Leaving, Staying, or Living There?
Duration Commonsense Reasoning with Human-LLM Co-annotations

Anonymous ACL submission

Abstract

Gardening in the backyard implies living there while checking in an Airbnb more likely means staying for a few days. Inferring the duration of human activities from visual and language cues is fundamental to human intelligence. However, this important problem so far has received limited attention. In this paper, we investigate if vision language models possess such duration commonsense. We create a new corpus of geotagged tweets that are co-annotated with both human and LLM's duration commonsense to serve as a test set for this task. We found that after finetuning, open-source VLMs perform comparably to reasoning-based closed-source VLMs in understanding duration commonsense. Additionally, we provide error analyses that highlight the critical commonsense knowledge necessary for accurate duration inference.

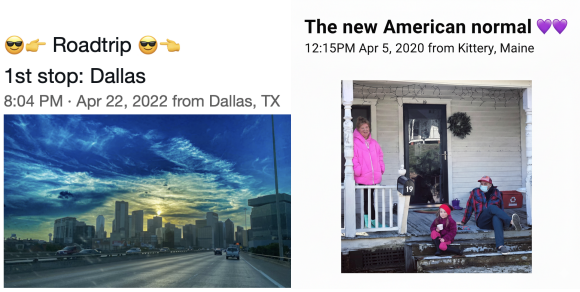

Figure 1: Examples of tweets in which the author will stay at the tagged places for a short (left) or long (right) time. The author of the left tweet will stay at *Dallas* for a short time during the road trip, while the author of the right tweet will stay at *Kittery* for a long time because the tweet shows people at a residential house, which suggests the author is living at this location and documenting their daily life during a period of extended stay-at-home measures.

1 Introduction

Duration commonsense, the skill of deciphering the duration associated with human activities, forms a foundational aspect of human intelligence, allowing us to navigate and understand the temporal aspects of various activities and situations. This skill relies heavily on the ability to understand social contexts and interpret visual and language cues from the scene. Humans often rely on nuances in text and images, which can be challenging for vision language models.

Consider the tweets in Figure 1. The author of the tweet on the left is likely to stay for short time in Dallas, as people typically do on their road trips. Although the text alone could arguably be enough to conclude so, the image of the Dallas skyline and the fact that the author is driving provide additional evidence. The author of the tweet on the right may spend a couple of years in Kittery, as it can be inferred that the author is living at this location and documenting their daily life during a period of extended stay-at-home measures.

Prior research in duration understanding is centered on event duration inference (Gusev et al., 2011; Leeuwenberg and Moens, 2018), aiming at predicting how long events last regardless if people participate or how long they participate in the events. A few works have been done with formal texts (Vempala and Blanco, 2016; Chinnappa et al., 2020) to understand commonsense related to people's duration.

In this paper, we focus on the problem of using noisy user-generated content to determine how long people will stay at the places they tag in their tweets posted on X. X (formerly known as Twitter) has quickly become one of the most popular social media platforms: it has 95.4 million users in the US, spending an average of 30.9 minutes per day on it (Woodward, 2023). People talk about their vacation trips, work, and opinions in tweets. Tweets contain not only text (e.g., hashtags, emojis, and URLs), but also multimedia content (e.g., images and videos) and geotags. Indeed, 42% of tweets have images (Lee, 2015), Tweets with im-

ages receive 150% more retweets than tweets without them (Alexis, 2022). On the contrary, geotags are not consistently available because sharing location information when tweeting is not enabled by default due to privacy concerns.

Duration commonsense understanding from social media can be useful in many applications, including transportation planning, personalized recommendation systems, and emergency response. For example, understanding the duration of people staying in places that they check in can help understand travel patterns and estimate travel demand. As another example, during emergency situations, knowing who are likely to be around the impacted area can help to direct warning or evacuation messages to the right hands.

The main contributions of this paper are: (a) We **formulate a novel and socially grounded problem**: inferring duration commonsense about people from text and images in social media. The task is inherently challenging, requiring nuanced interpretation of social context that even humans find difficult; (b) We construct a **Human–LLM co-annotated Corpus** to address the challenges of annotating socially complex phenomena. By combining human judgment with the consistency and scalability of large language models, we create **a corpus of 4,312 geotagged tweets** with duration annotations indicating how long tweet authors are likely to stay at the locations mentioned in their posts[1]; (c) We study **duration commonsense understanding capabilities of the traditional as well as recent largest vision and language models**. Results show that providing samples from our dataset improves traditional models and recent VLM's performance to a similar level. Finetuning improves the recent instruction-tuned large language models by over **85%**; (d) We perform error analyses providing **insights** into what types of commonsense knowledge contributes to duration commonsense understanding.

