Evaluating Lexical Aspect with Large Language Models

Bolei Ma

LMU Munich & Munich Center for Machine Learning

bolei.ma@lmu.de

Abstract

In this study, we explore the proficiency of large language models (LLMs) in understanding two key lexical aspects: duration (durative/stative) and telicity (telic/atelic). Through experiments on datasets featuring sentences, verbs, and verb positions, we prompt the LLMs to identify aspectual features of verbs in sentences. Our findings reveal that certain LLMs, particularly those closed-source ones, are able to capture information on duration and telicity, albeit with some performance variations and weaker results compared to the baseline. By employing prompts at three levels (sentence-only, sentence with verb, and sentence with verb and its position), we demonstrate that integrating verb information generally enhances performance in aspectual feature recognition, though it introduces instability. We call for future research to look deeper into methods aimed at optimizing LLMs for aspectual feature comprehension.

1 Introduction

Aspect is a verbal category that is closely linked to concepts such as tense, temporality, verbal semantics, and quantification. In linguistics, aspect refers to different perspectives on the internal temporal constitution of a situation (Comrie, 1976; Leiss, 1992; Klein, 1994; Xiao and McEnery, 2004). There is two main sub groups of aspect, the grammatical aspect which refers to the verbal flexion in languages such as Slavic Languages, and the lexical aspect which contains the semantics of the event or state of a verb phrase situated in time.

In this paper, we focus on the lexical aspect with two important aspect features: duration and telicity. Duration (durative/stative) is the property of a verb or verb phrase that presents a state or an action, regardless of their endpoints. Durative aspect denotes the reading of an action, while stative aspect denotes the reading of a state. Telicity (telic/atelic) distinguishes between verbs that describe an action or event as having a specific endpoint. Telic aspect denotes the reading of the endpoint of an action or event, while atelic aspect denotes the reading of no endpoint. Table 1 shows examples for each feature in English.

Label	Sentence
durative	The boxer **is hitting** his opponent.
stative	Bread **consists** of flour, water and yeast.
telic	I **ate** a fish for lunch.
atelic	Cork **floats** on water.

Table 1: Examples of the two aspect features: duration (durative/stative) and telicity (telic/atelic) (Metheniti et al., 2022).

The identification of the aspectual features of the verbs in the sentence could be difficult as other verb categories or sentence elements such as tense, temporal adverbials, and context could affect the reading of the aspect (Zhang, 1995). Using computational models to identify the aspectual features could be therefore more challenging. There are various existing works on building datasets for lexical aspect and training models to classify the sentences in terms of their aspectual features (Friedrich and Palmer, 2014; Friedrich and Pinkal, 2015; Friedrich et al., 2016; Friedrich and Gateva, 2017; Kober et al., 2020; Metheniti et al., 2022). Nowadays, the vast expanse of LLMs has also opened the chance to study linguistics using LLMs (Opitz et al., 2024). Therefore, it is interesting to probe the proficiency of LLMs on aspectual features.

In this paper, based on a dataset on duration and telicity (Metheniti et al., 2022), we evaluate the ability of 6 different LLMs to identify the two aspectual features of sentences by zero-shot prompting the LLMs in three different levels: sentence-only, sentence with verb, and sentence with verb and its position. Our experimental results show that some LLMs are capable of capturing aspectual information, while there are some variations and

Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, pages 123–131

weaker performance compared to the fine-tuning baseline. In addition, adding verb information generally improves the prediction performance of LLMs. Overall, our study provides valuable insights into the challenges and opportunities in leveraging LLMs for evaluating lexical aspect.

2 Related Work

The evaluation and classification of aspectual features of verbs using NLP have been explored extensively in previous research. Siegel and McKeown (2000) are the first to employ supervised machine learning methods for aspectual classification.

Friedrich and Palmer (2014) introduced a semi-supervised approach that combined linguistic and distributional features to predict a verb's stativity/duration, also providing two annotated datasets for stativity. Furthermore, Friedrich and Pinkal (2015) focused on classifying clauses based on their aspectual properties, and expanded the scope to include situation entity types in Friedrich et al. (2016). Friedrich and Gateva (2017) contributed two English datasets with gold and silver annotations of telicity and duration, utilizing an L1-regularized multi-class logistic regression model.

