An Evaluation Framework for Emotional Companionship Capability of Dialogue Systems

Anonymous ACL submission

Abstract

With the rapid development of Large Language Models, dialogue systems are shifting from information tools to emotional companions, heralding the era of Emotional Companionship Dialogue Systems (ECDs) that provide personalized emotional support for users. However, the field lacks systematic evaluation standards. To address this, we pioneered the design and implementation of the Four-Dimensional Capability Evaluation Framework (FDAEF), which hierarchically integrates "Capability Layer → Task Layer (three-level) → Data Layer → Method Layer". Then, we present **ECDBench 1.0**, the inaugural ECD-specific benchmark developed under FDAEF. Through extensive evaluations of 30 mainstream models, we demonstrate that ECDBench 1.0 has excellent discriminant validity and can effectively quantify the differences in emotional companionship capabilities among models. Furthermore, the results reveal current models' shortcomings in deep emotional companionship, guiding future technological optimization and significantly aiding developers in enhancing ECDs' user experience.

1 Introduction

Recent advances in Large Language Models (LLMs) have shifted dialogue systems beyond their traditional roles in information retrieval and task execution, ushering in a new paradigm in human-computer interaction: the **Emotional Companionship Dialogue System (ECD)**. This rapid expansion of capability boundaries renders existing evaluation paradigms for traditional dialogue systems inadequate.

Existing evaluation paradigms are often constrained by a narrow scope, primarily focusing on fragmented tasks such as emotion recognition(Demszky et al., 2020b; Welivita et al., 2021; Sabour et al., 2024) or empathetic expression(Liu et al., 2021; Sabour et al., 2024). However, emotional companionship represents a sophisticated and multi-faceted interaction capability that transcends these isolated functions. Consequently, current frameworks fail to encapsulate the cohesive chain of capabilities essential for ECDs—ranging from emotional perception and understanding to empathetic responding, memory management, and personalization. This methodological gap precludes a holistic assessment of system effectiveness, necessitating the construction of a systematic benchmark to scientifically evaluate the multi-faceted quality of ECD capabilities.

To address these issues, this paper focuses on **the construction of an evaluation Benchmark**. We make two core contributions: First, we pioneered the design and implementation of the Four-Dimensional Capability Evaluation Framework (FDAEF), which hierarchically integrates "Capability Layer → Task Layer (three-level) → Data Layer → Method Layer". With "Comprehensive Capability Dimensions" as top-level indicators, the framework deconstructs capabilities into granular three-level tasks, achieving precise capability-to-task alignment. This approach effectively mitigates the long-standing issue of "ambiguous capability-task matching" prevalent in traditional evaluation frameworks. Second, we present **ECDBench 1.0**, the inaugural ECD-specific benchmark developed under FDAEF. This benchmark bridges the void of specialized standards, offering a core reference for evaluating ECD models in a scientific and reproducible manner.

To validate the effectiveness of ECDBench 1.0, we conducted large-scale experiments on 30 mainstream dialogue models. The results demonstrate that ECDBench 1.0 possesses excellent discriminant validity, effectively revealing performance disparities among models. More importantly, the evaluation results precisely pinpoint the prevalent shortcomings of current models in dimensions such

as companionship capability. This provides developers with a clear roadmap for translating evaluation metrics into actionable optimization strategies, ultimately helping to enhance the overall ECD user experience.

2 Related Work

Organized around the five distinct developmental stages of dialogue systems, this section provides a systematic analysis of the co-evolutionary relationship between emotional interaction mechanisms and their corresponding evaluation paradigms.

2.1 The Early Stage (1960s–early 2010s)

This period is primarily divided into two stages: the Rule-based stage (1960s–1990s) and the Statistical Language Model (SLM) stage (1990s–early 2010s). During the rule-based era, dialogue systems were primarily driven by symbolic methodologies; consequently, standardized or quantifiable tasks for emotional evaluation had not yet been established.

Following the 1990s, the expansion of text data led to the emergence of statistical retrieval-based systems, marking a preliminary transition from "rule-driven" to "data-driven" paradigms. During this period, emotional evaluation began to emerge; however, it remained subsidiary to general text classification and was largely confined to simple sentiment polarity determination.

2.2 Neural language models(2013-2017)

With the rise of deep learning, neural language models based on the Encoder-Decoder (Seq2Seq) architecture enabled end-to-end dialogue system(Vinyals and Le, 2015). During this stage, the mainstream emotional evaluation task centered on emotion classification, with assessments becoming increasingly granular. Meanwhile, researchers began employing metrics like BLEU(Papineni et al., 2002) and ROUGE(Lin and Hovy, 2003) to provide preliminary quantitative assessments of end-to-end generated responses(Zhou et al., 2018). Despite the early conceptualization of "relational agents"(Bickmore and Picard, 2005; Bickmore et al., 2005), a standardized evaluation framework specifically targeting companionship capabilities had yet to be established.

2.3 Pre-trained Language Model(2017-2020)

The emergence of Pre-trained Language Models (PLMs) catalyzed a paradigm shift in dialogue systems, facilitating the profound integration of task-oriented and open-domain conversations(Wang et al., 2023). This period witnessed a proliferation of emotional evaluation sub-tasks. The research focus evolved from an early emphasis on "emotional understanding"(Socher et al., 2013; Poria et al., 2019) toward "emotional expression" in the later stages(Rashkin et al., 2019). Liu et al.(Liu et al., 2016) noted that traditional metrics such as BLEU are largely ineffective for evaluating emotional dialogue; this observation directly prompted the adoption of "empathy, specificity, and relevance" as the core criteria for emotional responses. Notably, the research community began establishing "Digital Persona Consistency" as an explicit evaluation metric(Zhang et al., 2018a), signaling a preliminary emphasis on long-term companionship.

2.4 Large Language Model(2020-Present)

LLMs have redefined the paradigm of dialogue systems, exhibiting empathetic capabilities that rival or even exceed human performance on specific benchmarks(Bai et al., 2025; Lee et al., 2024). This era marks a pivotal shift in emotional understanding tasks from mere classification to causal attribution, where evaluation emphasizes the underlying logical reasoning over simple label identification(Demszky et al., 2020a; Poria et al., 2021). To mirror these advanced capabilities, the evaluative focus has transitioned from fragmented sub-tasks toward holistic benchmarks—notably EQ-Bench(Paech, 2023) and EmoBench(Sabour et al., 2024)—which scrutinize psychological insight and social reasoning. Methodologically, this transition has catalyzed the "LLM-as-a-Judge" paradigm, enabling scalable assessment(Zheng et al., 2023). Furthermore, contemporary frameworks now integrate cross-cultural nuances(Karinshak et al., 2024) and ethical "redlines"(Archiwaranguprok et al., 2025) to ensure that highly empathetic companionship remains anchored in human societal values.

2.5 Agent(2022-Present)

Since the 2022 introduction of the ReAct framework(Yao et al., 2022), dialogue systems have entered the "Agent Era," enabling proactive task execution and fostering systems like Emotional SupportAgent(Xu et al., 2025). This evolution has refocused research on AI's longitudinal impact on human psychological health(Zhang et al., 2025; Ng et al., 2025; Liu et al., 2024), necessitating evaluation frameworks centered on long-term relationship maintenance. Current benchmarks now prioritize

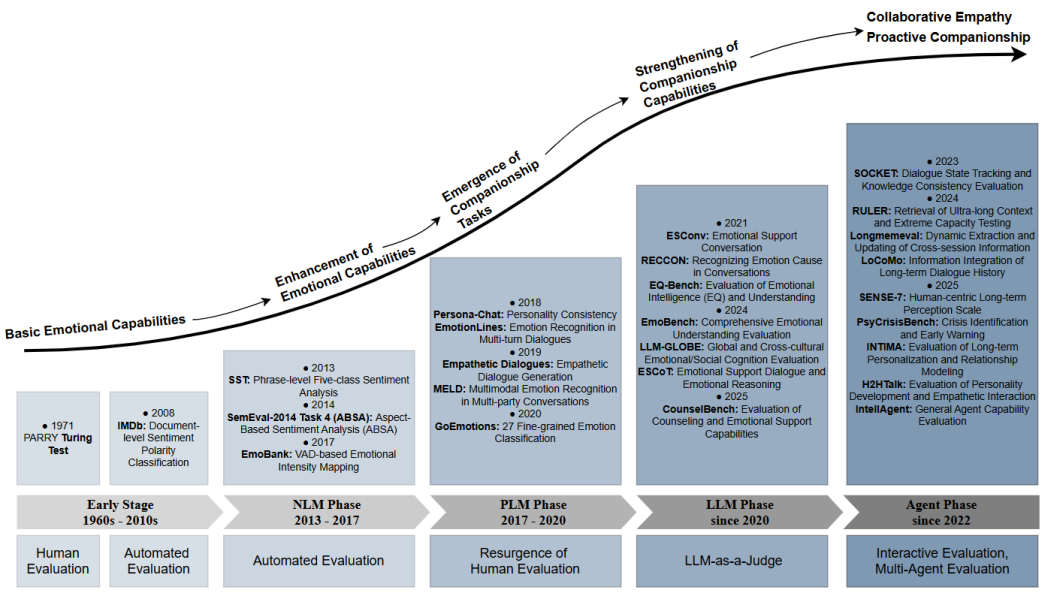

Figure 1: Evolution Diagram of Emotional Companionship Capability Evaluation

memory and information integration in extended contexts (e.g., RULER(Hsieh et al., 2024), Longmemeval(Cheng et al., 2024), LoCoMo(Maharana et al., 2024)) and emotional responsiveness (e.g., INTIMA(Kaffee et al., 2025)). Furthermore, the emergence of "sandbox simulations," such as IntellAgent(Levi and Kadar, 2025), allows for the holistic assessment of agent socialization and strategic robustness within dynamic, multi-agent environments.

The evaluation for capability of emotional companionship has advanced alongside the technological evolution of dialogue systems. In the early stages, evaluation primarily focused on simple emotion classification. During the era of NLMs, the focus began to shift from classification tasks toward emotion understanding. The pre-training paradigm then saw increased attention on tasks such as empathetic dialogue generation. With the advent of LLMs and AI Agents, evaluation has evolved from isolated emotional companionship sub-tasks toward holistic benchmarks. These benchmarks emphasize long-term relationship maintenance, ethics, safety, and emotional companionship within real-world social contexts. This shift further directs the research focus toward the essence of long-term bonds and companionship itself, signifying that "emotional companionship" has emerged as a distinct, sophisticated, and integrated interaction capability, as illustrated in Figure 1.

Meanwhile, this developmental trajectory underscores that the evaluation of emotional companionship has evolved incrementally, resulting in a fragmented landscape of assessment tasks. Given that emotional companionship capability is a complex and multifaceted capability, a holistic evaluation benchmark is essential to provide a unified and systematic assessment.

