Towards Evaluation of Language Models with Skill Dimensions: A Case Study on Narrative Question Answering

Emil Kalbaliyev
Institute of Computer Science
University of Tartu
Tartu, Estonia
emil.kalbaliyev@ut.ee

Kairit Sirts
Institute of Computer Science
University of Tartu
Tartu, Estonia
kairit.sirts@ut.ee

Abstract

Large language models have demonstrated varying levels of competence across a range of reasoning tasks, but coarse-grained evaluations often do not reflect their specific strengths and weaknesses, particularly in complex tasks such as Narrative Question Answering. In this paper, we advocate for a multi-dimensional skill-based evaluation that assesses models across distinct core skill dimensions. Our proposed skill-focused evaluation framework offers a granular and more realistic measure of model performance, revealing targeted areas for improvement and guiding future development. Experiments on Narrative Question Answering demonstrate that dimension-level analysis captures the multifaceted nature of the task and informs more effective model evaluation.

1 Introduction

Large language models (LLMs) have achieved impressive results across a variety of reasoning tasks. However, current evaluation practices predominantly rely on coarse-grained evaluation that aggregates performance into a single score that does not reflect the strengths and weaknesses of models across different reasoning skills. This limitation poses a significant challenge: without a detailed understanding of where models excel or struggle on a task, it becomes difficult to identify targeted areas for improvement or to accurately estimate their readiness for real-world applications.

Earlier works on fine-grained skill evaluation (Sugawara et al., 2017a,b) have focused primarily on challenging skills. While this offers useful insight into models' upper limits, it similarly captures only part of the broader skill landscape, leaving room for more comprehensive approaches that consider the full spectrum of reasoning abilities. Ye et al. (2024) makes progress toward addressing issues in coarse-grained evaluation by selecting and evaluating a subset of fine-grained skills drawn from an existing skill taxonomy (Rogers et al., 2023). This approach enhances the granularity of evaluation by focusing on specific, essential top skills. However, it does not attempt to cover the full range of skills that might be involved in answering each question. As a result, the evaluation may overlook other relevant skills, providing only a partial view of model capabilities.

To fill this gap, we propose a dimensional skill evaluation framework that systematically assesses model performance across distinct core skill dimensions. While earlier works have focused on defining skill dimensions (Rogers et al., 2023; Schlegel et al., 2020; Kalbaliyev and Sirts, 2024), they have not leveraged these dimensions to structure evaluation in a way that more accurately reflects model capabilities. In contrast, our framework introduces a three-level evaluation approach: (1) skill-level evaluation, (2) dimension-level evaluation, and (3) multi-dimensional evaluation. This structure ensures that each skill and dimension contributes proportionally to the final assessment. By explicitly accounting for both skill-level and dimension-level variation, our framework addresses challenges such as skill imbalance and the diverse range of reasoning abilities required in complex question answering tasks.

We validate the effectiveness of this skill-focused framework through evaluation of LLMs on the Narrative Question Answering task. The results demonstrate that dimensional evaluation captures the multifaceted nature of the task and reveals nuanced insights, such as uneven improvements across skill dimensions when scaling model size, and highlights developmental priorities masked by coarse-grained scores. Overall, the proposed framework offers a more granular and actionable measure of LLM performance, guiding future model development and deployment for complex reasoning tasks.

*Proceedings of the 14th Joint Conference on Lexical and Computational Semantics (*SEM 2025)*, pages 430–440

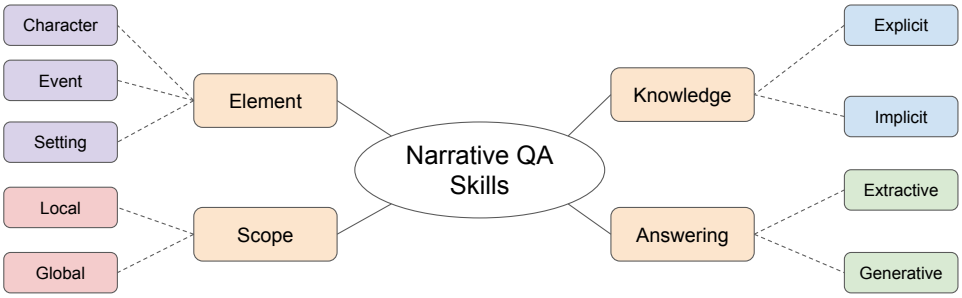

Figure 1: Narrative Question Answering Skills based on Kalbaliyev and Sirts (2024).

2 Background

2.1 Fine-grained Skill Evaluation

In the context of question answering (QA), skills can be understood as learned patterns that enable the system to comprehend and generate accurate responses. Evaluating QA through skills provides more diagnostic insight than overall coarse-grained metrics, as it reveals which reasoning abilities a model possesses or lacks.