Hermes et al. (2015) computationally modeled Vendler classes (Vendler, 1957) for 95 German verbs, combining distributional vectors with supervised classification. Additionally, Ramm et al. (2017) developed the first open-source tool for annotating morphosyntactic tense, mood, and voice for verbal complexes in multiple languages. Kober et al. (2020) introduced a dataset for tense and aspect concepts using natural language inference and proposed modeling aspect of English verbs in context using compositional distributional models.

In a more recent study by using a bunch of transformer-based models, Metheniti et al. (2022) conducted experiments on transformer models to identify aspectual features, revealing biases towards verb tense and word order. However, in the current era of the advances of LLMs, it is still unexplored whether the LLMs are able to capture the aspectual features.

A more detailed introduction to aspect concepts and their computational approaches can be found in this survey (Friedrich et al., 2023).

3 Experiments

Dataset. We use the dataset with telicity and duration-annotated sentences created by Metheniti et al. (2022). The dataset was built upon two previous datasets from Friedrich and Gateva (2017) and Alikhani and Stone (2019). It has two main subsets, one for duration and the other for telicity. Each subset contains sentences with the main verbs and their positions in sentences, as well as binary labels for durative ('1') or stative aspect ('0') in the duration subset, and telic ('1') or atelic ('0') aspect in the telicity subset. The label distribution in the test sets is presented in §A.1.

Prompt. Each question consists of a general instruction with a choice of answers (e.g. durative or stative) and the example sentence. We include the sentence, verb and verb information into the prompt. In addition, to test the robustness of the models as well as the ability of the models to comprehend the aspectual features both in the sentence level (without explicitly mentioning the verb) and the verb level (with explicitly mentioning the verb), we conduct the experiments in three different levels with different prompt formats. Table 2 shows the prompt formats of the three levels in the examples of duration subset with durative and stative aspects. In level 1, we only provide the sentence and ask for the aspect features. In level 2, we include the verb into the prompt. In level 3, we include the verb along with its position in the sentence into the prompt. The prompts are outlined in Table 2.

Models. We evaluate the aspect tasks with the following close- and open-source instruction-tuned LLMs: `GPT-3.5` (Brown et al., 2020) and `GPT-4` (OpenAI et al., 2024), `Llama-2-13b-chat-hf` and `Meta-Llama-3-8B-Instruct` (Touvron et al., 2023), `Gemma-7b-it` (Team et al., 2024), and `Mixtral-8x7B-Instruct-v0.1` (Jiang et al., 2024).

Baseline. We compare our zero-shot prompting of LLMs with the baselines of fine-tuning BERT-based models (Devlin et al., 2019) on the training data with and without adding information on the verb position as in Metheniti et al. (2022). We select the best performing model `bert-large-cased` in their work for fine-tuning as baseline.

LLM Output Extraction. Although we prompt the LLMs to give answers with single tokens of telic/atelic and durative/stative, in most cases, the LLMs respond with more tokens in different formats, and sometimes with explanation of their choices. We use a string matching method using `RegEx` to map the responses to the categories,

Level	Prompt
Level 1	`Does this sentence have durative aspect or stative aspect? Answer with durative or stative.\n Sentence:\n {sentence}`
Level 2	`Does the verb {verb} in this sentence have durative aspect or stative aspect? Answer with durative or stative.\n Sentence:\n {sentence}`
Level 3	`Does the verb {verb} in position {position} of this sentence have durative aspect or stative aspect? Answer with durative or stative.\n Sentence:\n {sentence}`

Table 2: Instruction prompt in three different constraint levels for the durative and stative aspects. Level 1 only shows the sentence, level 2 shows the sentence and the main verb of the sentence, level 3 shows the sentence, the main verb and its position of the sentence.

which is commonly used in extracting LLM outputs (Argyle et al., 2023). Afterwards, we manually evaluate the coded outputs and in case of uncertain responses, we note them accordingly.

4 Results

4.1 Main Results

We summarize the main results of the six LLMs in Table 3 on the duration test set and the Telicity test set, as well as the performance of the fine-tuned model `bert-large-cased` as the baseline for comparison.

On the duration test set, GPT-4 achieves the highest performance among the LLMs with an accuracy of 0.74 and an F1 score of 0.76. This is followed by GPT-3.5 and Llama-2, which both show comparable results in the 0.67 to 0.69 range for both metrics. The Llama-3 and Mixtral models also perform similarly with slightly lower scores. The Gemma model demonstrates the lowest performance among the LLMs with an accuracy of 0.54 and an F1 score of 0.42. Notably, the baseline large bert model significantly outperforms all LLMs, achieving an accuracy and F1 score of 0.96.