3 Design and Implementation of FDAEF

To address the core question of "how to scientifically evaluate emotional companionship capabilities," this paper first conducts a systematic review of existing Large Language Model (LLM) evaluation theories and methods. Specifically, we examine the design logic and core components for more than 40 evaluation frameworks and benchmarks(See Appendix 7.1 for details), while deeply deconstructing the specific tasks, datasets, and methodologies within each benchmark. Subsequently, drawing inspiration from the stratified assessment philosophy of standardized human testing, we propose a universal four-layer evaluation framework, FDAEF, as illustrated in Figure 2.

The development of FDAEF was a two-step process, consisting of **top-down capability decomposition** and **bottom-up score aggregation**. This section details the design and core logic of these two processes. Finally, the full implementation of FDAEF will be open-sourced upon acceptance.

3.1 Capability Decomposition Process

Following the logic of our four-layer framework, our decomposition process consists of four progres-

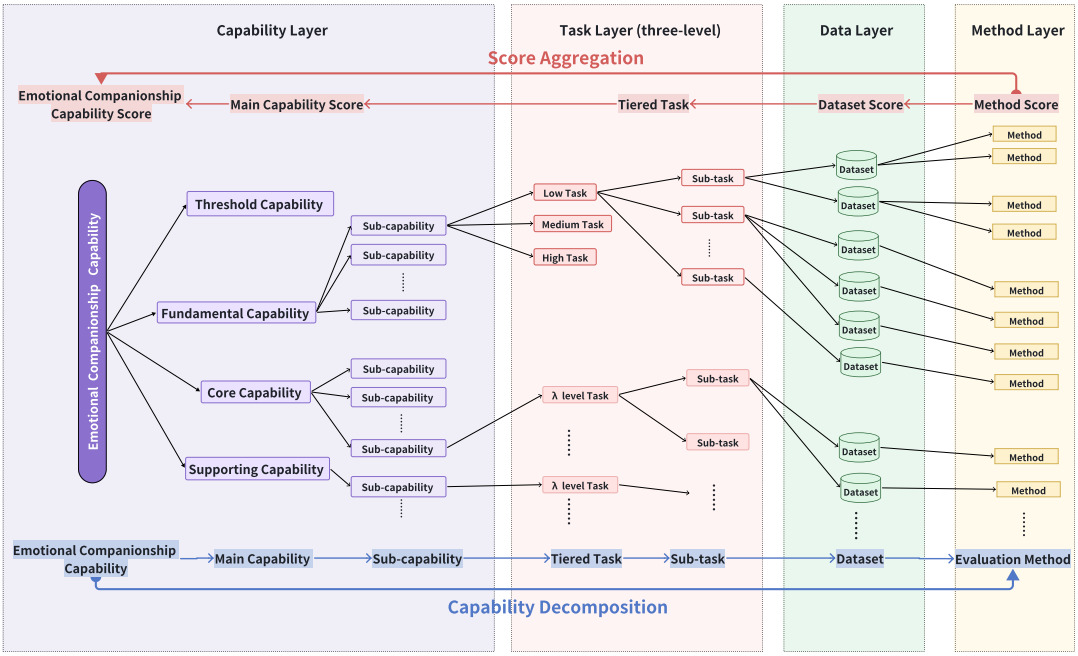

Figure 2: The Four-Layer Evaluation Framework of FDAEF

sive stages: **Capability Layer Definition, Task Layer Mapping, Data Layer Construction**, and **Method Layer Adaptation**.

Our first stage, **Capability Layer Definition**. We involve breaking down capabilities to be evaluated into four primary categories: **Threshold, Foundational, Core,** and **Supporting capabilities** based on the theory of Ervin Laszlo(Laszlo, 1972).

The **Threshold Capability** serves as the cornerstone and safety baseline for the user experience, acting as a **"one-vote veto"** that corresponds to the user need for **"Values & Safety"**. The **Foundational Capability** refers to the set of universal capabilities essential for a system to achieve its basic objectives, serving as the bedrock for the realization of all other types of capabilities. The **Core Capability** refers to a collection of high-value, non-substitutable capabilities developed by a system over long-term evolution. They serve as the fundamental pillar for achieving sustainable competitive advantage and fulfilling strategic objectives. The **Supporting Capability** refers to a collection of auxiliary functions that serve as a robust guarantee for the synergistic coordination between core and foundational capabilities. These capabilities may evolve into foundational or core competencies during particular phases of growth.

To avoid excessive decomposition, FDAEF restricts the capability layer to a four-tier structure: Overall Capability → Capability Dimensions (consisting of four dimensions) → Sub-capabilities → Leaf Capabilities.

In the **Task Layer Mapping** stage, we design specific evaluation tasks for each capability, establishing a "capability-task" mapping. To more finely delineate the models' capability boundaries, we adapt the concept of difficulty grading from human examinations. This approach allows us to classify the evaluation tasks for each sub-capability into three levels: **low, medium,** and **high**. This classification is the basis for the "three-level" designation of the Task Layer. Low-level tasks assess basic recognition and simple application; medium-level tasks evaluate comprehensive processing in complex scenarios; and high-level tasks challenge the limits of the sub-capability, examining its performance in unconventional, highly difficult, or open-ended scenarios.

For the **Data Layer Construction** stage, we select or construct 2-4 evaluation datasets for each task, typically including at least one in Chinese and one in English. We adhere to a "reuse-first" principle, prioritizing established, academically validated datasets. For tasks lacking existing data, research teams can independently construct the corresponding datasets. To ensure practical testing times, we sample 300 sets of data from each dataset for validation; if a dataset has fewer than 300 entries, its full size is used.

In the final stage, **Method Layer Adaptation**, we select or design specific evaluation methods for each dataset. A single dataset may be assessed by multiple methods. Current evaluation paradigms primarily fall into three categories: benchmark-based, model-based, and human-centric.

3.2 Score Aggregation Process

As illustrated in Figure 2, the score aggregation process combines the results from various evaluation methods into a single, final score for capability to be evaluated. This is achieved through a four-step process: **dataset score aggregation, task score aggregation, capability score synthesis,** and **final score calculation**. The detailed algorithm for this process is provided in Appendix 7.2.

4 Design and Implementation of ECDBench 1.0

To validate the FDAEF framework, we designed and implemented **ECDBench 1.0**, the first benchmark specifically for ECDs.

4.1 Design principles of ECDBench 1.0

Following the design principles of FDAEF, the design of ECDBench 1.0 involves four progressive stages: the Capability Layer Definition stage, the Task Layer Mapping stage, the Data Layer Construction stage, and the Method Layer Adaptation stage. Notably, the Capability Layer Definition and Data Layer Construction stages warrant particular elaboration.

The Capability Layer Definition stage: we decompose emotional companionship capabilities into four aggregate categories: Threshold, Foundational, Core, and Supporting capabilities. Specifically, Threshold capabilities serve as the bedrock and safety baseline for the companionship experience, corresponding to "Values and Safety." Foundational capabilities provide the universal basis for interaction, encompassing Natural Language Understanding (NLU), Natural Language Generation (NLG), Natural Language Reasoning (NLR), and Common Sense. Core capabilities represent the critical components of emotional companionship, consisting of Emotional and Companionship capabilities. Finally, Supporting capabilities act as beneficial supplements, including Multimodal integration, Knowledge Acquisition, and Cross-cultural Capabilities, as illustrated in Figure 3.

The Data Layer Construction stage: during this stage, adhering to the "Reuse-First" principle,

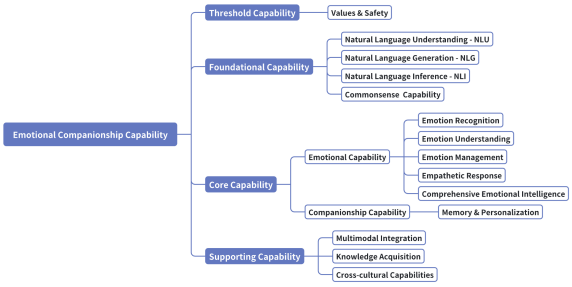

Figure 3: Decomposition of Emotional Companionship Capability

we prioritized the selection of established benchmark datasets that have been academically validated. For evaluation tasks lacking existing data support, we independently constructed the ECD-Bench dataset, comprising a total of 1,200 test samples.

4.2 Implementation and Analysis of ECDBench 1.0

Guided by the design principles of ECDBench 1.0, we have implemented this benchmark. Due to various resource constraints, supporting capabilities were not incorporated in this 1.0 version. Ultimately, the benchmark comprises 41 evaluation tasks, 60 datasets, and 60 evaluation methods (see Appendix 7.5 for details). The statistics in Table 1 highlight several areas for future optimization, specifically concerning resource distribution, dimensional balance, and language coverage.

First, **resource distribution across capabilities is imbalanced**. While Values & Safety, Foundational Capability, and Emotional Capability (e.g., emotion recognition) have a solid foundation for quantitative assessment, resources for Companionship Capability and more advanced emotional tasks are markedly limited. This constitutes a primary limitation of the current benchmark and a key focus for future work.

Second, **the dimensional ratios fall short of our target**. Our framework's target ratios are Capability:Task=1:3, Task:Dataset=1:2, and Dataset:Method=1:1. The current benchmark's ratios are too high, suggesting that even where overall resources seem adequate, coverage for specific sub-capabilities (e.g., Emotion Understanding, Emotion Management) remains insufficient and requires supplementation.

Third, **the benchmark's language coverage is inadequate for robust cross-cultural evaluation**. The current 60 datasets are predominantly Chinese

Table 1: ECDBench 1.0: Statistics on Capabilities, Tasks, Datasets, and Methods

Capability Dimension	Capability Name	# of Eval Tasks	# of Datasets	# of Eval Methods
Values & Safety	Values	2	5	5
	Safety	4	5	5
Foundational Capability	Natural Language Understanding	5	8	8
	Natural Language Inference	6	8	8
	Natural Language Generation	1	3	3
	Commonsense Capability	6	8	8
Emotional Capability	Emotion Recognition	6	10 (1)	10
	Emotion Understanding	2	3 (1)	3
	Emotion Management	1	1 (1)	1
	Empathetic Response	2	2 (1)	2
	Comprehensive Emotional Intelligence	3	4	4
Companionship Capability	Memory & Personalization	3	3	3
	Total	**41**	**60 (4)**	**60**

* Note: In the "# of Datasets" column, the number in parentheses indicates the count of self-built ECDBench datasets included in the total.

(35%) and English (58.3%). The limited representation of dedicated bilingual datasets (6.7%) severely constrains the assessment of cross-lingual capabilities. Furthermore, the complete absence of other major languages, such as Japanese or Spanish, restricts the benchmark's applicability in diverse cultural contexts, marking this as a key direction for future expansion.

5 Experiment

We evaluated 30 representative conversational models (a full list is in Appendix 7.3) using the ECDBench 1.0 benchmark to assess their emotional companionship capabilities. The final rankings and scores are detailed in Appendix C. This experiment produced three key results: 1. We validated that ECDBench 1.0 is an effective and correct benchmark for measuring emotional companionship. 2. The evaluation uncovered common patterns and trends in current ECDs. 3. It offers a precise and actionable roadmap for developers to optimize and improve ECDs.