2 Related Work

Duration has been investigated since it was proposed long ago (Pan et al., 2011). In their pioneer work, they create a corpus consisting of news articles in which events are annotated for estimated bounds on their duration, as well as providing simple machine learning baseline models that take into account word n-grams, syntactic relations, and WordNet hypernyms (Miller et al., 1990). Since neural networks started to be adopted in NLP (Socher et al., 2011; Chen and Manning, 2014; Kim, 2014), many neural network-based models are proposed to predict event durations using formal text, including clinical documents and news articles (Yang et al., 2020; Leeuwenberg and Moens, 2020; Ma et al., 2021; Dasgupta et al., 2025). In the realm of social media data, Williams and Katz (2012) create a dataset of 14 million tweets and build a bootstrapping classifier for the event duration. However, the duration of the event differs from the duration of people's stays. People may or may not participate in the events they mention in their tweets (Sanagavarapu et al., 2017). Even if people participate in the mentioned events, the duration of those events may not correspond to the duration of people's stays. Different from previous efforts aiming at understanding commonsense knowledge about event-centric duration (Zhou et al., 2019, 2020; Pereira et al., 2020, 2021), our work presented in this paper focuses on understanding people-centric duration commonsense.

There have been few previous works studying duration related to people. For instance, Chinnappa et al. (2020) determine the duration of the possession relations, in which possessions are defined as someone possessing something. More related to our work, Vempala and Blanco (2021) extract where people are located and for how long, creating a corpus based on OntoNotes (Weischedel et al., 2010), which consists of formal text. In contrast, our work is complementary in two ways. First, we work with informal text, which makes our task more challenging. Second, we consider not only text but also images from which duration commonsense could be inferred.

Annotation of such socially complex task is itself challenging. While crowdsourcing is relatively inexpensive and flexible, the resulting quality can be inconsistent. Recent advances in large language models (LLMs) have made them attractive for annotation. There is a significant body of research explores using LLMs to support or replace human annotators (Wang et al., 2021; Li, 2024), with promising results across a range of domains. Notably, LLMs have shown strong performance in annotating socially nuanced tasks (Gilardi et al., 2023), sometimes even surpassing human annotators (He et al., 2024). These capabilities enable the design of

[1] Corpus and code available at https://anonymous.4open.science/r/loc_tmp3-7D5B

`0-24 hours` `1-30 days` `more than 30 days`

Figure 2: Annotation examples. The author of the leftmost tweet is departing *SeaTac* ("be safe this time" in the text and the airport waiting area in the image), indicating a stay of `0-24 hours`. The middle tweet exemplifies `1-30 days`, in which the author of the tweet is enjoying her roadtrip to *St Louis*, indicating she will stay there for 1-30 days, as the time spent at one location during the roadtrip would most likely be less than one week. The author of the right tweet lives in *Iowa City*, as the image shows that his family members (mentioned in the tweet text) are in front of his house, leading to `more than 30 days`.

automated annotation pipelines that are accessible without extensive technical expertise, broadening their applicability.

Our work is motivated by these previous efforts. However, to the best of our knowledge, we are the first to study the duration commonsense about people using text and images on social media.

3 A Corpus of Duration Commonsense in Tweets

We create a new corpus as there is not an existing dataset. To capture duration commonsense as understood by both humans and large language models (LLMs), we adopt a hybrid annotation strategy that integrates crowdworker and LLM annotations. Human annotators contribute socially grounded judgments shaped by real-world experience, while LLMs provide scalable and consistent interpretations, having shown strong performance on socially nuanced tasks. Combining both sources allows us to leverage their complementary strengths: human contextual sensitivity and LLM consistency. We then introduce a process for combining these human and LLM labels into a cohesive and high-quality dataset.

3.1 Collecting Tweets

We choose to work with GeoCov19Tweets (Lamsal, 2020), which consists of approximately 420,000 tweets during Covid-19 at the time of download. Those tweets cover a broad range of human activities, as shown in Figure 2. GeoCov19Tweets does not contain tweet metadata but rather contains tweet IDs. With tweet IDs, we download tweet metadata using the official X Developer API. What sets this dataset apart from other Twitter datasets is that each tweet in GeoCov19Tweets comes with geotags in its metadata, which correctly labels the authors' locations.

Tweets Filtering To filter out tweets not suitable for our purposes, we followed three criteria. First, we only downloaded tweets with a geotag at the US city level, which means high-level geotags (e.g., *California, USA*) and low-level geotags (e.g., *Pacific Plaza*) are excluded. Second, each tweet must contain text and images. Finally, tweets published by companies or organizations were excluded. To do so, we applied a text-based filter and an image-based filter. The text-based filter uses a manually defined list of keywords (e.g., *daily*, *sports*, and *sales*) to identify usernames of potential company Twitter users, while the image-based filter uses CRAFT (Baek et al., 2019) and manually defined rules to identify poster-like images. Additionally, we filtered out tweets containing explicit duration-related expressions, such as *5 days* and *one year*.