On the telicity test set, GPT-4 again leads among the LLMs with an accuracy of 0.71 and an F1 score of 0.72. GPT-3.5 and Llama-3 also show strong performances with scores around the 0.65 to 0.67 range for both metrics. The Mixtral model has slightly lower scores, and Llama-2 and Gemma exhibit the lowest performance. The fine-tuning bert-large-cased baseline still outperforms all LLMs.

We show that prompting LLMs to recognize the two aspectual features in verbs results in lower performance compared to the fine-tuning baseline, which exhibits high performance. This suggests that LLMs might lack the capability to probe the deep linguistic features of given words and may require adaptation (i.e., fine-tuning) to effectively perform the task. When comparing the two aspectual features, we observe that the performance of most models is slightly lower on the telicity test set than on the duration test set, indicating that recognizing a/telic aspects is more challenging. Additionally, among the LLMs, the closed-source models (GPT-3.5 and GPT-4) demonstrate better performance than the open-source models.

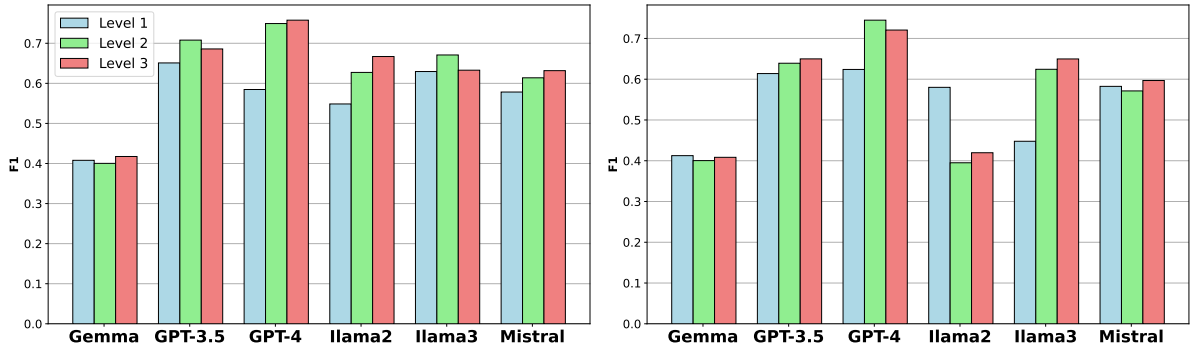
Model	Duration		Telicity	
	Acc	F1	Acc	F1
Gemma	0.54	0.42	0.52	0.41
GPT-3.5	0.68	0.69	0.67	0.65
GPT-4	0.74	0.76	0.71	0.72
Llama-2	0.67	0.67	0.53	0.42
Llama-3	0.64	0.63	0.65	0.65
Mixtral	0.62	0.63	0.59	0.60
`bert-large-cased`	0.96	0.96	0.88	0.87

Table 3: Accuracy and F1 scores for various zero-shot prompted LLMs vs. the fine-tuned baseline model `bert-large-cased` on duration and telicity test sets.

4.2 Verb and Verb Position Can Influence the Evaluation

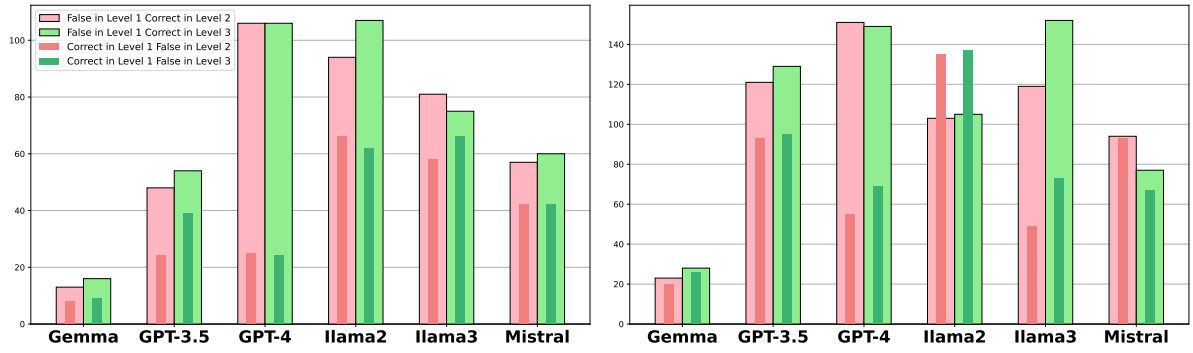
In this section, we analyze the impact of including the verb and its position in the sentence on the evaluation of aspectual features by LLMs in the duration and telicity test sets. we present F1 scores across three levels of prompting (sentence-only, sentence with verb, and sentence with verb and its position) in the bar plots in Figure 1 and they reveal significant insights partially.