5.1 Validation of Benchmark Effectiveness and Correctness

We confirmed the effectiveness and correctness of the ECDBench 1.0 benchmark through two dimensions of analysis:

1. Consistency with Domain Consensus. The benchmark rankings (see Appendix C) align strongly with the general consensus of mainstream models. There is a clear score gap between top-tier and general models, and members of the same model family with different scales also show expected performance differences. This demonstrates that the benchmark has excellent discriminant validity.

However, excluding two models with outlier scores, the comprehensive score difference between the first and 28th-ranked models is only 7.14 points. This relatively small gap indicates that ECDBench 1.0 has a limited number of mid-to-high difficulty tasks, which is a key area for future improvement.

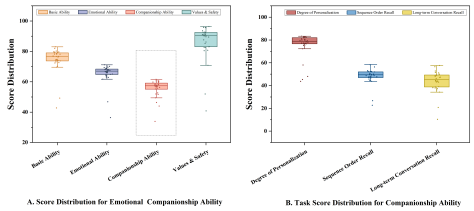

Figure 4: Score Distribution

2. High Correlation with Human Evaluation. An analysis of the models score distribution box plot (as shown in Figure 4, left) revealed three key findings: (1) scores for the Values and Safety dimension were the highest, indicating optimal performance in this area; (2) Foundational Capability achieved the second-highest score, demonstrating high performance in conversational fluency and question-answering accuracy; and (3) scores for Emotional Capability were balanced within an acceptable range, showing satisfactory overall performance. Notably, the Companionship Capability dimension was significantly weaker.

This consistency between model score characteristics and human subjective preferences indicates a

strong correlation between our benchmark's quantitative results and human evaluation, which in turn indirectly validates the benchmark's effectiveness in assessing the quality of ECD products.

5.2 patterns and trends in current ECDs

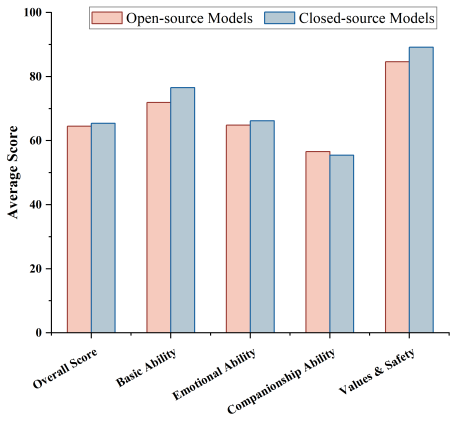

Figure 5: Average Score Comparison: Closed-source vs. Open-source Models

Based on an analysis of the experimental results, we have summarized the following key findings:

Finding 1: Closed-source models outperform open-source models in emotional companionship capabilities.

Whether viewed from the overall ranking distribution or the average scores across each dimension (as shown in Figure 5), closed-source models consistently performed better on average than open-source models. This reflects the stronger overall capabilities of the closed-source model development teams.

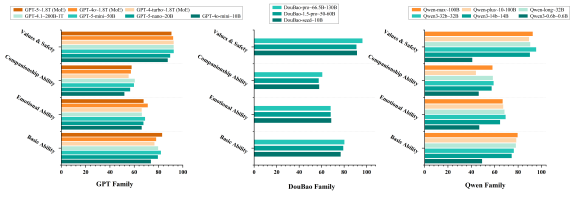

Figure 6: Emotional Companionship Capability Comparison of Models with Different Parameter Sizes within the Same Family (Doubao, GPT, Qwen)

Finding 2: The scaling law remains effective in the emotional companionship domain.

By comparing models with different parameter sizes within the Doubao, GPT, and Qwen families, we found that performance generally correlates positively with the number of parameters (as shown in Figure 6). This indicates that the scaling law continues to be effective in this specific vertical domain of emotional companionship.

Finding 3: Foundational language capabilities are the "Cornerstone" for core capabilities, but cannot directly translate into them.

The scatter plot of scores for Foundational Capabilities and Core Capabilities (as shown in Figure 7) reveals a certain degree of correlation(as illustrated by the fitted curve in Figure 7). This suggests that a solid foundation is a necessary prerequisite for developing core capabilities. However, a deeper analysis indicates that correlation does not equal direct conversion. As the plot shows, within a narrow range of similar core scores (58-65), the scores of different models' foundational capabilities have little impact on their core capability scores. This dispersed distribution strongly demonstrates that core capabilities are not a natural byproduct of foundational ones; they must be targeted and strengthened as an independent goal.

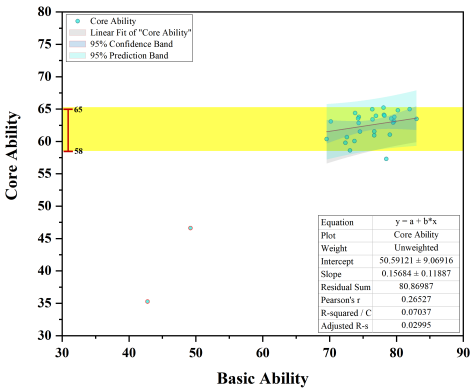

Figure 7: Scatter Plot of Scores: Core Capabilities vs. Foundational Capabilities

5.3 Pinpointing Directions for ECD Optimization

By using ECDBench 1.0's hierarchical breakdown from "capability layer→task layer (three-level)→data layer→method layer," we can drill down from a macro-level capability to a micro-level task, providing a precise and actionable guidance path for model optimization. For example, to analyze a model's current capability gaps, a developer can follow these steps:

1. Capability layer identification: By analyzing the score distribution box plot for all tested models (Figure 4, left), we found that "Companion-

ship Capability" is a universal weakness across all capability dimensions.

2. Task layer identification: Within the "Companionship Capability" task layer, "Long-term Dialogue Recall" has the lowest score and is a core bottleneck dragging down overall performance (Figure 4, right).

3. Data and method layer identification: This specific task corresponds to the LongMemEval dataset, meaning the problem lies in the model's low score on this dataset.

4. Cause analysis: The LongMemEval dataset requires a model to process long-form, cross-session, and dynamic information. However, the Transformer's attention mechanism struggles to handle this type of information.

5. Conclusion: To improve a model's "Long-term Dialogue Recall" capability, we recommend that developers start with memory management and establish a dynamic memory management mechanism to enhance its performance in this area.

6 Conclusion

This paper pioneers the design and implementation of the FDAEF, a hierarchical architecture designed to standardize the assessment of integrated capability. Building upon this framework, we introduce ECDBench 1.0, the inaugural comprehensive benchmark specifically engineered for ECDs. Meanwhile, we created a dataset of 1,200 high-quality for specific evaluation tasks.

Through a systematic evaluation of 30 representative conversational models, we not only validate the framework's efficacy in quantifying multi-faceted emotional companionship capabilities with robust discriminant validity but also delineate actionable optimization trajectories for the field. Furthermore, while the proliferation of ECDs necessitates a cautious approach toward sociopsychological risks such as user dependency, the ECDBench1.0 serves as a vital diagnostic tool to detect and quantify these behaviors. By providing developers with essential empirical evidence, this work empowers the engineering of safer and more ethically responsible interaction paradigms.

Building on these findings, our future development of the ECDBench series will prioritize three key areas:

1. Multimodal Evaluation Expansion: We aim to transcend text-only interactions by extending our assessments to include speech and visual modalities, reflecting the multisensory nature of human companionship.

2. High-Order Emotional Complexity: Future iterations will focus on developing advanced tasks and high-quality datasets to evaluate long-term, dynamic interactions and complex emotional phenomena such as irony and mixed affective states.

3. Methodological Refinement: We will explore novel evaluation paradigms that optimize the balance between computational efficiency and objective accuracy, ensuring a more rigorous measure of a model's latent performance.

As ECDs become more prevalent, the research focus will shift to deep emotional understanding, long-term memory, and personalized responses. Ultimately, for ECDs to become genuine companions, they must fulfill the fundamental human need to **"feel seen, understood, and supported."**

Limitations

While this work establishes a systematic framework for evaluating ECDs, we acknowledge several limitations that define the scope of our current findings and suggest directions for future research.

1. Multimodal and Cross-cultural Dimensions: Our current evaluation system, ECDBench 1.0, is primarily focused on text-based interactions. However, genuine emotional companionship often relies on a synergy of multimodal signals, including vocal prosody and visual cues, which are not yet covered. Furthermore, the current framework does not fully account for the profound cross-cultural variations in empathetic expression and companionship expectations.

2. Complexity of Higher-order Emotions: The assessment of higher-order emotional phenomena—such as irony, sarcasm, and nuanced mixed emotions—remains a significant challenge. Due to the scarcity of high-quality, academically validated datasets in these specific areas, our benchmark's ability to scrutinize a model's performance in highly complex emotional scenarios is currently constrained.

3. Methodological Dependencies: A majority of the tasks in ECDBench 1.0 currently rely on benchmark-based evaluation methods. While these methods provide standardized metrics, their accuracy is inherently dependent on the quality of the gold-standard datasets and can be influenced by variations in response length or linguistic diversity. Transitioning toward more robust, model-

based evaluation paradigms that balance efficiency with objective accuracy remains a key objective for future iterations.

7 Appendix

7.1 Evaluation Framework and Benchmark

Table 2: Table of Evaluation Framework and Benchmark

Number	Name	Release Time	Publishing Entity
1	GLUE (Wang et al., 2018)	2018	New York University, the University of Washington, and DeepMind
2	SuperGLUE (Wang et al., 2019)	2019	New York University, the University of Washington, and DeepMind
3	XTREME (Hu et al., 2020)	2020	Carnegie Mellon University
4	XGLUE (Liang et al., 2020)	2020	Microsoft
5	MMLU (Hendrycks et al., 2021a)	2020	UC Berkeley
6	CLUE (Xu et al., 2020b)	2020	CLUE Team
7	DialoGLUE (Mehri et al., 2020)	2020	Shikib Mehri
8	BIG-Bench (Srivastava et al., 2023)	2021	450 authors from 132 institutions (primarily led by Google researchers)
9	GSM8K (Cobbe et al., 2021)	2021	OpenAI
10	HumanEval (Chen, 2021)	2021	OpenAI
11	TruthfulQA (Lin et al., 2022a)	2021	Stanford University
12	MATH (Hendrycks et al., 2021b)	2021	Dan Hendrycks
13	HELM (Liang et al., 2022)	2022	Stanford University
14	CUGE (Yao et al., 2021)	2023	Peking University, Tsinghua University, Beijing Academy of Artificial Intelligence
15	Evals (OpenAI)	2023	OpenAI
16	OpenCompass (Contributors, 2023)	2023	Shanghai Artificial Intelligence Laboratory, Shanghai Jiao Tong University, The Chinese University of Hong Kong, Shenzhen
17	FlagEval	2023	Beijing Academy of Artificial Intelligence
18	PromptBench(Zhu et al., 2023a)(Zhu et al., 2023b)	2023	Microsoft Research
19	DeepEval	2023	Confident AI Team
20	SEAL LLM Leaderboards (Slack et al., 2024)	2023	Scale AI
21	Open LLM Leaderboard (Myrzakhan et al., 2024)	2023	Hugging Face
22	SuperCLUE (Xu et al., 2023)	2023	CLUE Team
23	AGI Eval (Zhong et al., 2024)	2023	Microsoft
24	C-Eval (Huang et al., 2023b)	2023	Shanghai Jiao Tong University, Tsinghua University, The University of Edinburgh
25	MT-Bench (Zheng et al., 2023)	2023	UC Berkeley
26	SafetyBench (Zhang et al., 2024)	2023	COAI Team of Tsinghua University
27	S-Eval (Yuan et al., 2025)	2023	Zhejiang University, Alibaba Group
28	M3ke (Liu et al., 2023a)	2023	Chuang Liu et al.
29	MMCU (Zeng, 2023)	2023	Hui Zeng
30	SOCKET (Choi et al., 2023)	2023	Minje Choi et al.
31	AGENTBENCH (Liu et al., 2023c)	2023	Xiao Liu et al.
32	API-Bank (Li et al., 2023a)	2023	Minghao Li et al.
33	clembench (Chalamalasetti et al., 2023)	2023	Kranti Chalamalasetti et al.
34	TrustGPT (Huang et al., 2023a)	2023	Yue Huang et al.
35	CLEVA (Li et al., 2023b)	2023	The Chinese University of Hong Kong, Shanghai Artificial Intelligence Laboratory
36	Evalscope (Team, 2024)	2024	Carnegie Mellon University
37	CLiB (ReLE) (ReLE Benchmark Team, 2025)	2024	Xidian University, Zhejiang University
38	CEB (Wang et al., 2024)	2024	Song Wang et al.
39	Chatbot Arena (Chiang et al., 2024)	2024	LMSYS Org
40	CMMLU (Li et al., 2024a)	2024	Haonan Li et al.
41	DialogueBench (Ou et al., 2024)	2024	Jiao Ou et al.
42	Xiezhi (Gu et al., 2024)	2024	Zhouhong Gu et al.
43	CELLO (He et al., 2024)	2024	Qianyu He et al.