Fine-grained skill evaluation involves several stages. First, a skill taxonomy is either defined or adopted to capture the range of abilities relevant to the target QA domain. Then, questions are annotated based on skill taxonomy with the skills they require, either via experts, crowdworkers, automated heuristics, or LLMs. Lastly, model performance is analyzed per skill, revealing gaps in skills and informing targeted improvements.

Skill definition plays a crucial role in determining how deep and precise an evaluation can be. In traditional fine-grained skill evaluations for QA, the focus is often limited to a set of selectively chosen skills. This selective approach can overlook important competencies, leading to incomplete assessments. In contrast, a dimensional skill perspective defines skills along well-structured dimensions, each representing a high-level category of related competencies that are essential for QA. By organizing skills into dimensions, this approach ensures that all core competencies are systematically captured. While prior work has proposed various dimensional skill taxonomies (Rogers et al., 2023; Schlegel et al., 2020; Kalbaliyev and Sirts, 2024), skill dimensions have not been used to structure evaluations. In the following subsection, we highlight a dimensional skill taxonomy for Narrative Question Answering, both to illustrate how such skills are defined and to provide the foundation for our case study evaluation.

2.2 Narrative QA Skill Taxonomy

As depicted in Figure 1, Narrative QA skills can be categorized into four core skill dimensions, each encompassing distinct skills, with every question attributable to one specific skill within each dimension (Kalbaliyev and Sirts, 2024). In the following paragraphs, we briefly review each dimension and its associated skills. The taxonomy contains four skill dimensions: the first three related to narrative understanding, and the last one focused on question answering.

Element Dimension. Narrative understanding requires comprehending core narrative elements, character, event, and setting, which are also the primary focus of Narrative Question Answering. ***Character*** questions examine identities, traits, and relationships of story figures. ***Event*** questions focus on what happens in the story and how events relate to each other. ***Setting*** questions address the time, place, and environment in which the narrative unfolds. This dimension enables an element-wise evaluation of a model's comprehension of narrative context.

Scope Dimension. Answering questions based on a story requires forming adequate narrative representations, which vary in scope depending on the extent of the text needed. This scope also determines the type of reasoning skill required. ***Local*** questions pertain to a specific part of the story and require making local inferences, while ***Global*** questions span multiple parts of the story and require broader comprehension, synthesis, or summarization of information across the multiple parts.

Knowledge Dimension. Another key comprehension skill dimension is a model's ability to understand both explicitly stated and implied information in a narrative. ***Explicit*** questions can be answered directly using clearly presented information from the text, while ***Implicit*** questions require the model to infer unstated information using com-

monsense knowledge and read between the lines.

Answering Dimension. A crucial skill dimension in Narrative Question Answering is the ability to express understanding through sufficient answer formulation. ***Extractive*** questions can be adequately answered by identifying and using spans from the text, while ***Generative*** questions require the model to generate additional words or phrases to form a complete answer or enhance extracted spans.

3 Dimensional Skill Evaluation

Traditional evaluation in QA often provides a single coarse-grained score that aggregates performance across all of the questions. However, these scores do not account for different skills and skill dimensions, potentially masking a model's strengths and weaknesses in specific skills. To address this, we propose a dimensional skill evaluation framework that assesses the model at three levels: skill-level, dimension-level, and multi-dimensional evaluations.

3.1 Skill-level Evaluation

Each question $q \in Q$ is associated with a specific skill $S_{i,j}$, where $S_{i,j}$ denotes the j-th skill in the i-th skill dimension D_i. Let $Q_{i,j}$ denote the set of questions associated with skill $S_{i,j}$.

The performance of the model on a single skill $S_{i,j}$ is evaluated using a base metric m (e.g., accuracy), defined as:

$$m_{i,j} = \frac{1}{|Q_{i,j}|} \sum_{q \in Q_{i,j}} \text{Score}(q) \tag{1}$$

where $\text{Score}(q)$ is the model's score for question q based on the chosen evaluation metric.

3.2 Dimension-level Evaluation

A skill dimension D_i consists of K_i skills, $D_i = \{S_{i,1}, S_{i,2}, \ldots, S_{i,K_i}\}$. The objective in this evaluation is to assess the model's performance on the level of skill dimension.

Let $w_{i,j}$ denote the weight assigned to j-th skill in i-th dimension D_i, with following condition:

$$\sum_{j=1}^{K_i} w_{i,j} = 1 \tag{2}$$

The performance of the model on each dimension D_i is computed as:

$$M_i = \sum_{j=1}^{K_i} w_{i,j} \cdot m_{i,j} \tag{3}$$

By default, all skills within a dimension are equally weighted as:

$$w_{i,j} = \frac{1}{K_i} \tag{4}$$

This ensures that each skill within the dimension contributes equally to the dimension-level performance, regardless of the number of questions associated with each skill. However, if there is a significant skill imbalance within a dimension, the weights can be adjusted to account for it.