In the duration set, Gemma's performance remains consistent across different levels of context, while GPT-3.5 and GPT-4 show substantial improvements with additional contextual information, although GPT-3.5 experiences a slight drop at the



(a) Duration (b) Telicity

Figure 1: F1 results of models in three different levels



(a) Duration (b) Telicity

Figure 2: Count of prediction differences in three different levels. The dark bars represent the count of correct predictions in level 1 which are falsely predicted in level 2&3. The light bars represent the count of false predictions in level 1 which are correctly predicted in level 2&3.

highest level. Llama-2 and Llama-3 generally benefit from more context, but Llama-3's performance slightly decreases at the highest level. Mistral demonstrates modest but consistent improvements. In the telicity set, Gemma again shows minimal variation, and GPT-3.5 and GPT-4 improve with more context, although GPT-4 has a slight decline at the highest level. Llama-2 struggles with additional context, while Llama-3 shows significant improvement. Mistral exhibits stable performance with slight gains.

We further examine the count of prediction differences across the three levels for both the duration and telicity sets in Figure 2. For the duration set, models like GPT-4 and GPT-3.5 showed substantial improvements with more corrections than regressions when additional context was provided, indicating a positive impact from including the verb and its position. For instance, GPT-4 had 106 corrections at both Levels 2 and 3, compared to 25 and 24 regressions. Llama-2 and Llama-3 also benefited from the additional context, though to a lesser extent., with Llama-2 showing 94 and 107 corrections versus 66 and 62 regressions. Mistral saw steady but modest improvements, while Gemma displayed minimal changes, suggesting limited sensitivity to the added information. On the telicity set, GPT-4 and GPT-3.5 again demonstrated significant improvements with more corrections (151 and 129) than regressions (55 and 95). However, Llama-2 struggled with additional context, showing more regressions than corrections, indicating potential confusion. Llama-3 and Mistral exhibited notable but smaller benefits from the detailed prompts.

Overall, our analysis underscores that while additional context including the verbs and verb positions generally enhances performance, the degree of improvement varies across models.

4.3 Error Analysis

In this section, we conduct a brief analysis of certain specific error predictions we observed during inference, which are outlined in Table 4.

Case 1 illustrates instances at prompt level 1,

Case, Input, Output
Case 1, Mistral on duration set, in level1 **Sentence:** Should be alright then . **Label:** durative **Output:** This sentence is stative in nature. It expresses a state of being or a condition, rather than an ongoing action. The word ""then"" suggests a conclusion or a result, which is also indicative of a stative aspect.
Case 2, Llama-2 on duration set, in level3 **Sentence:** Bake at 180 degrees of 40 - 50 minutes . Do n't think highly too that cheese was n't overroasted . **Label:** durative **Output:** The verb "think" in position 12 of the sentence does not have a durative or stative aspect. It is a stative aspect.
Case 3, Gemma on telicity set **Sentence:** I was born here , and I was lucky enough not to get attached to any particular river . **Label:** telic **Output (level1):** The sentence is telic because the verb ""was born"" is telic and the verb ""get attached"" is also telic. **Output (level2):** The verb ""attach"" in this sentence has telic aspect because the verb is completed with a specific object, which is ""river"". **Output (level3):** The verb attached in position 13 is ""attach"". The answer is atelic.
Case 4, GPT-4 on telicity set, level3 **Sentence:** A TV mounted above a fireplace in a nicely furnished living room . **Label:** telic **Output:** The verb ""mounted"" in position 2 in the sentence ""A TV mounted above a fireplace in a nicely furnished living room"" does not by itself convey a clear aspect because the sentence is a noun phrase, not a complete sentence with a tense that would allow us to determine aspect.

Table 4: Example error cases.

where the absence of verb ("be alright") or verb position information leads the model to predict the aspect for the entire sentence, resulting in inaccurate predictions.

Case 2 highlights instances of model hallucinations, where initially, the model denies the presence of probed aspects in the verb but subsequently provides an aspect in the following sentence.

Case 3 demonstrates the prediction disparity across the three prompt levels. While the predictions are accurate and nuanced in the first two levels, they become nonsensical and incorrect in level 3, underscoring the model's tendency towards hallucinations and instability.