7.2 Evaluation Algorithm for Emotional Companionship Capability

This appendix provides detailed pseudocode for the algorithm and a complete reference table for the symbols, functions, and data structures used. To calculate the final score for emotional companionship capability, ECDBench 1.0 implements a four-step process: progressing from the method layer to the data layer, from the data layer to the task layer, from the task layer to the capability layer, and finally calculating the total score. This process systematically transforms a model's scattered performance on specific tasks into a final comprehensive score that measures its emotional companionship capability (for the detailed process, see Algorithm 1; for explanations of specific functions, see Table 3).

7.2.1 Step 1: Composite Score Algorithm for Datasets(Method Layer → Data Layer)

Objective : Starting from the most fundamental evaluation methods, to calculate a unified and standardized composite score for each dataset.

Process :

1. **Raw Score Generation and Normalization**: For a specific dataset (D) and evaluation method (M), the model's raw score, E_{raw}, is first obtained. Since raw scores can come in various formats (e.g., numerical values, grades, ratios), **Algorithm 2 (NormalizeScore)** is called to uniformly translate them into a standardized score, S_{norm}, on a 0-100 scale, thereby ensuring comparability.

2. **Dataset Score Aggregation**: If a single dataset is evaluated by multiple methods, it will generate multiple standardized scores. **Algorithm 3 (AggregateScores)** is then called to perform a weighted average of these scores, ultimately yielding the composite score for the dataset, $S(D)$.

$$S(D) = \frac{\sum_{M \in \mathcal{M}_D} w(M) \cdot S_{norm}(M, D)}{\sum_{M \in \mathcal{M}_D} w(M)} \tag{1}$$

   Here, $S(D)$ is the composite score of dataset D, $\mathcal{M}_D$ is the set of all evaluation methods applied to dataset D, $S_{norm}(M, D)$ is the standardized score for the corresponding method, and $w(M)$ is the weight of method M.

7.2.2 Step 2: Task Layer Score Aggregation Algorithm (Data Layer → Task Layer)

Objective : To hierarchically aggregate the scattered dataset scores into "sub-task scores" and "difficulty-level scores" according to the definitions in the evaluation framework.

Process :

1. **Sub-task Score Aggregation**: The algorithm iterates through all sub-tasks (T_{sub}) defined in the configuration file, finds the scores of the datasets that constitute each sub-task, and calculates the sub-task score $S(T_{sub})$ by performing a weighted average using **Algorithm 3 (AggregateScores)**.

$$S(T_{sub}) = \frac{\sum_{D \in \mathcal{D}_{sub}} w(D, T_{sub}) \cdot S(D)}{\sum_{D \in \mathcal{D}_{sub}} w(D, T_{sub})} \tag{2}$$

2. **Difficulty-Level Score Aggregation**: Through a three-level loop, the algorithm iterates through Main Capability → Sub-capability → three difficulty levels ("Low (L)," "Medium (M)," "High (H)") (λ). It then aggregates the scores of all sub-tasks belonging to the same difficulty level under a sub-capability, again using **Algorithm 3 (AggregateScores)**, to obtain the total score for that difficulty level, $S(T^{\lambda})$.

$$S(T^{\lambda}) = \frac{\sum_{T_{sub} \in \mathcal{T}^{\lambda}} w(T_{sub}) \cdot S(T_{sub})}{\sum_{T_{sub} \in \mathcal{T}^{\lambda}} w(T_{sub})} \tag{3}$$

Algorithm 1 Overall Benchmark Evaluation Flow

function CALCULATEBENCHMARKSCORE(Model)
▷ Initialize configurations, e.g., datasets, tasks, capabilities, weights
Config ← LoadConfiguration()
Scores ← InitializeScoreStorage()
▷ **Step 1: Calculate the composite score for each dataset (Method Layer → Data Layer)**
for each Dataset D in Config.Datasets **do**
method_scores ← []
for each Method M used for D **do**
E_{raw} ← GetRawScore(Model, M, D)
S_{norm} ← NORMALIZESCORE(E_{raw}, M.type, M.params) ▷ Call Algorithm 2
Append (S_{norm}, M.weight) to method_scores
end for
Scores.dataset[D] ← AGGREGATESCORES(method_scores) ▷ Call Algorithm 3
end for
▷ **Step 2: Aggregate task scores (Data Layer → Task Layer)**
for each SubTask T_{sub} in Config.Tasks **do**
dataset_scores ← GetRelevantScores(Scores.dataset, T_{sub}.datasets)
Scores.subtask[T_{sub}] ← AGGREGATESCORES(dataset_scores)
end for
for each MainAbility A_i in Config.Capabilities.main_abilities **do** ▷ Iterate through main capabilities
for each SubAbility A_{ij} in A_i.sub_abilities **do** ▷ Iterate through sub-capabilities
for each Level λ in {L, M, H} **do**
subtask_scores ← GetRelevantScores(Scores.subtask, A_{ij}.tasks(λ)) ▷ Get tasks for the sub-capability at a specific difficulty level
Scores.task_level[A_{ij}, λ] ← AGGREGATESCORES(subtask_scores)
end for
end for
end for
▷ **Step 3: Synthesize capability scores (Task Layer → Capability Layer)**
for each SubAbility A_{ij} in Config.Capabilities **do**
Scores.sub_ability[A_{ij}] ← SYNTHESIZESUBABILITYSCORE(Scores.task_level, A_{ij}) ▷ Call Algorithm 4
end for
for each MainAbility A_i in Config.Capabilities **do**
sub_ability_scores ← GetRelevantScores(Scores.sub_ability, A_i.sub_abilities)
Scores.main_ability[A_i] ← AGGREGATESCORES(sub_ability_scores)
end for
▷ **Step 4: Calculate the final total score (Capability Layer → Total Score)**
S_{total} ← CALCULATEFINALSCORE(Scores.main_ability, Config.penalty_rule) ▷ Call Algorithm 5
return S_{total}
end function

Algorithm 2 Score Normalization (NormalizeScore)

function NORMALIZESCORE(E_{raw}, type, params)
  if type is "numeric" **then** ▷ e.g., a five-point scale with a range of [1, 5]
    $(S_{min}, S_{max}) \leftarrow$ params
    return $(E_{raw} - S_{min})$ / $(S_{max} - S_{min}) \times 100$
  else if type is "grade" **then** ▷ e.g., "Excellent" -> 95
    mapping_table $\leftarrow$ params
    return mapping_table[E_{raw}]
  else if type is "ratio" **then** ▷ e.g., accuracy, with a range of [0, 1]
    return $E_{raw} \times 100$
  else
    return E_{raw} ▷ If no specific type, input is assumed to be on a 0-100 scale
  end if
end function

Here, $S(T_{sub})$ represents the score of a single sub-task; $\mathcal{D}_{sub}$ is the set of all datasets that constitute sub-task T_{sub}; and $S(D)$ is the composite score of dataset D. $S(T^{\lambda})$ refers to the task score for difficulty level λ, while $\mathcal{T}^{\lambda}$ is the set of all sub-tasks at that difficulty level. $w(D, T_{sub})$ and $w(T_{sub})$ are the weights of the dataset and the sub-task, respectively. In ECDBench 1.0, their weights are provisionally set to 1.

Algorithm 3 General Weighted Aggregation (AggregateScores)

function AGGREGATESCORES(scored_items) ▷ Input is a list of (score, weight) pairs
  total_weighted_score $\leftarrow 0$
  total_weight $\leftarrow 0$
  for each (score, weight) in scored_items **do**
    total_weighted_score $\leftarrow$ total_weighted_score + score $\times$ weight
    total_weight $\leftarrow$ total_weight + weight
  end for
  if total_weight = 0 **then**
    return 0
  else
    return total_weighted_score / total_weight
  end if
end function

7.2.3 Step 3: Capability Layer Score Synthesis Algorithm (Task Layer → Capability Layer)

Objective : To map and synthesize specific task performance scores into a capability score that can measure a certain capability of the model.

Process :

1. **Sub-capability Score Synthesis**:

   This stage serves as the bridge connecting "tasks" and "capabilities." The algorithm calls the specialized **Algorithm 4 (SynthesizeSubAbilityScore)** to perform a weighted sum of the task scores for the three difficulty levels, $S(T^{\lambda})$, based on preset asymmetric difficulty weights (e.g., Low 30%, Medium 55%, High 15%), to obtain the final score for the "sub-capability" (A_{ij}), which is $S(A_{ij})$.

$$S(A_{ij}) = \omega^{L} \cdot S(T_{ij}^{L}) + \omega^{M} \cdot S(T_{ij}^{M}) + \omega^{H} \cdot S(T_{ij}^{H}) \tag{4}$$

   Here, $\omega^{L}, \omega^{M}, \omega^{H}$ are the difficulty weights for the Low, Medium, and High tasks, respectively, and their sum is 1. The weights are set as follows:

- **Standard Weights**: When a sub-capability includes tasks of all three difficulty levels, the weights are set to $\omega^L = 0.3$, $\omega^M = 0.55$, and $\omega^H = 0.15$.
- **Special Cases (Missing Tasks)**: When a sub-capability only includes tasks of some difficulty levels, the weights are dynamically adjusted proportionally to ensure their sum is 1.
  (a) If there is only **one level** of evaluation tasks, the weight for that level is 1.
  (b) If there are only **Low and Medium** tasks, the weights are adjusted to $\omega^L = 0.4$ and $\omega^M = 0.6$.
  (c) If there are only **Low and High** tasks, the weights are adjusted to $\omega^L = 0.6$ and $\omega^H = 0.4$.
  (d) If there are only **Medium and High** tasks, the weights are adjusted to $\omega^M = 0.7$ and $\omega^H = 0.3$.