3.3 Multi-Dimensional Evaluation

The multi-dimensional evaluation aggregates the performance across all N skill dimensions, $\mathcal{D} = \{D_1, D_2, \ldots, D_N\}$. The objective is to ensure that each dimension contributes equally to the overall task-level coarse-grained evaluation.

Let v_i denote the weight assigned to dimension D_i. If all dimensions are to be equally weighted, we define:

$$v_i = \frac{1}{N} \tag{5}$$

Alternatively, if the dimensions are to be weighted differently due to domain/dimension importance, the weights v_i can be adjusted accordingly, ensuring that:

$$\sum_{i=1}^{N} v_i = 1 \tag{6}$$

The overall multi-dimensional skill-balanced evaluation metric M is then calculated as:

$$M = \sum_{i=1}^{N} v_i \cdot M_i \tag{7}$$

Expanding M_i, we can express M as:

$$M = \sum_{i=1}^{N} v_i \cdot \left(\sum_{j=1}^{K_i} w_{i,j} \cdot m_{i,j} \right) \tag{8}$$

The proposed evaluation framework ensures that the model's performance is assessed not only at the skill-balanced coarse-grained level but also at the skill and dimension levels. This provides a more nuanced understanding of the model's capabilities, enabling targeted improvements and better analysis of model behavior.

	Models	Llama-3			Gemma-3			GPT-4o
Dimension	Skills/Size	1B	3B	8B	1b	4b	12b	-
Element	Character	62.83	86.48	88.83	79.92	85.18	87.13	91.82
	Event	56.79	83.07	86.45	67.18	82.23	83.07	91.43
	Setting	74.20	90.14	94.20	79.13	92.75	95.65	96.81
	Dim	***64.61***	***86.56***	***89.83***	***75.41***	***86.72***	***88.62***	***93.36***
Scope	Local	64.03	88.01	91.67	74.03	87.50	89.66	94.27
	Global	38.91	68.09	69.07	58.03	66.45	63.61	81.20
	Dim	***51.47***	***78.05***	***80.37***	***66.03***	***76.97***	***76.63***	***87.74***
Knowledge	Explicit	67.48	90.08	92.49	76.50	88.85	89.64	95.18
	Implicit	39.93	70.51	75.56	58.02	71.06	73.45	83.89
	Dim	***53.71***	***80.30***	***84.03***	***67.26***	***79.95***	***81.54***	***89.54***
Answering	Extractive	63.92	88.14	90.80	74.70	86.68	88.57	94.52
	Generative	42.65	70.24	75.36	57.63	72.32	71.18	81.99
	Dim	***53.29***	***79.19***	***83.08***	***66.16***	***79.50***	***79.88***	***88.26***
All	***Multi-Dim***	***55.77***	***81.02***	***84.33***	***68.72***	***80.79***	***81.67***	***89.72***
Coarse-grained		59.46	84.39	87.57	71.12	83.67	84.93	91.90

Table 1: LLM-as-a-Judge evaluation of Large Language Models on the test of the FairytaleQA dataset. Dimension column indicates the sections of the table corresponding to specific skill dimensions. Skills show the skill-level evaluation of models. ***Dim*** refers to results of dimension-level evaluation, while ***Multi-Dim*** represents a multi-dimensional evaluation of the task. The last row provides a single coarse-grained evaluation.

4 Evaluation Setup

We evaluate the Narrative Question Answering abilities of Large Language Models (LLMs) on the test set of the FairytaleQA (Xu et al., 2022) dataset. The test set contains 1,007 questions. We automatically annotated the questions based on skill taxonomy by Kalbaliyev and Sirts (2024). More details on the annotation process can be found in Appendix A.

For evaluation, we choose LLMs with different parameter sizes: 1B, 3B, and 8B instruction-tuned variants of the Llama-3 (Grattafiori et al., 2024) and 1B, 4B, and 12B instruction-tuned variants of Gemma-3 (Team et al., 2025) models, as well as the GPT-4o (OpenAI et al., 2024). We use Flow Judge v0.1 (FlowAI, 2024), an LLM-as-a-Judge model, to evaluate models' predictions. We report the average LLM-as-a-Judge-based accuracy score of 5 model runs. Dimensional Skill Evaluation is conducted with equal skill and dimension weights. Additional details on the evaluation setup can be found in Appendix B.

5 Results and Discussion

Skill-level, dimension-level, multi-dimensional, and coarse-grained evaluations of LLMs on FairytaleQA are presented in Table 1.

Skills within dimensions provide a good comparison point. Analyzing skill-level performance within a dimension, such as skill-level results in the scope dimension section of Table 1, allows for more meaningful comparisons between models or model variants, revealing trade-offs or uneven gains that overall metrics, such as coarse-grained results in the last column of Table 1, might mask. From the LLM-as-a-Judge-based results of Gemma-3 models on the FairytaleQA test set, we can observe an overall coarse-grained performance improvement from the 4B to the 12B model. However, disaggregating results by the Scope dimension reveals that this improvement is primarily driven by gains on local questions, while performance on global questions slightly declines. This finding suggests that increasing model size does not uniformly enhance all reasoning capabilities, showing the advantage of dimensional skill-level evaluation.

