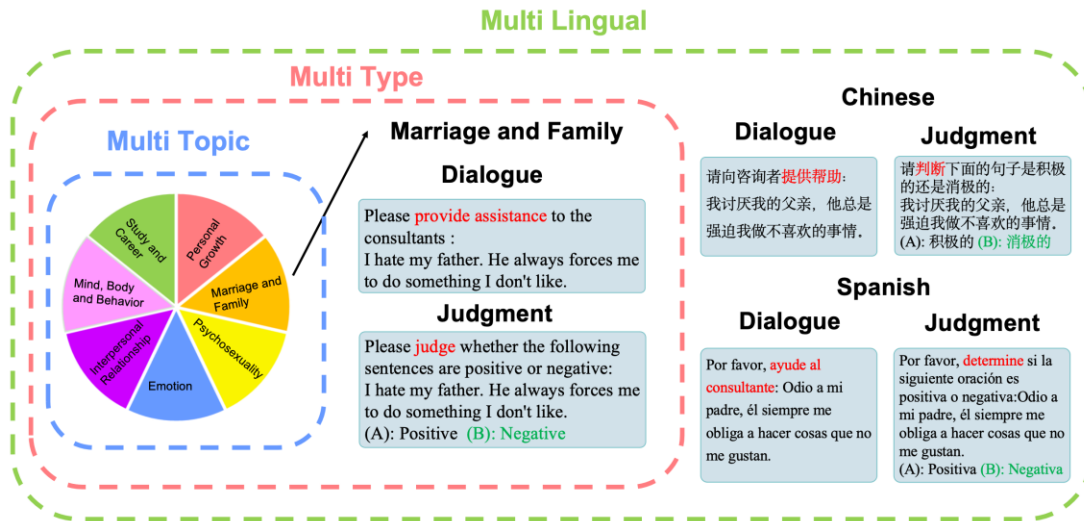
PesTest: A Comprehensive Benchmark for Psychological Emotional Support Capability of Large Language Models

**Tianwei LAN, Yuhang GUO, Siyuan GAO, Hongfei XIA,
Zeming LIU, Heyan HUANG**

Frontiers of Computer Science, DOI: [10.1007/s11704-026-50643-w](https://doi.org/10.1007/s11704-026-50643-w)

Problems & Ideas

- Problems of conventional evaluation approaches for LLMs' psychological emotional support capability:
 - Conventional evaluations adopt single task modalities, monolingual approaches and restricted topical domains.
 - These limitations lead to misleading assessments with models performing well on benchmarks but acting ineffectively in practice.
- Ideas: Propose PesTest, a multidimensional benchmark with multi-task, multilingual and extensive topical coverage for comprehensive evaluation and model capability enhancement.



This figure outlines PesTest, the LLM psychological support benchmark—by displaying its core Multi-Topic (7 expert-curated domains like “Marriage and Family”), Multi-Task (Dialogue for support interactions, Judgment for emotional understanding), and Multi-Lingual (Chinese, English, Spanish) features, paired with sample task instances across languages to show its comprehensive evaluation design.

Main Contributions

- Contributions:
 - Identify the key challenge in evaluating LLMs' psychological emotional support capabilities: the need for a comprehensive assessment across multiple task types, languages, and mental health topics.
 - A comprehensive benchmark for evaluating psychological emotional support capabilities of LLMs, featuring multiple task types, multilingual support, and a professionally curated taxonomy of mental health topics.
 - Our comprehensive evaluation using the PesTest assesses 12 leading LLMs, revealing consistent shortcomings in dialogue-based support tasks and notable performance disparities across different psychological topics.

Model	Interpersonal Relationship	Psychosexuality	Marriage & Family	Personal Growth	Study & Career	Emotion	Mind, Body & Behavior	All
Llama2	86.4	47.5	48.3	44.9	48.4	80.8	75.7	74.1
Llama3-8B	85.2	81.3	70.4	83.7	82.9	82.0	86.7	83.9
Qwen	65.4	89.8	72.8	96.6	97.0	61.4	70.8	70.4
Qwen2-0.5B	54.8	45.7	46.4	64.6	68.7	47.0	56.4	54.1
Qwen2-1.5B	67.3	84.8	67.0	94.6	97.5	67.1	72.0	72.0
Qwen2-7B	81.6	88.1	73.7	94.6	95.7	73.5	63.8	73.3
Qwen3-8B	91.9	93.2	79.4	98.7	99.4	86.7	83.4	87.1
Baichuan2	61.8	59.3	64.1	64.7	81.6	61.2	43.1	55.1
ChatGLM3	95.9	86.4	73.2	95.2	95.7	81.6	84.2	86.3
GPT-3.5-Turbo	92.2	86.5	72.8	96.6	96.3	87.7	83.6	86.7
GPT-4o	93.5	88.1	77.1	95.9	97.0	88.5	84.6	87.9
DeepSeek	91.2	88.1	79.4	93.9	95.7	91.4	87.9	89.4