2. **Main Capability Score Aggregation**: The algorithm iterates through all "main capabilities" (A_i) and aggregates the scores of all their subordinate sub-capabilities using **Algorithm 3 (AggregateScores)** (typically by arithmetic mean) to obtain the main capability score, $S(A_i)$.

$$S(A_i) = \frac{\sum_{A_{ij} \in \mathcal{A}_i} w(A_{ij}) \cdot S(A_{ij})}{\sum_{A_{ij} \in \mathcal{A}_i} w(A_{ij})} \tag{5}$$

Here, $S(A_i)$ represents the score of main capability i; $\mathcal{A}_i$ is the set of all sub-capabilities that constitute main capability A_i; $S(A_{i,j})$ is the score of sub-capability j; and $w(A_{i,j})$ is the weight of sub-capability j. In ECDBench 1.0, we provisionally consider each sub-capability to be of equal importance, thus setting $w(A_{i,j}) = 1$.

Algorithm 4 Sub-capability Score Synthesis (SynthesizeSubAbilityScore)

function SYNTHESIZESUBABILITYSCORE(task_level_scores, sub_ability_config)
  ▷ Get scores for each difficulty level; if missing, the score is 0
  score_L ← GetScoreForLevel(task_level_scores, L, sub_ability_config)
  score_M ← GetScoreForLevel(task_level_scores, M, sub_ability_config)
  score_H ← GetScoreForLevel(task_level_scores, H, sub_ability_config)
  ▷ Get dynamically adjusted difficulty weights based on existing task levels
  (w_L, w_M, w_H) ← GetAdjustedDifficultyWeights(score_L, score_M, score_H)
  $S_{A_{ij}} \leftarrow w_L \cdot$ score_L $+ w_M \cdot$ score_M $+ w_H \cdot$ score_H
  return $S_{A_{ij}}$
end function

7.2.4 Step 4: Final Score Calculation Algorithm (Capability Layer → Total Score)

Objective : To calculate a final total score that represents the model's overall performance by applying weights to each main capability score.

Process :

1. **Threshold Check**: First, the algorithm checks if the score for the "threshold capability" (A_k), "Values & Safety," meets the preset minimum threshold, τ (set to 60 in ECDBench). If the model fails to meet this threshold, a "one-vote veto" mechanism is triggered, and the total score is recorded as 0.

2. **Weighted Total Score**: If the model passes the threshold check, a weighted sum of the scores for each main capability is calculated based on their final weights (in ECDBench, these are set to Foundational Capability 30%, Emotional Capability 40%, and Companionship Capability 30%) to derive the final total score, S_{total}.

$$S_{total} = \begin{cases} \frac{\sum_{A_i \in \mathcal{A}_{main}} w(A_i) \cdot S(A_i)}{\sum_{A_i \in \mathcal{A}_{main}} w(A_i)} & \text{if } S(A_k) \geq \tau \\ 0 & \text{if } S(A_k) < \tau \end{cases} \tag{6}$$

Algorithm 5 Final Score Calculation (CalculateFinalScore)

function CALCULATEFINALSCORE(main_ability_scores, penalty_rule, main_ability_weights)
    $A_k \leftarrow$ penalty_rule.gate_ability ▷ threshold capability, e.g., "Values & Safety"
    $\tau \leftarrow$ penalty_rule.threshold ▷ Minimum passing threshold, e.g., 60
    if main_ability_scores[A_k] < τ **then**
        return 0 ▷ Trigger "one-vote veto"; total score is 0
    else
        ▷ Prepare a list for weighted aggregation of main capabilities
        scored_items $\leftarrow$ []
        for each capability_name, score in main_ability_scores **do**
            if capability_name is not A_k **then**
                weight $\leftarrow$ main_ability_weights[capability_name] ▷ Get the corresponding weight
                Append (score, weight) to scored_items
            end if
        end for
        $S_{total} \leftarrow$ AGGREGATESCORES(scored_items)
        return S_{total}
    end if
end function

Reference Table for Symbols, Functions, and Data Structures as tab 3.

Table 3: Reference Table for Symbols, Functions, and Data Structures

Symbol/Name	Conceptual Definition
Core Data Structures	
Config	The evaluation configuration object. Stores the "blueprint" for the entire benchmark, including task hierarchies, datasets, capability definitions, and all weights and other metadata.
Scores	The dynamic score storage object. Used to cache scores at various levels during the calculation process, such as Scores.dataset, Scores.subtask, etc.
Main & Helper Functions	
CalculateBenchmarkScore	The main function for the overall evaluation process. It takes a model to be tested (Model) as input and serves as the entry point for the entire algorithm.
LoadConfiguration	An initialization function responsible for loading configuration information from an external file (e.g., YAML/JSON) and constructing the Config object.
InitializeScoreStorage	An initialization function responsible for creating an empty Scores object before the calculation begins, which is used to store subsequent results.
GetRawScore	The raw score retrieval function. It calls the specific evaluation program to obtain the model's raw, unprocessed score for a particular dataset and method.
NormalizeScore	The score normalization function. It converts raw scores of various formats (e.g., five-point scale, grade-based) into a standardized 0-100 score.
AggregateScores	The general weighted aggregation function. It takes a list of (score, weight) pairs and calculates their weighted average score.
GetRelevantScores	The data filtering and retrieval function. Based on a given list of names, it precisely extracts a relevant subset of scores from a larger score pool to prepare for the next aggregation step.
SynthesizeSubAbilityScore	The sub-capability score synthesis function. It calculates the weighted sum of task scores based on preset asymmetric difficulty weights (Low/Medium/High) to derive the sub-capability score.
GetScoreForLevel	The score retrieval function. Used during the synthesis of sub-capability scores to obtain the task score for a specific difficulty level (L/M/H), returning 0 if it does not exist.
CalculateFinalScore	The final total score calculation function. It is responsible for performing the "threshold capability" check and calculating a weighted sum of the scores of the main capabilities that pass the check, based on preset weights, to derive the final total score.
Mathematical & Logical Symbols	
$S(\cdot)$	Scoring Function, calculates the score of the object in the parentheses.
$E_{raw}(M, D)$	The raw score of method M on dataset D.
$S_{norm}(M, D)$	The 0-100 standardized score of method M on dataset D.
S_{min}, S_{max}	The minimum/maximum value range for numerical scores.
$w(\cdot)$	General Weight, the weight of different objects is distinguished by context.
w_{A_i}	The weight of main capability A_i.

(Continued) Reference Table for Symbols, Functions, and Data Structures

Symbol/Name	Conceptual Definition
ω^{λ}	Task difficulty weight, $\lambda \in \{\text{L, M, H}\}$.
D	Dataset.
M	Evaluation Method.
$\mathcal{M}_D$	The set of all evaluation methods used for dataset D.
T_{sub}	Sub-task.
$\mathcal{D}_{sub}$	The set of datasets that constitute sub-task T_{sub}.
T^{λ}	Task at a specific difficulty level λ.
$\mathcal{T}^{\lambda}$	The set of all sub-tasks at difficulty level λ.
λ	Difficulty Level, with values L (Low), M (Medium), H (High).
A_i	The i-th Main Capability.
A_{ij}	The j-th Sub-capability under the main capability A_i.
$\mathcal{A}_i$	The set of all sub-capabilities contained in main capability A_i.
$\mathcal{A}_{main}$	The set of all main capabilities participating in the total score calculation (excluding the threshold capability).
A_k	Threshold capability, specifically refers to "Values & Safety".
τ	The minimum passing threshold for the threshold capability.

7.3 Information on Models Tested

Our selection of 30 models for evaluation was guided by three key principles (see Table 4):

- **Geographic and Technical Diversity:** To include models from both domestic and international sources representing different technical routes and cultural backgrounds.
- **Varied Development Paradigms:** To compare the Capability differences between closed- and open-source models.
- **Scale Validation:** To select models with different parameter scales from the same family (e.g., Doubao, Qwen, GPT series) for validating the effectiveness of Scaling Laws in the emotional companionship domain.

Table 4: Basic Information of Models Tested

Model Name	Institution	Type	Parameters
Doubao Series			
doubao-seed-1-6-flash-250615	ByteDance	Closed-source	10B (Estimated)
doubao-1.5-pro-32k-character-250715	ByteDance	Closed-source	50-60B (Estimated)
doubao-pro-32k-241215	ByteDance	Closed-source	66.5B-130B (Estimated)
OpenAI Series			
gpt-4o-mini	OpenAI	Closed-source	8B (Estimated)
gpt-5-nano-2025-08-07	OpenAI	Closed-source	20B (Estimated)
gpt-5-mini-2025-08-07	OpenAI	Closed-source	50B (Estimated)
gpt-4-turbo	OpenAI	Closed-source	1.8T (Total MoE Parameters)
gpt-4o	OpenAI	Closed-source	1.8T (Total MoE Parameters)
gpt-4.1-2025-04-14	OpenAI	Closed-source	1.8T (Total MoE Parameters)
gpt-5-2025-08-07	OpenAI	Closed-source	1.8T (Total MoE Parameters)
Alibaba (Qwen) Series			
qwen3-0.6b	Alibaba Tongyi Qianwen Team	Open-source	0.6B
qwen3-14b	Alibaba Tongyi Qianwen Team	Open-source	14B
qwen3-32b	Alibaba Tongyi Qianwen Team	Open-source	32B
qwen-long	Alibaba Tongyi Qianwen Team	Open-source	32B
qwen-plus	Alibaba Tongyi Qianwen Team	Closed-source	Hundreds of Billions (Estimated)
qwen-max	Alibaba Tongyi Qianwen Team	Closed-source	Hundreds of Billions (Estimated)
Baichuan Series			

Table 4: Basic Information of Models Tested

Model Name	Institution	Type	Parameters
Baichuan2-turbo	Baichuan Inc.	Closed-source	100B (Estimated)
Baichuan4-Air	Baichuan Inc.	Closed-source	400B/170B (Total/Activated Parameters, Estimated)
DeepSeek Series			
deepseek-v3.1	DeepSeek	Open-source	671B/37B (Total/Activated Parameters)
Anthropic Series			
claude-3-sonnet-20240229	Anthropic	Closed-source	70B (Estimated)
claude-opus-4-20250514	Anthropic	Closed-source	1.2T (Estimated)
Google Series			
gemini-1.5-flash-002	Google	Closed-source	500B (Estimated)
gemini-2.0-flash	Google	Closed-source	1.2T (Estimated)
Meta (LLaMA) Series			
llama-4-scout	Meta	Open-source	109B/17B (Total/Activated Parameters)
llama-4-maverick	Meta	Open-source	400B/170B (Total/Activated Parameters)
Zhipu AI (GLM) Series			
glm-4-9b	Zhipu AI	Open-source	9B
glm-4-flash	Zhipu AI	Closed-source	100B (Estimated)
glm-4v-plus	Zhipu AI	Closed-source	200B (Estimated)
Baidu (ERNIE) Series			
ernie-4.5-21b-a3b	Baidu	Closed-source	21B
ABAB Series			
abab6.5t-chat	MiniMax	Closed-source	1T

7.4 Overall Leaderboard of Evaluation Results

The final evaluation rankings are presented in Table 5, where, for each Capability, the highest score (and those within 1 point) is marked in bold and underlined, while the lowest score (and those within 1 point) is marked in red.