Dimension-level scores could indicate development priorities. The aggregated dimension-level scores (denoted as ***Dim*** in the Table 1) provide a high-level summary of model performance across different skill areas. These scores highlight which dimensions excel and which ones lag behind. Notably, the Element dimension consistently shows stronger performance compared to others, suggest-

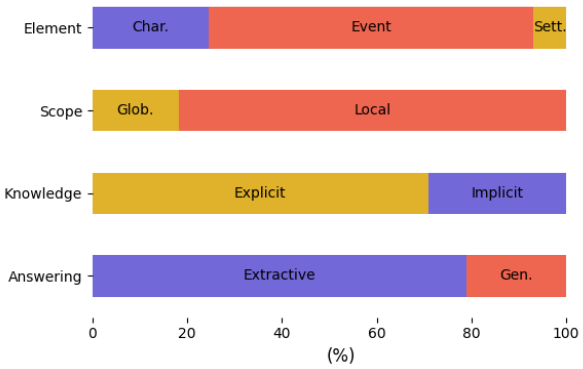

Figure 2: Skill imbalance in the test set of FairytaleQA

ing that models are better at recognizing common skill patterns within this dimension. While previous research (Bao et al., 2023; Peng et al., 2023) has primarily focused on enhancing the reasoning skills of language models by adding skill representations along a dimension similar to the Element dimension, our findings suggest that prioritizing the improvement of other underperforming dimensions could better address the models' developmental needs.

Multi-dimensional aggregates provide more realistic performance estimates. When aggregating performance across all dimensions (denoted as ***Multi-Dim*** in the Table), models achieve lower scores compared to evaluations based on a single coarse-grained metric (last row in the Table). For example, GPT-4o scores 91.90 based on the coarse-grained metric but only 89.72 under the multi-dimensional evaluation in the Table 1. This gap is even larger for smaller models like LLaMA-3 1B, which scores 59.46 in coarse-grained versus 55.77 in the multi-dimensional evaluation setting.

This discrepancy highlights that coarse-grained evaluations tend to overestimate model capabilities by masking weaknesses in specific skills. Moreover, skill imbalance across dimensions, such as in FairytaleQA shown in Figure 2, can further distort these coarse-grained metrics, as models may perform well on overrepresented or easier skills, inflating overall scores. The multi-dimensional aggregate, by capturing performance variability across underrepresented and challenging skill dimensions and skills, provides a more nuanced and accurate assessment of overall task performance. Consequently, multi-dimensional evaluation offers a more representative measure of model readiness for complex, real-world applications.

6 Related Works

Dimension-based evaluation has been explored in other NLP tasks such as summarization and open-ended dialogue, where metrics are typically defined by dimensions. For example, summarization is often assessed in terms of consistency, relevance, fluency, and coherence (Jain et al., 2023), while open-domain dialogue is evaluated along dimensions such as appropriateness, content, grammar, and relevance (Lin and Chen, 2023). In contrast, fine-grained evaluation of question answering has largely focused on selected skills (Sugawara et al., 2017a,b; Ye et al., 2024). Although these skills can be grouped into broader dimensions that reflect higher-level competencies, prior work has typically discussed such dimensions only at a conceptual level, without explicitly structuring evaluation or analysis around them (Schlegel et al., 2020; Rogers et al., 2023; Kalbaliyev and Sirts, 2024). Our work moves beyond selective skill-level assessment to demonstrate how organizing evaluation around dimensions yields clearer insights into model behavior and capabilities.

Another line of research focuses on analyzing the instance-level complexity of benchmarks, either to gain a deeper understanding of the skills being evaluated or to identify informative subsets of examples that better capture task diversity and LLM performance (Rodriguez et al., 2021; Ye et al., 2023; Gor et al., 2024; Cook et al., 2025). While our work emphasizes the importance of evaluating models across all core skills and dimensions, we leave for future research the exploration of how instance-level complexity analysis can be combined with dimensional skill evaluation.

7 Conclusion

This study argues that a dimension-focused skill evaluation offers a more accurate and insightful assessment of large language models on complex tasks like Narrative Question Answering. Unlike coarse-grained evaluation, this approach uncovers specific strengths and weaknesses across skill dimensions, revealing that model improvements are often uneven and concentrated in certain areas. The findings emphasize the limitations of coarse-grained evaluation and advocate for dimension-level analysis to guide model development priorities and better reflect real-world performance readiness.