3.2 Annotation Guidelines

We ask both crowdworders and LLM to annotate:

- Amazon Mechanical Turk (AMT): We utilized the Amazon Mechanical Turk platform

to obtain annotations from crowd-workers.

- LLM Annotation: Six Large Language Models (LLMs), Qwen 2.5 (Bai et al., 2025), GPT5-mini (OpenAI, 2025), Grok4 (X.AI, 2025), Claude 3, Gemini 2.5 flash and Gemini 2.5 pro (Deepmind, 2025) were employed to generate annotations.

For crowdworkers, we present each tweet with text and image(s) and ask them a straightforward question: "How long will the author of the tweet stay in the *city* continuously after publishing this tweet?" The *city* in question is the geotagged city of the tweet. For instance, given a tweet with the text "*I am gonna join abortion law protest in Dallas tomorrow*", its associated image, and a geotag of *Plano, TX*, we can infer that the author of the tweet will stay in Plano for less than 24 hours and then go to Dallas. Crowdworkers choose among four options corresponding to *leaving*, *staying*, *living*, and *can not determine*:

1. `0-24 hours`: the author will continuously stay in the *city* for 0–24 hours after publishing the tweet.
2. `1-30 days`: the author will continuously stay in the *city* for 1–30 days after publishing the tweet;
3. `more than 30 days`: the author will continuously stay in the *city* for more than 30 days after publishing the tweet; or
4. `I cannot determine`: it is unclear how long the author will continuously stay in the *city* after publishing the tweet.

For LLMs, we use a task-specific system prompt: "You are a helpful assistant to infer the duration of human activities." Each tweet, along with text and image(s), is presented, and the same question with four options is posed to the model. A detailed example of the prompt can be found in Figure 6.

3.3 Annotation Quality Control of Crowdworkers and LLMs

Ensuring annotation quality is critical for both crowdsourcing effort or LLM annotation. To ensure quality of crowdsourcing annotation: (1) We recruit crowdworkers and with previous approval rates above 95%; (2) Furthermore, we have set a minimum completion time of 5 seconds per Human Intelligence Task, and do not allow crowdworkers to continue if their average completion time falls below this threshold. This decision was made based on observations during pilot annotations.

For crowdsourcing annotation, we collect five annotations per Human Intelligence Task/LLM annotation task and filter out unreliable annotations until we obtain *substantial* inter-annotator agreement. We apply Multi-Annotator Competence Estimation (MACE) (Hovy et al., 2013) to weight annotators by inferred competence and derive aggregated labels, avoiding the limitations of majority voting. Inter-annotator reliability is then quantified using Krippendorff's α(Krippendorff, 2011), where $\alpha = 0$ indicates chance-level agreement and $\alpha = 1$ indicates perfect agreement. Values of $\alpha \geq 0.6$ are considered substantial and $\alpha > 0.8$ near-perfect. We repeat the following steps for each batch until $\alpha \geq 0.6$:

1. Use MACE to calculate the competence score for each crowdworker and sort them in descending order.
2. Discard all the annotations by the crowdworker with the lowest score.
3. Check Krippendorff's α on the remaining annotations.

Additionally, we employ Fleiss' Kappa (FK) (Fleiss, 1971) to assess inter-rater agreement among LLM annotations. This metric provides a robust measure of response consistency across models. Table1 reports the corresponding kappa values and their interpretations. Models with $\kappa < 0.2$ (Qwen 2.5 and GPT5-mini) are treated as low-quality annotators and excluded from further aggregation. The final adjudicated labels are then derived using MACE on the remaining four LLMs.

Model	Kappa Value	Interpretation
Qwen 2.5	0.1659	Moderate agreement
GPT5-mini	0.1892	Moderate agreement
Grok 4	0.2985	Fair agreement
Claude 3	0.3256	Fair agreement
Gemini 2.5 flash	0.5302	Substantial agreement
Gemini 2.5 pro	0.5572	Substantial agreement

Table 1: Inter-model agreement measured using Cohen's Kappa.

3.4 LLM and Human Annotation Alignment

Table 2 presents a confusion matrix detailing the joint distribution of human and LLM annotations, where values on the main diagonal represent agreement. From the table, we can see that among the pairs where human and LLM agree, "more than 30 day" has the most agreement. This suggests that, in general, both humans and LLMs share the same background knowledge regarding typical quarantine durations.

LLM	Human		
	0-24hours	1-30days	more than 30 days
0-24hours	**2.00%**	0.73%	18.36%
1-30days	4.57%	**2.65%**	3.20%
more than 30 days	34.43%	2.37%	**31.69%**

Table 2: Confusion matrix of duration labels assigned by human and LLM annotators. Diagonal cells (in **bold**) indicate consensus labels between annotators.