Model	Metrics	Interpersonal Relationship	Psychosexuality	Marriage & Family	Personal Growth	Study & Career	Emotion	Mind, Body & Behavior	All
Llama2	BLEU	8.47	8.23	8.46	8.60	8.53	8.51	8.29	8.45
	D-1	41.6	44.8	41.2	45.1	43.0	42.7	43.9	42.3
	SemanticSim	43.8	40.1	42.7	41.4	42.0	42.5	42.2	42.7
Llama3-8B	BLEU	8.74	8.24	8.71	8.49	8.20	8.67	8.42	8.63
	D-1	36.2	42.0	37.1	39.4	36.8	38.0	40.7	37.8
	SemanticSim	50.0	44.4	47.7	45.6	47.7	48.5	47.4	48.3
Qwen	BLEU	8.85	8.46	8.83	8.57	8.50	8.76	8.54	8.74
	D-1	33.5	38.9	31.1	36.2	34.8	36.3	39.5	34.8
	SemanticSim	42.3	40.4	40.4	39.7	39.5	41.8	41.8	41.3
Qwen2-0.5B	BLEU	8.66	8.51	8.71	8.81	8.58	8.77	8.62	8.70
	D-1	37.3	39.5	37.6	38.6	37.3	38.1	40.2	38.1
	SemanticSim	43.4	44.0	42.4	42.6	43.0	44.7	47.1	44.0
Qwen2-1.5B	BLEU	8.61	8.45	8.59	8.43	8.49	8.71	8.46	8.60
	D-1	36.8	41.7	37.2	39.5	38.7	39.0	41.3	38.4
	SemanticSim	45.1	41.6	43.1	44.1	45.0	44.3	43.3	43.9
Qwen2-7B	BLEU	8.87	8.49	8.89	8.76	8.69	8.83	8.63	8.81
	D-1	37.3	41.2	37.7	40.7	38.8	38.6	42.1	38.7
	SemanticSim	52.3	45.6	51.0	47.4	50.2	51.5	49.8	51.0
Qwen3-8B	BLEU	8.88	8.66	8.98	8.77	8.84	8.89	8.66	8.87
	D-1	38.0	42.2	39.3	40.1	39.0	39.3	40.8	39.3
	SemanticSim	61.4	54.6	61.1	55.1	59.2	59.9	59.8	60.2
Baichuan2	BLEU	8.27	7.85	8.27	8.03	8.03	8.28	8.04	8.22
	D-1	33.2	36.2	33.3	34.1	33.3	34.4	36.4	34.1
	SemanticSim	46.8	43.0	45.7	45.5	44.3	46.2	44.7	45.9
ChatGLM3	BLEU	8.56	7.96	8.35	8.36	8.10	8.37	8.21	8.37
	D-1	33.9	38.2	34.1	36.8	34.8	35.4	38.7	35.2
	SemanticSim	45.4	42.0	44.2	44.2	45.6	45.6	45.7	45.1
GPT-3.5-Turbo	BLEU	9.01	8.45	9.03	8.74	8.72	8.91	8.87	8.94
	D-1	37.6	42.1	37.7	40.0	38.1	38.7	42.1	38.7
	SemanticSim	59.9	57.3	64.1	57.9	62.5	61.2	62.5	61.7
GPT-4o	BLEU	8.40	8.14	8.41	8.16	8.14	8.41	8.11	8.34
	D-1	34.7	38.8	34.9	36.9	35.1	38.3	37.5	36.3
	SemanticSim	63.7	57.3	64.1	57.8	62.4	62.3	62.5	62.9
DeepSeek	BLEU	8.45	8.37	8.89	8.54	8.71	8.98	8.67	8.75
	D-1	40.5	39.7	38.1	38.5	40.1	41.2	39.4	39.8
	SemanticSim	61.7	58.2	62.9	58.1	61.8	63.3	61.6	62.3

Table 4 The models' overall score and sub-topic scores on the Psychological Support Dialogue Task.

The models' overall score and sub-topic scores on the Emotional Tendency Judgment Task and Psychological Support Dialogue Task