Limitations

In this paper, we focus our analysis on a single skill taxonomy in order to clearly demonstrate the applicability of skill dimensions. While our evaluation method is inherently flexible and can be adapted to alternative taxonomies with similar structural properties, its reliance on the availability of a well-defined taxonomy poses a limitation. Applying the framework to new tasks may require the design or refinement of task-specific taxonomies, which can be both time-consuming and non-trivial. In particular, ensuring consistency, coverage, and granularity across different domains could introduce additional challenges.

Another limitation is related to the skill annotation. The process of labeling questions according to skill taxonomy inevitably depends on the expertise, perspective, and assumptions of the annotator. While the dimensional skill taxonomy is conceptually differentiable across multiple dimensions, the boundaries between individual skills within each dimension may be interpreted differently depending on the annotator. For example, determining whether a question requires implicit versus explicit reasoning often hinges on how much background knowledge the annotator, a large language model in our case, considers "commonsense." This introduces a degree of subjectivity and potential variability in the labeling process. Despite these limitations, the skill taxonomies provide a structured framework for categorizing questions and allow for systematic analysis of model performance across a range of skill types. Consequently, even with inherent labeling variability, it offers valuable insights into the strengths and weaknesses of models in handling diverse reasoning and knowledge-intensive tasks.

Additionally, since the primary objective of this work is to introduce and justify a dimensional skill evaluation methodology, we do not use computationally demanding, larger language models or reasoning models.

Acknowledgments

This work was supported by the Estonian Research Council Grant PSG721.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, and 110 others. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *Preprint*, arXiv:2404.14219.

Meikai Bao, Qi Liu, Kai Zhang, Ye Liu, Linan Yue, Longfei Li, and Jun Zhou. 2023. Keep skills in mind: Understanding and implementing skills in commonsense question answering. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, pages 5012–5020. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. 2021. Evaluation of text generation: A survey. *Preprint*, arXiv:2006.14799.

Ryan A. Cook, John P. Lalor, and Ahmed Abbasi. 2025. No simple answer to data complexity: An examination of instance-level complexity metrics for classification tasks. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2553–2573, Albuquerque, New Mexico. Association for Computational Linguistics.

FlowAI. 2024. Flow Judge: An Open Small Language Model for LLM System Evaluations.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, and 5 others. 2024. The language model evaluation harness.

Albert Gatt and Emiel Krahmer. 2018. Survey of the state of the art in natural language generation: core tasks, applications and evaluation. *J. Artif. Int. Res.*, 61(1):65–170.

Maharshi Gor, Hal Daumé Iii, Tianyi Zhou, and Jordan Lee Boyd-Graber. 2024. Do great minds think alike? investigating human-AI complementarity in question answering with CAIMIRA. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21533–21564, Miami, Florida, USA. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Sameer Jain, Vaishakh Keshava, Swarnashree Mysore Sathyendra, Patrick Fernandes, Pengfei Liu, Graham Neubig, and Chunting Zhou. 2023. Multi-dimensional evaluation of text summarization with in-context learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8487–8495, Toronto, Canada. Association for Computational Linguistics.

Emil Kalbaliyev and Kairit Sirts. 2022. Narrative why-question answering: A review of challenges and datasets. In *Proceedings of the 2nd Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 520–530, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Emil Kalbaliyev and Kairit Sirts. 2024. On narrative question answering skills. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 814–820, Mexico City, Mexico. Association for Computational Linguistics.

Yash Kumar Lal, Nathanael Chambers, Raymond Mooney, and Niranjan Balasubramanian. 2021. TellMeWhy: A dataset for answering why-questions in narratives. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 596–610, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yen-Ting Lin and Yun-Nung Chen. 2023. LLM-eval: Unified multi-dimensional automatic evaluation for open-domain conversations with large language models. In *Proceedings of the 5th Workshop on NLP for Conversational AI (NLP4ConvAI 2023)*, pages 47–58, Toronto, Canada. Association for Computational Linguistics.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, and 401 others. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Alison H. Paris and Scott G. Paris. 2003. Assessing narrative comprehension in young children. *Reading Research Quarterly*, 38(1):36–76.

Wei Peng, Wanshui Li, and Yue Hu. 2023. Leader-generator net: Dividing skill and implicitness for conquering fairytaleqa. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, page 791–801, New York, NY, USA. Association for Computing Machinery.

Pedro Rodriguez, Joe Barrow, Alexander Miserlis Hoyle, John P. Lalor, Robin Jia, and Jordan Boyd-Graber. 2021. Evaluation examples are not equally informative: How should that change NLP leaderboards? 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 4486–4503, Online. Association for Computational Linguistics.

Anna Rogers, Matt Gardner, and Isabelle Augenstein. 2023. Qa dataset explosion: A taxonomy of nlp resources for question answering and reading comprehension. *ACM Comput. Surv.*, 55(10).

Viktor Schlegel, Marco Valentino, Andre Freitas, Goran Nenadic, and Riza Batista-Navarro. 2020. A framework for evaluation of machine reading comprehension gold standards. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5359–5369, Marseille, France. European Language Resources Association.