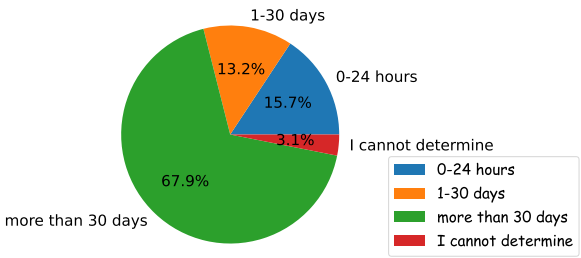

Figure 3: Label distribution.

We analyzed the labels where LLMs and human annotators disagreed and identified four main reasons for these discrepancies:

- R1: COVID-19 quarantine duration commonsense. LLMs or human annotators may misinterpret the typical duration in the COVID-19 quarantine period.
- R2: Duration not calculated at the city level. The annotator may overly focus on the specific location mentioned in the tweet, attempting to predict the duration based on that location, rather than at the city level.
- R3: Tweet intent unclear. The tweet might express intent, memory, or aspiration rather than an actual visit. For example, a tweet might include a picture from a past trip, but if the annotator doesn't know that the person is under quarantine, he might mistakenly interpret the post as indicating that the person is currently traveling.
- R4: Author residency status unclear. It's hard to interpret whether the author is a tourist or a resident. For example, a tweet showing a man cycling in a city could suggest he is either traveling in another city (tourist) or in his hometown (resident). This ambiguity might lead the annotator to predict a stay of 1-30 days or more than 30 days.

The first two reasons (R1 and R2) represent the most common errors made by the annotators, where annotators who make predictions within 30 days can be directly rejected. Table 3 presents examples of these errors. R3 and R4, being more ambiguous, cannot be definitively resolved, so we discarded those tweets. Examples of these cases are shown in Table 5.

To determine the final duration label, we apply the rejection and discard strategy based on the aforementioned reasoning. The remaining annotations, whether from the LLM or human annotator, are kept as the final label, while discarded annotations are excluded. The distribution of disagreement reasons for each label pair is displayed in Table 4. After fitering, the final corpus consists of 4,312 annotated tweets. In the rest of this paper, we work with these tweets.

4 Corpus Analysis

Figure 3 depicts the distribution of labels in our corpus, revealing that for approximately 67.9% of tweets, their authors will stay in the geotagged city for `more than 30 days`. Labels of the remaining tweets are nearly evenly split between `1-30 days` and `0-24 hours`. Less than 4% of tweets are labeled `I cannot determine`. We exclude them in the rest of this paper due to the lack of duration commonsense.

5 Experiments and Results

Our problem can be modeled as a classification task. The input is a (tweet, city) pair, and the output is a label indicating how long the authors of tweets will stay at the places they tag in their tweets after publishing tweets. We create stratified training and test splits (80% / 20%). We only use the first image if the tweet includes more than one image. Our models do not take user information (e.g., username and profile image) into account. They make predictions based exclusively on the content of tweets. We conduct all experiments on NVIDIA A100 GPUs in a week. The evaluation of different models is categorized into three types: **(1) Human:** The adjudicated labels by mechanical turk **(2) PTLM:** Next, we assess the performance of pretrained language models (PTLM). We use RoBERTa (Liu et al., 2019) as the text encoder and Vision Transformer (Dosovitskiy et al., 2021) as the image encoder and evaluate them on the testing set. We integrate the text and image representations using three methods through concatenating, adding, and multiplying the representations of the text and images to obtain the representation of the tweet. We finetuned using LoRA. **(3) LLM (Open-Source):** We finetune Qwen 2.5 (Bai et al., 2025),

R1: COVID-19 quarantine duration commonsense	R2: Duration not calculated at the city level
Quarantine Birthday for me Thanks for all your wishes and hoping next year will be a not virus birthday!	A beautiful sunny Saturday afternoon working nearby this area... much quiet and less people here during this crisis Stay safe stay healthy!!!
LLM: `1-30 days` The reason is that the tweet text explicitly mentions #Quarantine Birthday and uses hashtags like #lockdown and #coronavirus. During the COVID-19 pandemic, a quarantine period was a mandated or recommended period of isolation, typically lasting for 14 days. This falls within the 1-30 day range, making it the most logical duration for the author's stay in Manhattan.	LLM: `more than 30 days` The reason is that the tweet mentions the author is "working nearby this area". This suggests that the author either lives in or near Hermosa Beach or has a regular job there, implying they are a local resident and not a tourist on a short visit. Therefore, their stay in the city is long-term.
Human: `more than 30 days`	Human: `1-30 days`

Table 3: Two most common system errors made by LLM and Human annotators.