Table 5: Overall Leaderboard of Model Emotional Companionship Capability

Rank	Model Name	Overall Score	Foundational Capability	Emotional Capability	Companionship Capability	Values & Safety
1	gpt-5-mini-2025-08-07	70.09	81.98	68.90	59.79	92.53
2	doubao-pro-32k-241215	69.44	80.21	68.02	60.59	96.33
3	gpt-5-2025-08-07	69.34	83.00	67.71	57.87	90.70
4	gpt-4o	69.09	78.07	71.19	57.31	92.02
5	gpt-4.1-2025-04-14	68.57	79.68	66.30	60.47	92.48
6	qwen3-32b	68.40	76.38	69.39	59.11	95.50
7	qwen-long	68.32	78.12	68.37	58.44	90.42
8	deepseek-v3.1	68.26	78.20	67.08	59.90	95.77
9	doubao-1.5-pro-32k-character-250715	68.25	79.22	68.11	57.45	90.97
10	qwen-max	68.07	79.65	66.83	58.15	92.60
11	gpt-5-nano-2025-08-07	67.89	79.50	67.60	56.66	89.62
12	doubao-seed-1-6-flash-250615	67.85	76.93	68.54	57.86	91.43

13	gemini-2.0-flash	67.31	76.42	64.80	61.56	90.52
14	Baichuan4-Air	67.22	73.81	70.16	56.72	82.60
15	Baichuan2-turbo	66.98	74.37	70.35	55.10	82.62
16	llama-4-scout	66.78	74.26	66.92	59.11	95.68
17	claude-3-sonnet-20240229	66.45	79.01	68.12	51.66	83.08
18	llama-4-maverick	66.31	74.35	66.67	57.78	93.53
19	gpt-4-turbo	66.10	76.68	66.15	55.45	92.47
20	claude-opus-4-20250514	65.67	76.67	69.60	49.42	70.82
21	qwen3-14b	65.43	74.54	64.61	57.42	90.15
22	glm-4-9b	65.23	70.18	65.02	60.56	80.68
23	glm-4v-plus	64.25	72.58	63.97	56.30	83.12
24	gpt-4o-mini	64.16	73.72	66.18	51.92	87.58
25	qwen-plus	63.65	78.48	67.31	43.97	89.42
26	ernie-4.5-21b-a3b	63.57	72.37	61.67	57.31	85.23
27	glm-4-flash	63.13	69.56	62.84	57.10	80.47
28	gemini-1.5-flash-002	62.95	73.05	63.12	52.63	91.00
*** Note: Models are not included in the final ranking for their threshold capability score was below 60.**						
*	qwen3-0.6b	47.40	49.20	46.83	46.36	40.83
*	abab6.5t-chat	37.53	42.77	36.36	33.85	51.93

7.5 Overview of ECDBench 1.0

Table 6 provides a detailed list of the low, medium, and high-level evaluation tasks corresponding to each sub-capability, from Threshold to Core Capabilities, and specifies the datasets and evaluation methods used for each task.

Table 6: Overview of ECDBench 1.0 Tasks, Datasets, and Evaluation Methods

Capability Dimension	Sub-capability	Task Difficulty	Evaluation Task	Dataset	Evaluation Method	Metrics	Question Type
Values & Safety	Values	Low	Bias Detection	CrowS-Pairs	Benchmark	Accuracy	MCQ
				StereoSet (Nadeem et al., 2021)	Benchmark	Language Model Score, Bias Score	MCQ
				BBQ (Parrish et al., 2022)	Benchmark	Accuracy, Bias Score	MCQ
				SafetyBench7 - Unfairness and Bias (Zhang et al., 2023)	Benchmark	Accuracy	MCQ
		Medium	Morality Detection	SafetyBench1 - Ethics and Morality (Zhang et al., 2023)	Benchmark	Accuracy	MCQ
	Safety	Low	Content Safety	SafetyBench3-Offensiveness(Zhang et al., 2023)	Benchmark	Accuracy	MCQ

Capability Dimension	Capability Name	Task Difficulty	Task	Dataset	Evaluation Method	Metrics	Question Type
		Medium	Information Security	SafetyBench4 - Privacy and Property (Zhang et al., 2023)	Benchmark	Accuracy	MCQ
			User Safety	SafetyBench2 - Mental Health (Zhang et al., 2023)	Benchmark	Accuracy	MCQ
				SafetyBench6 - Physical Health (Zhang et al., 2023)	Benchmark	Accuracy	MCQ
			High-Risk Harmful Behavior	SafetyBench5 - Illegal Activities (Zhang et al., 2023)	Benchmark	Accuracy	MCQ
Foundational Capability	Natural Language Understanding	Low	Text Classification	AG News (Zhang et al., 2015)	Benchmark	Accuracy	MCQ
				THUCNews(Sun et al., 2016)	Benchmark	Accuracy	MCQ
			Text Matching	LCQMC(Liu et al., 2018)	Benchmark	Accuracy	MCQ
				QQP (Iyer et al., 2017)	Benchmark	Accuracy	MCQ
		Medium	Reading Comprehension (Extractive)	CMRC(Cui et al., 2019)	Benchmark	Rouge	Open-ended
			Word Sense Disambiguation	WiC(Pilehvar and Camacho-Collados, 2019)	Benchmark	Accuracy	MCQ
		High	Reading Comprehension	MultiRC (Khashabi et al., 2018)	Benchmark	F1 / EM	Multiple Answer
				ReCoRD(Zhang et al., 2018b)	Benchmark	F1 / EM	Multiple Answer
	Natural Language Reasoning	Low	Textual Entailment	OCNLI (Xu et al., 2020a)	Benchmark	Accuracy	MCQ
				MNLI (Williams et al., 2018)	Benchmark	Accuracy	MCQ
				RTE (Dagan et al., 2005)	Benchmark	Accuracy	MCQ
			Simple Causal Reasoning	COPA (Roemmele et al., 2011)	Benchmark	Accuracy	MCQ
		Medium	Coreference Resolution & Commonsense Reasoning	WSC (Levesque et al., 2012)	Benchmark	Accuracy	MCQ
			Multi-turn Dialogue Reasoning	MuTual (Cui et al., 2020)	Benchmark	Accuracy	MCQ

Capability Dimension	Capability Name	Task Difficulty	Task	Dataset	Evaluation Method	Metrics	Question Type
		High	First-Order Logic Reasoning	FOLIO(Han et al., 2023)	Benchmark	Recall @ K, MRR	MCQ
			Complex Logical Reasoning	LogiQA (Liu et al., 2020)	Benchmark	Accuracy	MCQ
	Natural Language Generation	Low	Summarization	LCSTS (Hu et al., 2015)	Model	Completeness, conciseness, etc.	Open-ended
				CNewSum (Wang et al., 2022)	Model	Completeness, conciseness, etc.	Open-ended
				VCSum (Liu et al., 2023b)	Model	Completeness, conciseness, etc.	Open-ended
	Commonsense Capability	Low	Daily Scenario Q&A	HellaSwag(Zellers et al., 2019)	Benchmark	Accuracy	MCQ
			Commonsense Q&A	Cosmos QA (Huang et al., 2019)	Benchmark	Accuracy	MCQ
			Physical Commonsense	PIQA(Bisk et al., 2020)	Benchmark	Accuracy	MCQ
		Medium	Human Cognition & Exam Capabilities	AGIEval (Zhong et al., 2023)	Benchmark	Accuracy	MCQ & Open-ended
			Multi-task Knowledge Q&A	MMLU-pro (Chien et al., 2024)	Benchmark	Accuracy	MCQ
				C-MNLU(YAO et al., 2023)	Benchmark	Accuracy	MCQ
				C-Eval (Huang et al., 2023c)	Benchmark	Accuracy	MCQ
		High	Truthfulness Q&A	TruthfulQA v1(Lin et al., 2022b)	Benchmark	BLEURT, ROUGE, BLEU	Single & MCQ
Emotional Capability	Emotion Recognition	Low	Sentiment Polarity Recognition	IMDb (Maas et al., 2011)	Benchmark	Accuracy	MCQ
				SST-2 (Socher et al., 2013)	Benchmark	Accuracy	MCQ
				Dianping Dataset	Benchmark	Accuracy	MCQ
		Medium	Coarse-grained Emotion Recognition	ECDBench1	Benchmark	Accuracy	MCQ
			Irony Detection	SemEval-2018 Task 3(Van Hee et al., 2018)	Benchmark	F1	MCQ
			Metaphor Detection	VUA20(Leong et al., 2020)	Benchmark	F1	MCQ
		High	Fine-grained Emotion Recognition	GoEmotions (Demszky et al., 2020b)	Benchmark	F1	MCQ

Capability Dimension	Capability Name	Task Difficulty	Task	Dataset	Evaluation Method	Metrics	Question Type
				CPED (Chen et al., 2022)	Benchmark	Accuracy	MCQ
				EDOS(Anuradha et al., 2021)	Benchmark	Accuracy	MCQ
			Emotion Recognition in Complex Scenarios	EmoBench1(Li et al., 2024b)	Benchmark	Accuracy	MCQ
	Emotion Understanding	Low	Emotion Cause Analysis	ECDBench2	Benchmark	Accuracy	MCQ
				EmoBench2(Li et al., 2024b)	Benchmark	Accuracy	MCQ
		Medium	Emotion Intensity Understanding	SemEval-2018 Task 1 (Mohammad et al., 2018)	Benchmark	Accuracy	MCQ
	Emotion Management	High	Emotion Strategy Selection	ECDBench3	Benchmark	Accuracy	MCQ
	Empathetic Response	Medium	Empathetic Expression	EmoBench3(Li et al., 2024b)	Benchmark	Accuracy	MCQ
		High	Contextual Empathetic Response	ECDBench4	Benchmark	Accuracy	MCQ
	Comprehensive Emotional Intelligence	Low	EQ Test	EQ-60 Empathy Quotient Scale	Benchmark	Score	MCQ
				IRI 30-item EQ Questionnaire	Benchmark	Score	MCQ
		Medium	Interpersonal Relationship Test	Interpersonal Relationship Test	Benchmark	Score	MCQ
		High	Effective Communication Test	Effective Communication Test	Benchmark	Score	MCQ
Companionship Capability	Memory & Personalization	Low	Degree of Personalization	PersonaFeedback (Tao et al., 2024)	Model	Accuracy	MCQ
		Medium	Sequence Order Recall	Book-SORT (Levy et al., 2024)	Benchmark	Accuracy	MCQ
		High	Long-term Conversation Recall	LongMemEval (Cheng et al., 2024)	Benchmark	Recall@K	Open-ended

*In the Evaluation Method column: Benchmark = Benchmark-based, Model = Model-based.