Saku Sugawara, Yusuke Kido, Hikaru Yokono, and Akiko Aizawa. 2017a. Evaluation metrics for machine reading comprehension: Prerequisite skills and readability. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 806–817, Vancouver, Canada. Association for Computational Linguistics.

Saku Sugawara, Hikaru Yokono, and Akiko Aizawa. 2017b. Prerequisite skills for reading comprehension: Multi-perspective analysis of mctest datasets and systems. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1).

Kai Sun, Dian Yu, Jianshu Chen, Dong Yu, Yejin Choi, and Claire Cardie. 2019. DREAM: A challenge data set and models for dialogue-based reading comprehension. *Transactions of the Association for Computational Linguistics*, 7:217–231.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. 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 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Alon Talmor, Ori Yoran, Ronan Le Bras, Chandra Bhagavatula, Yoav Goldberg, Yejin Choi, and Jonathan Berant. 2021. Commonsenseqa 2.0: Exposing the limits of ai through gamification. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey

Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

University of Tartu. 2018. UT Rocket.

Leandro Von Werra, Lewis Tunstall, Abhishek Thakur, Sasha Luccioni, Tristan Thrush, Aleksandra Piktus, Felix Marty, Nazneen Rajani, Victor Mustar, and Helen Ngo. 2022. Evaluate & evaluation on the hub: Better best practices for data and model measurements. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 128–136, Abu Dhabi, UAE. Association for Computational Linguistics.

Ying Xu, Dakuo Wang, Mo Yu, Daniel Ritchie, Bingsheng Yao, Tongshuang Wu, Zheng Zhang, Toby Jia-Jun Li, Nora Bradford, Branda Sun, Tran Bao Hoang, Yisi Sang, Yufang Hou, Xiaojuan Ma, Diyi Yang, Nanyun Peng, Zhou Yu, and Mark Warschauer. 2022. Fantastic questions and where to find them: FairytaleQA – an authentic dataset for narrative comprehension. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 447–460, Dublin, Ireland. Association for Computational Linguistics.

Qinyuan Ye, Harvey Fu, Xiang Ren, and Robin Jia. 2023. How predictable are large language model capabilities? a case study on BIG-bench. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7493–7517, Singapore. Association for Computational Linguistics.

Seonghyeon Ye, Doyoung Kim, Sungdong Kim, Hyeonbin Hwang, Seungone Kim, Yongrae Jo, James Thorne, Juho Kim, and Minjoon Seo. 2024. FLASK: Fine-grained language model evaluation based on alignment skill sets. In *The Twelfth International Conference on Learning Representations*.

A Skills Annotation

As skills are categorized under distinct dimensions in the taxonomy proposed by Kalbaliyev and Sirts (2024), we classify questions separately based on each dimension. In the following subsections, we outline our chosen methods to facilitate efficient annotation.

A.1 Element Dimension Annotation

Xu et al. (2022) involves human annotators to label narrative elements from taxonomy of Paris and Paris (2003) for FairytaleQA dataset. Bao et al. (2023) use a set of common words or phrases, referred to as skill seeds, to automatically annotate questions in CommonsenseQA and CommonsenseQA 2.0 datasets (Talmor et al., 2019, 2021).

Since keyword-based annotation often fails to produce robust results, we opted to annotate the dataset along the element dimension by prompting a Large Language Model with element definitions. Specifically, we used GPT-4o (OpenAI et al., 2024) to classify the dataset. For this process, we provided the model with a System Prompt and passed both the context and the question as a User Prompt to guide the annotation.

System Prompt 1. Element Dimension

You are a helpful assistant. You will be provided with a context, and a question. Based on the definition below, determine if the question is Character, Event or Setting:

- It is **Event** question if it asks about an activity, an action, an event, or relationships between among events and characters, such as a reason. For example,'why', 'what happened?' questions are event questions.
- It is **Character** question if it directly asks about the identity, feeling or characteristics of the characters. For example, "who" is character question.
- It is **Setting** question if it asks about specific place, time, and environment in which the events take place. For example, "where" and "when" are setting questions.

For each question, provide a brief explanation of your reasoning and classify the question. Use the following format:
Explanation: <your explanation>
Classification: <Character or Event or Setting>

User Prompt 1. Annotation

CONTEXT: <context>
QUESTION: <question>

We generated responses via Azure OpenAI's API with the temperature value of 0. To validate our question annotations, we leveraged the existing annotated FairytaleQA dataset (Xu et al., 2022) as a reference point and approximately aligned its skill taxonomy with our chosen framework: Character (character and feeling), Event (causal relationship, outcome resolution, action), and Setting (setting). Using this alignment, the GPT-based annotation achieved an accuracy of 98% on the roughly mapped validation set from FairytaleQA.