Prediction: <LLM, Human>	R1	R2	R3	R4	# Tweets
<0-24 hours, 1-30 days>	0	0	3(D)	5(D)	8
<0-24 hours, more than 30 days>	0	20(H)	191(D)	0	201
<1-30 days, 0-24 hours>	0	0	26(D)	24(D)	50
<1-30 days, more than 30 days>	4(H)	0	0	31(D)	35
<more than 30 days, 0-24 hours>	0	290(L)	87(D)	0	377
<more than 30 days, 1-30 days>	3(L)	0	0	23(D)	26

Table 4: Prediction Pair Analysis. 'L' indicates that the LLM's prediction is the final decision, 'H' indicates that the human annotator's prediction is the final decision, and 'D' indicates that the annotation was discarded.

Llama 3.2 (Grattafiori et al., 2024), Glm-4v (Team et al., 2025) and evaluate them on the testing set. **(4) LLM (Close-Source):** Finally, we evaluate the performance of Claude 3, Gemini 1.5 Flash, GPT5-mini (OpenAI, 2025), Grok4 and Gemini 2.5 pro (Deepmind, 2025). which represent proprietary LLMs, using zero-shot and few-shot promptings on the full annotated testing set. See more details in Appendix A.3 and A.4.

5.1 Results

The evaluation results are presented in Table 6. Our observations include:

Performance of PTLM. We observe that combining text and images, regardless of how the modality interaction is encoded, is beneficial, indicating that incorporating information carried by different modalities is needed. Secondly, CLIP, pretrained on (text, image) pairs, outperforms the model that combines models pretrained on only text or images, suggesting that joint pretraining on (text, image) pairs can effectively bridge the modality gap.

Effect of finetuning and few-shot learning. Finetuning substantially enhances the performance of open-source LLMs in both zero shot and few shot setting. The relative gains are larger in zero-shot; for example, Glm-4v improves by 243% (0.23→0.79) under zero shot scenario and has the best performance among all the models. Moreover, LLMs generally perform better in the few-shot setting than in the zero-shot setting, indicating the importance of providing exemplars.

Effect of reasoning Generally, reasoning-based LLMs (GPT-5 mini, Grok4, and Gemini 2.5 Pro) outperform non reasoning-based models (Claude 3 and Gemini 1.5 flash), since they have better contextual understanding abilities.

5.2 Error Analysis of Glm-4v

To conduct a more detailed error analysis, we manually examined 200 sampled responses generated by GLM-4v and classified their errors into two main categories: (1) **Overreliance on Geotags:** For instance, consider the tweet: "The #Stella #thimble #ring is your #solution! #safewithstyle #brevard #design #thimble #sanitation #armor #armorring @ Miami, Florida". In this case, GLM-4v interprets the duration of the stay as 1-30 days, based on the hashtag #brevard, which refers to Brevard County in Florida, implying the author's business or base is located there. However, the tweet is geotagged in Miami, suggesting that the author is likely in Miami on a business trip. In reality, Brevard is relatively close to Miami, so it is more plausible that the author has not left their base, and the correct duration should be more than 30 days. (2) **Failure**

R3: Tweet intent unclear	R4: Author residency status unclear
gloom my day #locked #gates #cold	I am a Barbie girl ... in a Covid world!
LLM: `more than 30 days` The reason is that the tweet was posted during the COVID-19 pandemic, as indicated by the hashtags #covid-19 and #epedemic. The image and hashtags like #boardwalkclosed, #locked, and #fencedin show that public areas were shut down. During this period, travel was heavily restricted, and people were generally under stay-at-home orders. It is most likely that the author is a resident of Seaside Heights documenting the unusual state of their local area during the lockdown, rather than a tourist visiting a closed-down town. As a resident, they would be staying for more than 30 days.	LLM: `1-30 days` The tweet includes the hashtags #saturdayvibes and #saturdaynight, indicating that the author is dressed up for a specific weekend event. This, combined with being in a specific city 'Dallas', suggests that she is likely on a short trip, probably for the weekend. A weekend trip typically lasts for a few days, which falls into the 1-30 day category. She posted on Saturday night, so it is reasonable to assume she will stay at least until the next day, but is unlikely to be staying for more than a month for a weekend event.
Human: `0-24 hours`	Human: `more than 30 days`

Table 5: R3 and R4 made by LLM and Human annotators.

to Use Image Context Correctly: For instance, an image depicts a packed suitcase and a hotel room, but GLM-4v erroneously predicts a short stay, similar to a coffee visit, and categorizes the duration as less than 24 hours. But actually a packed suitcase typically suggests preparation for a more extended stay. So the correct duration should be 1-30 days.