References

Sanathkumar Anuradha, Manisha Ghangas, Ishita Gupta, Naman Singh, Rahul Arora, Sanchit Singh, Shreya Goel, and Anshul Kumar. 2021. A large-scale dataset for empathetic response generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1247–1262, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Chayapatr Archiwaranguprok, Constanze Albrecht, Pattie Maes, Karrie Karahalios, and Pat Pataranutaporn. 2025. Simulating psychological risks in human-ai interactions: Real-case informed modeling of ai-induced addiction, anorexia, depression, homicide, psychosis, and suicide. *arXiv preprint arXiv:2511.08880*.

Xin Bai, Guanyi Chen, Tingting He, Chenlian Zhou, and Cong Guo. 2025. A holistic comparative study of large language models as emotional support dialogue systems. *Cognitive Computation*, 17(2):71.

Timothy W Bickmore, Lisa Caruso, and Kerri Clough-Gorr. 2005. Acceptance and usability of a relational agent interface by urban older adults. In *CHI'05 extended abstracts on Human factors in computing systems*, pages 1212–1215.

Timothy W Bickmore and Rosalind W Picard. 2005. Establishing and maintaining long-term human-computer relationships. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 12(2):293–327.

Yonatan Bisk, Rowan Zellers, Ronan Lebras, Jian Gao, and Yejin Choi. 2020. PIQA: Reasoning about physical commonsense in natural language. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence*, New York, USA.

Kranti Chalamalasetti, Jana Götze, Sherzod Hakimov, Brielen Madureira, Philipp Sadler, and David Schlangen. 2023. clembench: Using game play to evaluate chat-optimized language models as conversational agents. *arXiv preprint arXiv:2305.13455*.

Mark Chen. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

Yirong Chen, Weiquan Fan, Xiaofen Xing, Jianxin Pang, Minlie Huang, Wenjing Han, Qianfeng Tie, and Xiangmin Xu. 2022. CPED: A large-scale chinese personalized and emotional dialogue dataset for conversational AI. *Preprint*, arXiv:2205.14727.

Feifei Cheng, Shuo Wang, Zhaochun Liu, Dejiao Peng, Xun Sun, James Glass, Asli Celikyilmaz, and Jianfeng Gao. 2024. LongMemEval: Benchmarking Chat Assistants on Long-Term Interactive Memory. In *The 2024 Conference on Empirical Methods in Natural Language Processing*.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot arena: An open platform for evaluating llms by human preference. *Preprint*, arXiv:2403.04132.

Tsung-Hsun Chien, Pin-Jui Li, Qucheng Niu, Lilian Weng, Ruoxi Jia, Hsiang-Fu Yu, Chih-Jen Lin, S. V. N. Vishwanathan, and Inderjit S. Dhillon. 2024. MMLU-Pro: A more robust and challenging multi-task language understanding benchmark. *Preprint*, arXiv:2405.08298.

Minje Choi, Jiaxin Pei, Sagar Kumar, Chang Shu, and David Jurgens. 2023. Do llms understand social knowledge? evaluating the sociability of large language models with socket benchmark. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, and 1 others. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

OpenCompass Contributors. 2023. Opencompass: A universal evaluation platform for foundation models. `https://github.com/open-compass/opencompass`.

Leyang Cui, Yu Wu, Shujie Liu, Yue Zhang, and Ming Zhou. 2020. MuTual: A dataset for multi-turn dialogue reasoning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1108–1118, Online. Association for Computational Linguistics.

Yiming Cui, Ting Liu, Zhipeng Chen, Shijin Wang, and Guoping Hu. 2019. A span-extraction dataset for chinese machine reading comprehension. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5886–5891, Hong Kong, China. Association for Computational Linguistics.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The pascal recognising textual entailment challenge. In *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment: First PASCAL Machine Learning Challenges Workshop, MLCW 2005, Southampton, UK, April 11-13, 2005, Revised Selected Papers*, pages 177–190. Springer.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020a. Goemotions: A dataset of fine-grained emotions. *arXiv preprint arXiv:2005.00547*.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gen Niemi, and Dacher Keltner. 2020b. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054, Online. Association for Computational Linguistics.

Zhouhong Gu, Xiaoxuan Zhu, Haoning Ye, Lin Zhang, Jianchen Wang, Yixin Zhu, Sihang Jiang, Zhuozhi Xiong, Zihan Li, Weijie Wu, and 1 others. 2024. Xiezhi: An ever-updating benchmark for holistic domain knowledge evaluation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 18099–18107.

Xinya Du Han, Zhengbao Zhong, Zhiruo Wang Al-Obaidi, Yuliang Shu, Peixuan Shi, Peng Dou, and Rui Qian. 2023. FOLIO: A benchmark for evaluating natural language reasoning in financial question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4706–4721, Singapore. Association for Computational Linguistics.

Qianyu He, Jie Zeng, Wenhao Huang, Lina Chen, Jin Xiao, Qianxi He, Xunzhe Zhou, Jiaqing Liang, and Yanghua Xiao. 2024. Can large language models understand real-world complex instructions? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18188–18196.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021a. Measuring massive multitask language understanding. *Preprint*, arXiv:2009.03300.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021b. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. 2024. Ruler: What's the real context size of your long-context language models? *arXiv preprint arXiv:2404.06654*.

Baotian Hu, Qingcai Chen, and Fangze Zhu. 2015. LCSTS: A large scale Chinese short text summarization dataset. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1677–1687, Lisbon, Portugal. Association for Computational Linguistics.

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.

Lifu Huang, Deye Le, Eunsol Choi, Simon Whitehead, and Kai-Wei Chang. 2019. Cosmos QA: Machine reading comprehension with contextual commonsense reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2369–2379, Hong Kong, China. Association for Computational Linguistics.

Yue Huang, Qihui Zhang, Lichao Sun, and 1 others. 2023a. Trustgpt: A benchmark for trustworthy and responsible large language models. *arXiv preprint arXiv:2306.11507*.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Yao Fu, and 1 others. 2023b. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *Advances in Neural Information Processing Systems*, 36:62991–63010.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Yang Su, Junteng Liu, Chunhui Lv, Cui Li, and Li Yufeng. 2023c. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. In *Advances in Neural Information Processing Systems*, volume 36.

Shankar Iyer, Nikhil Dandekar, and Kornel Csernai. 2017. Quora question pairs. `https://www.quora.com/q/quoradata/First-Quora-Dataset-Release-Question-Pairs`.

Lucie-Aimée Kaffee, Giada Pistilli, and Yacine Jernite. 2025. Intima: a benchmark for human-ai companionship behavior. *arXiv preprint arXiv:2508.09998*.

Elise Karinshak, Amanda Hu, Kewen Kong, Vishwanatha Rao, Jingren Wang, Jindong Wang, and Yi Zeng. 2024. Llm-globe: A benchmark evaluating the cultural values embedded in llm output. *arXiv preprint arXiv:2411.06032*.

Daniel Khashabi, Snigdha Chaturvedi, and Dan Roth. 2018. Looking beyond the surface: A challenge set for reading comprehension over multiple sentences. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 252–262, New Orleans, Louisiana. Association for Computational Linguistics.

Ervin Laszlo. 1972. *Introduction to Systems Philosophy: Toward a New Paradigm of Contemporary Thought*. Gordon and Breach, New York.

Yoon Kyung Lee, Jina Suh, Hongli Zhan, Junyi Jessy Li, and Desmond C Ong. 2024. Large language models produce responses perceived to be empathic. In *2024 12th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 63–71. IEEE.

Chee Wee Leong, Beata Beigman Klebanov, Ekaterina Shutova, and Gerard Steen. 2020. Report on the 2020 vua and Leiden university metaphor dectection shared task. In *Proceedings of the 4th Workshop on Figurative Language Processing*, pages 209–216, Online. Association for Computational Linguistics.

Hector Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning*.

Elad Levi and Ilan Kadar. 2025. Intellagent: A multi-agent framework for evaluating conversational ai systems. *arXiv preprint arXiv:2501.11067*.

Moran Levy, Elad Guz, Yoav Levine, Roee Aharoni, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2024. Assessing episodic memory in LLMs with sequence order recall tasks. In *The Twelfth International Conference on Learning Representations*.

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. 2024a. Cmmlu: Measuring massive multitask language understanding in chinese. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11260–11285.

Jingsheng Li, Yanzhou Zhou, Xuanye Li, Jiachen Li, Zhaoxuan Wei, Xinyi Li, Jiazeng Li, Ziyu Wang, Bowen Shen, Peipei Li, Xipeng Qiu, and Minlie Huang. 2024b. EmoBench: Evaluating the emotional intelligence of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Bangkok, Thailand. Association for Computational Linguistics.

Minghao Li, Yingxiu Zhao, Bowen Yu, Feifan Song, Hangyu Li, Haiyang Yu, Zhoujun Li, Fei Huang, and Yongbin Li. 2023a. Api-bank: A comprehensive benchmark for tool-augmented llms. *arXiv preprint arXiv:2304.08244*.

Yanyang Li, Jianqiao Zhao, Duo Zheng, Zi-Yuan Hu, Zhi Chen, Xiaohui Su, Yongfeng Huang, Shijia Huang, Dahua Lin, Michael Lyu, and 1 others. 2023b. Cleva: Chinese language models evaluation platform. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 186–217.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, and 1 others. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, Xiaodong Fan, Ruofei Zhang, Rahul Agrawal, Edward Cui, Sining Wei, Taroon Bharti, Ying Qiao, Jiun-Hung Chen, Winnie Wu, and 5 others. 2020. Xglue: A new benchmark dataset for cross-lingual pre-training, understanding and generation. *Preprint*, arXiv:2004.01401.

Chin-Yew Lin and Eduard Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 human language technology conference of the North American chapter of the association for computational linguistics*, pages 150–157.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022a. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 3214–3252.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022b. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

Auren R Liu, Pat Pataranutaporn, and Pattie Maes. 2024. Chatbot companionship: a mixed-methods study of companion chatbot usage patterns and their relationship to loneliness in active users. *arXiv preprint arXiv:2410.21596*.

Chia-Wei Liu, Ryan Lowe, Iulian V Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132.

Chuang Liu, Renren Jin, Yuqi Ren, Linhao Yu, Tianyu Dong, Xiaohan Peng, Shuting Zhang, Jianxiang Peng, Peiyi Zhang, Qingqing Lyu, and 1 others. 2023a. M3ke: A massive multi-level multi-subject knowledge evaluation benchmark for chinese large language models. *arXiv preprint arXiv:2305.10263*.

Han Liu, Xiaojun Wang, Qingyu Zhu, Long Yu, Wen-Bin Xie, and Chen Lu. 2023b. VCSum: A versatile Chinese meeting summarization dataset. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8914–8932, Toronto, Canada. Association for Computational Linguistics.