A.2 Scope Dimension Annotation

To annotate questions along the scope dimension, it is important to first define what constitutes a story part. A story part can be interpreted as a sentence, a paragraph, a scene, or another unit of narrative segmentation. If sentences are used as the unit of

segmentation, then single-sentence reasoning can be considered local, while multi-sentence reasoning can be attributed to global. If paragraphs or scenes are used as the unit of segmentation, especially in longer narratives, then local questions would be those that can be answered within a single paragraph or scene, while global questions would require information spanning across multiple paragraphs or scenes. The appropriate level of granularity depends largely on the nature of the dataset and the average length of the narratives involved. As we are focusing on short-form narratives, we chose local/global reasoning at the sentence level.

We prompted GPT-4o (OpenAI et al., 2024) to annotate the dataset based on the scope of text required to answer each question. We used the following System Prompt and passed context, question, and correct answers as a user prompt to annotate the dataset.

System Prompt 2. Scope Dimension

You are a helpful assistant. You will be provided with a context, a question, and the correct answer. Based on the definitions below, classify the question as either local or global:

- **Local** questions need information from only a single sentence to be answered correctly.
- **Global** questions need information from multiple sentences or the whole context to be answered correctly.

For each question, provide a brief explanation of your reasoning and classify the question. Use the following format:
Explanation: <your explanation>
Classification: <Local or Global>

We generated responses via Azure OpenAI's API with the temperature value of 0. To validate the classification accuracy, we used the skill-annotated portion of the Dream (Sun et al., 2019) dataset and mapped its skill taxonomy to our selected taxonomy. The GPT-based annotation achieved an accuracy and macro-F1 score of 91.5.

A.3 Knowledge Dimension Annotation

Similar to our approach for scope annotation, we prompted GPT-4o to classify datasets based on the implicitness of the question. We deliberately avoided providing a definition of explicitness, as doing so led to confusion between the two classes. To perform the annotation, we used the following System Prompt and supplied the context, question, and correct answer in the User Prompt.

System Prompt 3. Knowledge Dimension

You are a helpful assistant. You will be provided with a context, a question, and the correct answer. Based on the definition below, determine if the question is implicit:

- **Implicit** questions are those where the reader must use commonsense (world) knowledge or read between the lines to answer the question.

For each question, provide a brief explanation of your reasoning and classify the question. Use the following format:
Explanation: <your explanation>
Implicit: <Yes or No>

User Prompt 2. Annotation

CONTEXT: <context>
QUESTION: <question>
CORRECT ANSWERS: <answers>

To assess the classification accuracy, we evaluated the model on the skill-annotated portion of the DREAM dataset (Sun et al., 2019), aligning its original skill taxonomy with our selected taxonomy. The GPT-based annotation achieved an accuracy of 82.7% and a macro F1 score of 81%. Considering the inherent subjectivity in interpreting question implicitness, we found this result to be reasonable.

A.4 Answering Dimension Annotation

In order to annotate questions in the answering dimension, we automatically lemmatized and lowercased the context, question, and answers, removing articles and punctuation from the lemmas. Lemmatization was performed using spaCy's en_core_web_sm model. The context and question were then combined to form the input. If any lemmatized answer was a subset of the input lemmas, the question was classified as Extractive; otherwise, it was classified as Generative.

B Additional Details on Evaluation setup

B.1 Dataset

We selected the test set from the FairytaleQA (Xu et al., 2022) for evaluation. The questions were annotated according to the skill taxonomy proposed by Kalbaliyev and Sirts (2024), following the procedure detailed in Appendix A.

B.2 Language Model Prompting

We performed zero-shot prompting with Llama-3.2-1B-Instruct, Llama-3.2-3B-Instruct, and Llama-3.1-8B-Instruct models (Grattafiori et al., 2024) and gemma-3-1b-it, gemma-3-4b-it, and gemma-3-12b-it models (Team et al., 2025) models, as well as the GPT-4o (OpenAI et al., 2024). For the Gemma-3 and Llama-3, we used the LM Evaluation Harness library (Gao et al., 2024) with models loaded from the Hugging Face Hub. We run inferences on Tesla V100 GPUs in the High Performance Computing Center of the University of Tartu (University of Tartu, 2018). For GPT-4o, we generated responses via Azure OpenAI's API. The following user prompt was used for question answering.

User Prompt 3. Question Answering Prompt

Answer the question based on the context. Keep your answer concise, few words are enough.
CONTEXT: <context>
QUESTION: <question>
ANSWER:

For all models, responses were generated using a temperature of 1.0 and a maximum new tokens value of 1024. The reported results are presented as the average of 5 model runs.