6 Conclusions

In this paper, we have studied the problem of duration commonsense understanding. More specifically, we presented a novel task of determining how long authors of tweets will stay at the places they tag in their tweets. We created a new corpus (4,312 tweets) for this task. By analyzing our corpus, we observe that understanding the duration related to people requires a large degree of common sense (e.g., flight departure will take place in hours, and visiting Gateway Arch takes a few days). To the best of our knowledge, we are the first to tackle this challenging problem. Our experimental results highlight the difficulty of this task for some zero-shot, closed-source LLMs. However, fine-tuning open-source models demonstrated significant improvements, aligning with the capabilities of the latest thinking-based LLMs. Our detailed analyses of the most common errors made by the best PTLM model (i.e., CLIP) and LLM (i.e., Glm-4v) provide insights into various kinds of commonsense knowledge contributing to the duration commonsense understanding.

7 Ethical Considerations and Limitation

7.1 Ethical Considerations

Determining how long people will stay at the given places has the potential to open the door to malicious (or just unwanted) tracking and surveillance. For example, applications that track timeline information may turn around and sell that data, revealing people's stillness and movements—whether it is for commercial purposes or stalking. Equally important, Twitter users may not be aware that their tweets can be used for research purposes (Fiesler and Proferes, 2018). We are not interested in tracking people or surveillance. Instead, we are interested in investigating whether and how the task of duration commonsense understanding can be automated and how LLM's ability in duration commonsense understanding can be elicited.

In order to alleviate the issues above and preserve privacy, we implemented the following safeguards. First, the tweets in the corpus that we will publish do not contain user information, such as usernames and profile images. In addition, the tweet IDs will be excluded so that other people will not have access to these tweets. Second, in the case when an individual requests a specific tweet to be removed from our corpus, we have designed

Methods	Backbone	F1	
		Zero-shot	Few-shot
Human	N/A	0.37	
PTLM	RoBERTa-Large *340M*	0.18	0.21
	ViT-Large/16	0.05	0.07
	Concatenation	0.22	0.24
	Addition	0.16	0.17
	Multiplication	0.13	0.15
	CLIP	<u>0.38</u>	<u>0.38</u>
LLM(Open-Source, w/o FT)	Qwen 2.5	<u>0.28</u>	0.32
	Llama 3.2	0.22	0.41
	Glm-4v	0.23	<u>0.43</u>
LLM(Open-Source, w/ FT)	Qwen 2.5	**<u>0.80</u>** (↑185.7%)	0.82 (↑156.3%)
	Llama 3.2	0.65 (↑195.5%)	0.76 (↑85.4%)
	Glm-4v	0.79 (↑243.5%)	**<u>0.85</u>** (↑97.7%)
LLM(Close-Source, w/o reasoning)	Claude 3	0.22	0.25
	Gemini 1.5 Flash	<u>0.53</u>	<u>0.68</u>
LLM(Close-Source, w/ reasoning)	GPT5-mini	0.62	0.80
	Grok4	<u>0.70</u>	<u>0.82</u>
	Gemini 2.5 pro	0.70	0.80

Table 6: Weighted average F1 scores of various (L)LMs. The best performances within each method are <u>underlined</u> and the best among all methods are **bold-faced**.

a takedown request process via an online form following Mirowski et al. (2019).

7.2 Limitation

The work presented here has several limitations:

Corpus and annotations Our corpus only contains tweets with geotags at the US-city level. While accounting for any location worldwide is outside the scope of this paper, it is possible that the corpus does not generalize to other locales. Additionally, the corpus only contains Covid-related tweets. However, our approaches and analyses mainly focus on the relationship between location and stay duration of human activities and are not dependent on COVID-19. We believe the findings will likely generalize to other domains.

In addition to the bias contained in the tweets, another potential source of bias in our corpus is related to annotator selection. For instance, since the only criteria we use to select annotators are their locations (i.e., they must reside in the US) and previous approval rates (i.e., must be larger than 95%), their demographics, religions, and ages may vary significantly, potentially affecting their judgments on duration determination. Due to privacy concerns, crowdsourcing platforms like Amazon Mechanical Turk do not allow filters for demographics, religions, or ages. Therefore, we believe that our selection criteria are already the best possible approach, and the impact of annotator bias is at a minimal level.

Experiments and results The traditional models rely on pretrained encoders to generate the distributed representation of each modality. Since these encoders are pretrained on tasks different from ours, the distributed representations might be task-specific. This could degrade the model's performance.

References

Carine Alexis. 2022. 29 Incredible Stats that Prove the Power of Visual Marketing. https://movableink.com/blog/29-incredible-stats-that-prove-the-power-of-visual-marketing. Accessed: Nov 30, 2023.

Youngmin Baek, Bado Lee, Dongyoon Han, Sangdoo Yun, and Hwalsuk Lee. 2019. Character region awareness for text detection. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 9365–9374. Computer Vision Foundation / IEEE.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang

Lin. 2025. Qwen2.5-vl technical report. *Preprint*, arXiv:2502.13923.

Danqi Chen and Christopher Manning. 2014. A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 740–750, Doha, Qatar. Association for Computational Linguistics.