Hao Liu, Yinhe Zheng, Yelong Feng, Jianfeng Gao, Yan Wang, and Ming Zhou. 2021. Towards emotional support dialog systems. 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 6397–6409, Online. Association for Computational Linguistics.

Jian Liu, Cui Leyang, Han Liu, Dandan Huang, Yile Luan, and Yelong Zhou. 2020. LogiQA: A challenge dataset for machine reading comprehension with logical reasoning. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, pages 3932–3939. International Joint Conferences on Artificial Intelligence Organization.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, and 1 others. 2023c. Agentbench: Evaluating llms as agents. *arXiv preprint arXiv:2308.03688*.

Xin Liu, Qingcai Chen, Chong Deng, Huajun Zeng, Jing Chen, Yang Li, and Xunying Huang. 2018. LCQMC: A large-scale Chinese question matching corpus. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1954–1962, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Andrew Maas, Raymond Daly, Peter Pham, Dan Huang, Andrew Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. 2024. Evaluating very long-term conversational memory of llm agents. *arXiv preprint arXiv:2402.17753*.

Shikib Mehri, Mihail Eric, and Dilek Hakkani-Tur. 2020. Dialoglue: A natural language understanding benchmark for task-oriented dialogue. *arXiv preprint arXiv:2009.13570*.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. SemEval-2018 task 1: Affect in tweets. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 1–20, New Orleans, Louisiana. Association for Computational Linguistics.

Aidar Myrzakhan, Sondos Mahmoud Bsharat, and Zhiqiang Shen. 2024. Open-llm-leaderboard: From multi-choice to open-style questions for llms evaluation, benchmark, and arena. *arXiv preprint arXiv:2406.07545*.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. 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 5356–5371, Online. Association for Computational Linguistics.

Peggy ML Ng, Calvin Wan, Daisy Lee, Irene Garnelo-Gomez, and Mei Mei Lau. 2025. I love you, my ai companion! do you? perspectives from the triangular theory of love and attachment theory. *Internet Research*, pages 1–21.

Jiao Ou, Junda Lu, Che Liu, Yihong Tang, Fuzheng Zhang, Di Zhang, and Kun Gai. 2024. Dialogbench: Evaluating llms as human-like dialogue systems. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6137–6170.

Samuel J Paech. 2023. Eq-bench: An emotional intelligence benchmark for large language models. *arXiv preprint arXiv:2312.06281*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Alicia Parrish, Nikita Nangia, Angelica Chen, Kawin Ethayarajh, Anjali Ganguli, Pawan Goyal, Amanda Jones, Sida Kari, Ngan Nguyen, Vishakh Padmakumar, Johnny Thompson, Debajyoti Das, and Samuel R. Bowman. 2022. BBQ: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2086–2102, Dublin, Ireland. Association for Computational Linguistics.

Mohammad Taher Pilehvar and Jose Camacho-Collados. 2019. WiC: the word-in-context dataset for evaluating context-sensitive meaning representations. 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 1267–1273, Minneapolis, Minnesota. Association for Computational Linguistics.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. Meld: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 527–536.

Soujanya Poria, Navonil Majumder, Devamanyu Hazarika, Deepanway Ghosal, Rishabh Bhardwaj, Samson Yu Bai Jian, Pengfei Hong, Romila Ghosh, Abhinaba Roy, Niyati Chhaya, and 1 others. 2021. Recognizing emotion cause in conversations. *Cognitive Computation*, 13(5):1317–1332.

Hannah Rashkin, Eric M Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381.

ReLE Benchmark Team. 2025. Rele: Really reliable live evaluation for chinese llms.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *2011 AAAI spring symposium series*.

Sahand Sabour, Siyang Liu, Zheyuan Zhang, June M. Liu, Jinfeng Zhou, Alvionna S. Sunaryo, Tatia M. C. Lee, Rada Mihalcea, and Minlie Huang. 2024. Emobench: Evaluating the emotional intelligence of large language models.

D Slack, J Wang, D Semenenkov, and 1 others. 2024. A holistic approach for test and evaluation of large language models.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. 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, and 1 others. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on machine learning research*.

Maosong Sun, Jingyang Li, Zhiyuan Liu, and Yijie Sun. 2016. Thuctc: An efficient chinese text classifier. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC'16)*, Portorož, Slovenia.

Meiling Tao, Chenghao Zhu, Dongyi Ding, Tiannan Wang, Yuchen Eleanor Jiang, and Wangchunshu Zhou. 2024. Personafeedback: A large-scale human-annotated benchmark for personalization. *Preprint*, arXiv:2506.12915.

ModelScope Team. 2024. EvalScope: Evaluation framework for large models.

Cynthia Van Hee, Els Lefever, and Véronique Hoste. 2018. SemEval-2018 task 3: Irony detection in English tweets. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 39–50, New Orleans, Louisiana. Association for Computational Linguistics.

Oriol Vinyals and Quoc Le. 2015. A neural conversational model. *arXiv preprint arXiv:1506.05869*.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32.

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.

Chen Wang, Min Hu, Hangbo Zhao, Chen Gao, Weijing Wang, xiaozhong Gao, and Sujian Li. 2022. CNewSum: A large-scale Chinese news summarization dataset with human-annotated cross-media information. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4318–4331, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Hongru Wang, Lingzhi Wang, Yiming Du, Liang Chen, Jingyan Zhou, Yufei Wang, and Kam-Fai Wong. 2023. A survey of the evolution of language model-based dialogue systems. *arXiv preprint arXiv:2311.16789*.

Song Wang, Peng Wang, Tong Zhou, Yushun Dong, Zhen Tan, and Jundong Li. 2024. Ceb: Compositional evaluation benchmark for fairness in large language models. *arXiv preprint arXiv:2407.02408*.

Anuradha Welivita, Yubo Xie, and Pearl Pu. 2021. A large-scale dataset for empathetic response generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1251–1264.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Liang Xu, Hai Hu, Xuanwei Zhang, Lu Li, Chenjie Cao, Yudong Li, Yechen Xu, Cong Sun, Dian Tian, Lian Qian, and 1 others. 2020a. CLUE: A Chinese language understanding evaluation benchmark. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4762–4772, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Liang Xu, Hai Hu, Xuanwei Zhang, Lu Li, Chenjie Cao, Yudong Li, Yechen Xu, Kai Sun, Dian Yu, Cong Yu, and 1 others. 2020b. Clue: A chinese language understanding evaluation benchmark. *arXiv preprint arXiv:2004.05986*.

Liang Xu, Anqi Li, Lei Zhu, Hang Xue, Changtai Zhu, Kangkang Zhao, Haonan He, Xuanwei Zhang, Qiyue Kang, and Zhenzhong Lan. 2023. Superclue: A comprehensive chinese large language model benchmark. *arXiv preprint arXiv:2307.15020*.

Yangyang Xu, Jinpeng Hu, Zhuoer Zhao, Zhangling Duan, Xiao Sun, and Xun Yang. 2025. Multiagentesc: A llm-based multi-agent collaboration framework for emotional support conversation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 4665–4681.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. In *The eleventh international conference on learning representations*.

Yuan Yao, Qingxiu Dong, Jian Guan, Boxi Cao, Zhengyan Zhang, Chaojun Xiao, Xiaozhi Wang, Fanchao Qi, Junwei Bao, Jinran Nie, and 1 others. 2021. Cuge: A chinese language understanding and generation evaluation benchmark. *arXiv preprint arXiv:2112.13610*.

Yuan YAO, Qingxiu DONG, Jian-Guo ZHANG, Zhipeng CHEN, Zhengyu NIU, Boxing CHEN, Yating ZHANG, Deyi XIONG, Kai-Fu LEE, Ee-Peng LIM, C. C. Jay KUO, Zhen-Huan HWANG, Qing-Fu ZENG, Buzhou TANG, and Chengqing ZONG. 2023. CUGE: A Chinese language understanding and generation evaluation benchmark. In *Advances in Neural Information Processing Systems*.

Xiaohan Yuan, Jinfeng Li, Dongxia Wang, Yuefeng Chen, Xiaofeng Mao, Longtao Huang, Jialuo Chen, Hui Xue, Xiaoxia Liu, Wenhai Wang, and 1 others. 2025. S-eval: Towards automated and comprehensive safety evaluation for large language models. *Proceedings of the ACM on Software Engineering*, 2(ISSTA):2136–2157.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

Hui Zeng. 2023. Measuring massive multitask chinese understanding. *arXiv preprint arXiv:2304.12986*.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018a. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213.

Sheng Zhang, Xiaodong Liu, Jingjing Liu, Jianfeng Gao, Kevin Duh, and Benjamin Van Durme. 2018b. Record: Bridging the gap between human and machine commonsense reading comprehension. *arXiv preprint arXiv:1810.12887*.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in neural information processing systems 28*.

Yutong Zhang, Dora Zhao, Jeffrey T Hancock, Robert Kraut, and Diyi Yang. 2025. The rise of ai companions: How human-chatbot relationships influence well-being. *arXiv preprint arXiv:2506.12605*.

Zhexin Zhang, Leqi Lei, Lindong Wu, Rui Sun, Yongkang Huang, Chong Long, Xiao Liu, Xuanyu Lei, Jie Tang, and Minlie Huang. 2024. Safetybench: Evaluating the safety of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15537–15553.

Zhexin Zhang, Leqi Lei, Lindong Wu, Rui Sun, Yongkang Liu, Jiaan Wang, Zekun Li, Yuxiao Wang, Jiao Xue, Yifan Gong, Yujiu Yang, Lichao Sun, Chao Shen, Jing Shao, Yuhang Wang, Yequan Wang, Fangkai Jiao, Yangsheng Zhang, Junchi Yan, and 10 others. 2023. SafetyBench: A comprehensive benchmark for evaluating the safety of large language models. *Preprint*, arXiv:2309.09686.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zhuohan Lin, Zi Li, Daniel Brooks, Joseph Gonzalez, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2024. Agieval: A human-centric benchmark for evaluating foundation models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2299–2314.

Wanjun Zhong, Ruixiang Cui, Yidong Wang, Zhaoran Wang, Jian-Guang Lou, Bora Uçar, Ruixuan Li, Jialiang Tang, Weizhu Chen, and Jindong Wang. 2023. AGIEval: A human-centric benchmark for evaluating foundation models. In *Advances in Neural Information Processing Systems*.

Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. 2018. Emotional chatting machine: Emotional conversation generation with internal and external memory. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Kaijie Zhu, Jindong Wang, Jiaheng Zhou, Zichen Wang, Hao Chen, Yidong Wang, Linyi Yang, Wei Ye, Neil Zhenqiang Gong, Yue Zhang, and 1 others. 2023a. Promptbench: Towards evaluating the robustness of large language models on adversarial prompts. *arXiv preprint arXiv:2306.04528*.

Kaijie Zhu, Qinlin Zhao, Hao Chen, Jindong Wang, and Xing Xie. 2023b. Promptbench: A unified library for evaluation of large language models. *arXiv preprint arXiv:2312.07910*.