B.3 Metrics

We evaluated accuracy using an LLM-as-a-Judge to perform binary (Pass/Fail) assessments of generated outputs against reference answers. For this purpose, we used Flow Judge v0.1 (FlowAI, 2024), a specialized LLM-as-a-Judge model derived from further fine-tuning of the Phi-3.5-mini instruct model (Abdin et al., 2024). We selected Flow Judge v0.1 because of its compact size and performance comparable to the larger models, such as GPT-4o. This evaluation was conducted using the LM Evaluation Harness (Gao et al., 2024), with generation parameters used in the model's technical report: temperature of 0.1, Top P value of 0.95, and a maximum new tokens value of 1024. We used the following adapted scoring rubric:

- *Score 0 - Fail:* The generated response is completely incorrect or irrelevant to the query, with no overlap in information with any of the reference answers.
- *Score 1 - Pass:* The generated response matches one of the reference answers. The meaning conveyed by the generated response is equivalent to the reference. The generated response may leave out non-essential details compared to the references.

Additionally, we also evaluated the model with ROUGE-L (Lin, 2004) as an additional commonly used evaluation metric. We used the Hugging Face Evaluate library (Von Werra et al., 2022) implementation. The evaluation was conducted with the use_stemmer parameter set to True.

Language model performances based on LLM-as-a-Judge and ROUGE-L evaluation can be found in Tables 1 and 2.

C Results with ROUGE-L

Table 2 presents the ROUGE-L evaluation results. Although the scores differ from those obtained under the LLM-as-Judge setting, the main conclusions drawn from Table 1 remain valid. Analyzing performance at the skill level within each dimension, such as the element dimension section of Table 2, facilitates more fine-grained comparisons across models and model variants. This view makes it possible to identify trade-offs and uneven improvements that are often obscured when relying solely on aggregate metrics, such as those reported in the final column of Table 2.

Similarly, the element dimension consistently yields higher performance than other dimensions, suggesting that models are more adept at recognizing recurring skill patterns in this dimension. Additionally, with multi-dimensional evaluation, models achieve substantially lower scores compared to evaluations based only on a single coarse-grained metric. This discrepancy indicates that coarse-grained evaluations tend to overestimate model capabilities by concealing weaknesses in specific skills.

We note that the choice of metric for free-form evaluation can substantially influence the results. For instance, while LLM-as-a-judge-based evaluation of Gemma-3 models shows uneven gains in

	Models	Llama-3			Gemma-3			GPT-4o
Dimension	Skills/Size	1B	3B	8B	1b	4b	12b	-
Element	Character	36.76	55.27	62.68	54.62	60.47	65.58	66.86
	Event	36.06	49.36	55.27	37.66	42.53	44.96	60.09
	Setting	53.70	67.21	78.78	65.15	78.08	74.80	83.51
	Dim	***42.17***	***57.28***	***65.58***	***52.48***	***60.36***	***61.78***	***70.15***
Scope	Local	41.12	56.96	64.43	47.66	54.28	57.01	68.31
	Global	20.88	29.85	32.89	25.88	27.24	29.81	41.05
	Dim	***31.00***	***43.41***	***48.66***	***36.77***	***40.76***	***43.41***	***54.68***
Knowledge	Explicit	43.97	60.23	67.15	50.44	56.98	58.98	71.27
	Implicit	21.51	32.06	38.12	27.29	30.82	35.21	44.05
	Dim	***32.74***	***46.15***	***52.63***	***38.86***	***43.90***	***47.10***	***57.66***
Answering	Extractive	42.09	58.22	65.16	48.43	54.55	56.55	69.43
	Generative	19.90	28.71	34.34	25.85	29.81	35.16	40.43
	Dim	***31.00***	***43.47***	***49.75***	***37.14***	***42.18***	***45.85***	***54.93***
All	***Multi-Dim***	***34.23***	***47.58***	***54.16***	***41.31***	***46.80***	***49.54***	***59.36***
Coarse-grained		37.44	52.03	58.70	43.70	49.37	52.07	63.35

Table 2: Evaluation of Large Language Models on the FairytaleQA test set with the ROUGE-L metric. Dimension column indicates the sections of the table corresponding to specific skill dimensions. Skills show the skill-level evaluation of models. ***Dim*** refers to results of dimension-level evaluation, while ***Multi-Dim*** represents a multi-dimensional evaluation of the task. The last row provides a single coarse-grained evaluation.

the scope dimension, ROUGE-L scores indicate this difference in the element dimension. As LLM-as-a-judge offers greater flexibility in accommodating variations in model predictions, we report it as our main evaluation metric. Nevertheless, human evaluation remains the gold standard for text generation tasks, including Narrative QA (Celikyilmaz et al., 2021). However, conducting human evaluation is both costly and time-consuming (Lal et al., 2021), and its reliability has also been questioned (Gatt and Krahmer, 2018), particularly for QA instances that inherently involve multiple ambiguities (Kalbaliyev and Sirts, 2022).

D Data and Code Availability

The FairytaleQA dataset is publicly available on Hugging Face.[1] Code for annotation, inference, and evaluation is available on GitHub. [2]

[1] https://huggingface.co/datasets/WorkInTheDark/FairytaleQA

[2] https://github.com/EmilKalbaliyev/Dimensional-Skill-Evaluation