Dhivya Chinnappa, Srikala Murugan, and Eduardo Blanco. 2020. Beyond possession existence: Duration and co-possession. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8332–8341, Online. Association for Computational Linguistics.

Tirthankar Dasgupta, Manjira Sinha, and Sudeshna Jana. 2025. Predicting icu length of stay for patients using latent categorization of health conditions. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: Industry Track)*, pages 422–430.

Google Deepmind. 2025. Gemini 2.5: Our most intelligent ai model.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Casey Fiesler and Nicholas Proferes. 2018. "participant" perceptions of twitter research ethics. *Social Media + Society*, 4(1):2056305118763366.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76:378–382.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.

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, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit San-

gani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Andrey Gusev, Nathanael Chambers, Divye Raj Khilnani, Pranav Khaitan, Steven Bethard, and Dan Jurafsky. 2011. Using query patterns to learn the duration of events. In *Proceedings of the Ninth International Conference on Computational Semantics (IWCS 2011)*.

Zeyu He, Chieh-Yang Huang, Chien-Kuang Cornelia Ding, Shaurya Rohatgi, and Ting-Hao Kenneth Huang. 2024. If in a crowdsourced data annotation pipeline, a gpt-4. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. Learning whom to trust with MACE. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1120–1130, Atlanta, Georgia. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural*

Language Processing (EMNLP), pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability. *Retrieved from* https://repository.upenn.edu/asc_papers/43.

Rabindra Lamsal. 2020. Coronavirus (covid-19) geotagged tweets dataset.

Kevan Lee. 2015. What analyzing 1 million tweets taught us. https://thenextweb.com/socialmedia/2015/11/03/what-analyzing-1-million-tweets-taught-us/. Accessed: December 3, 2018.

Artuur Leeuwenberg and Marie-Francine Moens. 2018. Temporal information extraction by predicting relative time-lines. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1237–1246, Brussels, Belgium. Association for Computational Linguistics.

Artuur Leeuwenberg and Marie-Francine Moens. 2020. Towards extracting absolute event timelines from english clinical reports. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2710–2719.

Jiyi Li. 2024. A comparative study on annotation quality of crowdsourcing and llm via label aggregation. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6525–6529.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Mingyu Derek Ma, Jiao Sun, Mu Yang, Kung-Hsiang Huang, Nuan Wen, Shikhar Singh, Rujun Han, and Nanyun Peng. 2021. EventPlus: A temporal event understanding pipeline. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations*, pages 56–65, Online. Association for Computational Linguistics.

George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. 1990. Introduction to WordNet: An On-line Lexical Database*. *International Journal of Lexicography*, 3(4):235–244.

Piotr Mirowski, Andras Banki-Horvath, Keith Anderson, Denis Teplyashin, Karl Moritz Hermann, Mateusz Malinowski, Matthew Koichi Grimes, Karen Simonyan, Koray Kavukcuoglu, Andrew Zisserman, and Raia Hadsell. 2019. The streetlearn environment and dataset. *Preprint*, arXiv:1903.01292.

OpenAI. 2025. Introducing gpt-5. *OpenAI*.

Feng Pan, Rutu Mulkar-Mehta, and Jerry R. Hobbs. 2011. Annotating and Learning Event Durations in Text. *Computational Linguistics*, 37(4):727–752.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Lis Pereira, Fei Cheng, Masayuki Asahara, and Ichiro Kobayashi. 2021. ALICE++: Adversarial training for robust and effective temporal reasoning. In *Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation*, pages 373–382, Shanghai, China. Association for Computational Linguistics.

Lis Pereira, Xiaodong Liu, Fei Cheng, Masayuki Asahara, and Ichiro Kobayashi. 2020. Adversarial training for commonsense inference. In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 55–60, Online. Association for Computational Linguistics.

Krishna Chaitanya Sanagavarapu, Alakananda Vempala, and Eduardo Blanco. 2017. Determining whether and when people participate in the events they tweet about. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 641–646, Vancouver, Canada. Association for Computational Linguistics.

Richard Socher, Jeffrey Pennington, Eric H. Huang, Andrew Y. Ng, and Christopher D. Manning. 2011. Semi-supervised recursive autoencoders for predicting sentiment distributions. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 151–161, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014.

Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958.