Case 4 presents a scenario where the model fails to provide an aspectual feature, incorrectly concluding that the verb lacks an aspect.

These error cases underscore that employing LLMs may introduce unexpected errors due to model complexity and hallucinations. Additionally, the inconsistency of model output remains a pertinent question for further investigation.

5 Discussion & Conclusion

This preliminary study evaluates the performance of various LLMs in recognizing lexical aspects, specifically duration and telicity, in zero-shot scenarios. We notice while LLMs, especially the closed-source ones (GPT-3.5 and GPT-4), are capable of recognizing the lexical aspects of verbs in sentences, they lie behind the fine-tuned baselines, indicating the potential need for further adaptation to effectively probe deep linguistic features. We conduct experiments across three levels of prompting to assess the impact of including the verb and its position in the sentence. Our results reveal that LLMs, particularly the closed-source ones, benefit from the additional context of verbs. However, this added complexity sometimes introduced regressions, indicating that while context aids comprehension, it can also pose challenges. The case analysis also introduces concerns about the complexity of hallucinations within the models.

Future research could explore methods to optimize LLMs for aspectual feature recognition, such as fine-tuning LLMs or incorporating additional linguistic knowledge into model training. Currently, we only conduct the prompt in zero-shot settings, i.e. without context information. Previous work showed that prompt-based methods may underestimate the linguistic knowledge of LLMs (Hu and Levy, 2023). Therefore, we call for future exploration in different settings, such as few-shot prompting and Chain-of-Thought (CoT, Wei et al., 2023) prompting.

Overall, our study offers valuable insights into the challenges and opportunities of utilizing LLMs for linguistic feature recognition.

Limitations

The primary limitation of our preliminary work lies in the complexity and instability of LLMs, as detailed in §4.3. The models exhibit sensitivity to prompts and parameter settings. Our study tested only three curated prompts with varying information levels and observed significant variations across these conditions. Future research should delve deeper into these variations to provide explanations for these changes.

Additionally, as noted in previous work (e.g., Zhang, 1995), aspectual readings are sensitive to the context surrounding the verb. Our current study tested aspectual features using a single curated dataset with individual sentences and labels. Future research should explore data with longer texts containing more verbs and possibly provide sequential predictions on verbs within context. This would help to better understand the deeper linguistic comprehension capabilities of LLMs.

Acknowledgements

The research is supported by Munich Center for Machine Learning.

References

Malihe Alikhani and Matthew Stone. 2019. "caption" as a coherence relation: Evidence and implications. In *Proceedings of the Second Workshop on Shortcomings in Vision and Language*, pages 58–67, Minneapolis, Minnesota. Association for Computational Linguistics.

Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. 2023. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Bernard Comrie. 1976. *Aspect*. Cambridge Textbooks in Linguistics. Cambridge University Press, Cambridge.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Annemarie Friedrich and Damyana Gateva. 2017. Classification of telicity using cross-linguistic annotation projection. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2559–2565, Copenhagen, Denmark. Association for Computational Linguistics.

Annemarie Friedrich and Alexis Palmer. 2014. Automatic prediction of aspectual class of verbs in context. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 517–523, Baltimore, Maryland. Association for Computational Linguistics.

Annemarie Friedrich, Alexis Palmer, and Manfred Pinkal. 2016. Situation entity types: automatic classification of clause-level aspect. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1757–1768, Berlin, Germany. Association for Computational Linguistics.

Annemarie Friedrich and Manfred Pinkal. 2015. Automatic recognition of habituals: a three-way classification of clausal aspect. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2471–2481, Lisbon, Portugal. Association for Computational Linguistics.

Annemarie Friedrich, Nianwen Xue, and Alexis Palmer. 2023. A kind introduction to lexical and grammatical aspect, with a survey of computational approaches. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 599–622, Dubrovnik, Croatia. Association for Computational Linguistics.

Jürgen Hermes, Michael Richter, and Claes Neuefind. 2015. Automatic induction of german aspectual verb classes in a distributional framework. In *Proceedings of the International Conference of the German Society for Computational Linguistics and Language Technology (GSCL)*.

Jennifer Hu and Roger Levy. 2023. Prompting is not a substitute for probability measurements in large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5040–5060, Singapore. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las

Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts. *Preprint*, arXiv:2401.04088.

Wolfgang Klein. 1994. *Time in Language*. Routledge, London.

Thomas Kober, Malihe Alikhani, Matthew Stone, and Mark Steedman. 2020. Aspectuality across genre: A distributional semantics approach. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4546–4562, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Elisabeth Leiss. 1992. *Die Verbalkategorien des Deutschen*. De Gruyter, Berlin, Boston.

Eleni Metheniti, Tim Van De Cruys, and Nabil Hathout. 2022. About time: Do transformers learn temporal verbal aspect? In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 88–101, Dublin, Ireland. Association for Computational Linguistics.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Juri Opitz, Shira Wein, and Nathan Schneider. 2024. Natural language processing relies on linguistics. *Preprint*, arXiv:2405.05966.

Anita Ramm, Sharid Loáiciga, Annemarie Friedrich, and Alexander Fraser. 2017. Annotating tense, mood and voice for English, French and German. In *Proceedings of ACL 2017, System Demonstrations*, pages 1–6, Vancouver, Canada. Association for Computational Linguistics.

Eric V. Siegel and Kathleen R. McKeown. 2000. Learning methods to combine linguistic indicators:improving aspectual classification and revealing linguistic insights. *Computational Linguistics*, 26(4):595–627.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikuła, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vlad Feinberg, Wojciech Stokowiec, Yu hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Warkentin, Ludovic Peran, Minh Giang, Clément Farabet, Oriol Vinyals, Jeff Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. 2024. Gemma: Open models based on gemini research and technology. *Preprint*, arXiv:2403.08295.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Zeno Vendler. 1957. Verbs and times. *The Philosophical Review*, 66(2):143–160.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Richard Xiao and Tony McEnery. 2004. *Aspect in Mandarin Chinese: A corpus-based study*. John Benjamins.

Lihua Zhang. 1995. *A Contrastive Study of Aspectuality in German, English and Chinese*. Peter Lang, New York.

A Appendix

A.1 Label Distribution in Test Sets

We present the label distributions from the original test sets (Metheniti et al., 2022) in Table 5.

Test Set	Label '0'	Label '1'	Total
Duration	186	223	409
Telicity	315	292	607

Table 5: Test set statistics for duration and telicity. In the duration subset, '0' and '1' stand for stative and durative aspects, respectively. In the telicity subset, '0' and '1' stand for atelic and telic aspects, respectively.

A.2 Label Distribution in Predictions

Table 6 shows the label distribution from the model predictions. We notice some imbalanced label distribution especially in model Gemma on both duration and telicity sets across three prompt levels. This great imbalance also results in low prediction accuracies, as shown in §4.1. This indicates that the Gemma model might not be adequate to probe the aspectual features. The same imbalance can be found in Llama-2 in level3 on the telicity set.

Model	Level	'0'	'1'	'2'	'3'	'-99'
Gemma	level1	27	382	0	0	0
	level2	14	395	0	0	0
	level3	20	389	0	0	0
GPT-3.5	level1	211	195	1	0	2
	level2	194	214	0	0	1
	level3	193	215	0	0	1
GPT-4	level1	150	196	36	21	6
	level2	208	189	11	0	1
	level3	213	177	18	0	1
Llama2	level1	265	144	0	0	0
	level2	176	230	0	3	0
	level3	180	229	0	0	0
Llama3	level1	202	190	13	4	0
	level2	238	171	0	0	0
	level3	259	149	1	0	0
Mistral	level1	141	252	10	4	2
	level2	122	267	17	2	1
	level3	150	237	15	1	6

(a) Duration

Model	Level	'0'	'1'	'2'	'3'	'-99'
Gemma	level1	558	49	0	0	0
	level2	575	32	0	0	0
	level3	566	40	0	1	0
GPT-3.5	level1	252	348	0	0	7
	level2	454	147	0	0	6
	level3	450	153	0	0	4
GPT-4	level1	213	295	71	16	12
	level2	344	243	14	0	6
	level3	353	227	22	1	4
Llama2	level1	350	255	0	2	0
	level2	586	19	0	2	0
	level3	565	41	0	0	1
Llama3	level1	73	523	8	3	0
	level2	194	412	1	0	0
	level3	260	346	1	0	0
Mistral	level1	280	314	4	8	1
	level2	401	193	3	5	5
	level3	334	263	1	4	5

(b) Telicity

Table 6: Label distribution from the model predictions on duration and telicity set. '0' and '1' are the original binary labels of the dataset. '2' means model cannot find an aspect or thinks the verb doesn't have an aspect. '3' means nonsense output. '-99' means model refusal.