V Team, Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, Shuaiqi Duan, Weihan Wang, Yan Wang, Yean Cheng, Zehai He, Zhe Su, Zhen Yang, Ziyang Pan, Aohan Zeng, Baoxu Wang, Bin Chen, Boyan Shi, Changyu Pang, Chenhui Zhang, Da Yin, Fan Yang, Guoqing Chen, Jiazheng Xu, Jiale Zhu, Jiali Chen, Jing Chen, Jinhao Chen, Jinghao Lin, Jinjiang Wang, Junjie Chen, Leqi Lei, Letian Gong, Leyi Pan, Mingdao Liu, Mingde Xu, Mingzhi Zhang, Qinkai Zheng, Sheng Yang, Shi Zhong, Shiyu Huang, Shuyuan Zhao, Siyan Xue, Shangqin Tu, Shengbiao Meng, Tianshu Zhang, Tianwei Luo, Tianxiang Hao, Tianyu Tong, Wenkai Li, Wei Jia, Xiao Liu, Xiaohan Zhang, Xin Lyu, Xinyue Fan, Xuancheng Huang, Yanling Wang, Yadong Xue, Yanfeng Wang, Yanzi Wang, Yifan An, Yifan Du, Yiming Shi, Yiheng Huang, Yilin Niu, Yuan Wang, Yuanchang Yue, Yuchen Li, Yutao Zhang, Yuting Wang, Yu Wang, Yuxuan Zhang, Zhao Xue, Zhenyu Hou, Zhengxiao Du, Zihan Wang, Peng Zhang, Debing Liu, Bin Xu, Juanzi Li, Minlie Huang, Yuxiao Dong, and Jie Tang. 2025. Glm-4.5v and glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *Preprint*, arXiv:2507.01006.

Alakananda Vempala and Eduardo Blanco. 2016. Beyond plain spatial knowledge: Determining where entities are and are not located, and for how long. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1502–1512, Berlin, Germany. Association for Computational Linguistics.

Alakananda Vempala and Eduardo Blanco. 2021. Temporally anchored spatial knowledge: Corpora and experiments. *Natural Language Engineering*, 27(5):519–543.

Shuohang Wang, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2021. Want to reduce labeling cost? GPT-3 can help. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4195–4205, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*.

Ralph Weischedel, Mitch Marcus, Martha Palmer, Eduard Hovy, Robert Belvin, Sameer Pradhan, and Lance Ramshaw. 2010. Ontonotes: A large training corpus for enhanced processing. In *Handbook of Natural Language Processing and Machine Translation*. Springer, New York, NY.

Jennifer Williams and Graham Katz. 2012. Extracting and modeling durations for habits and events from Twitter. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 223–227, Jeju Island, Korea. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Matthew Woodward. 2023. Twitter user statistics 2023: What happened after "x" rebranding? https://www.searchlogistics.com/learn/statistics/twitter-user-statistics/.

X.AI. 2025. Grok 4 report. *X.AI*.

Zonglin Yang, Xinya Du, Alexander Rush, and Claire Cardie. 2020. Improving event duration prediction via time-aware pre-training. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3370–3378, Online. Association for Computational Linguistics.

Ben Zhou, Daniel Khashabi, Qiang Ning, and Dan Roth. 2019. "going on a vacation" takes longer than "going for a walk": A study of temporal commonsense understanding. 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 3363–3369, Hong Kong, China. Association for Computational Linguistics.

Ben Zhou, Qiang Ning, Daniel Khashabi, and Dan Roth. 2020. Temporal common sense acquisition with minimal supervision. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7579–7589, Online. Association for Computational Linguistics.

A Appendix

A.1 Annotation Details

There are approximately 2.5k annotators involved in our crowdsourcing efforts. Each tweet included in our corpus has an average of 2.5 annotations. Regarding the types of activities covered in our corpus, we do not see a direct connection between these activities and our error analysis because our error analysis categorizes errors based on different types of commonsense rather than different types of activities.

We discarded approximately 30% of the annotations collected through Amazon Mechanical Turk.

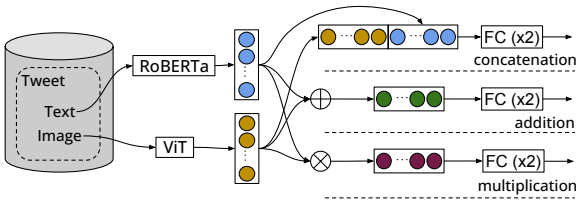

Figure 4: Architecture of the traditional model.

A.2 Difference between event duration inference and duration commonsense

The main difference is that the duration commonsense we focus on in this paper is not centered around events but around people. For example, a tweet like "Looks like the music festival downtown is absolutely packed this weekend! Everyone's sharing such amazing photos #FestivalVibes" discusses an event (the music festival), but it does not imply that the user is participating in it, and we cannot infer how long the user will stay there. In contrast, a tweet like "Finally made it to the tech conference! Excited to spend the next three days soaking up all the insights and networking #TechConf2025" provides clear duration information, as it mentions the user's plans to stay for three days.

A.3 Experiment setting for pretrained language models

We use RoBERTa-Large *340M* (Liu et al., 2019) as the text encoder and Vision Transformer (Dosovitskiy et al., 2021) as the image encoder as shown in Figure 4. [2] We integrate the text and image representations using three methods through concatenating, adding, and multiplying the representations of the text and images to obtain the representation of the tweet. With the tweet representation, we apply two fully connected layers (sizes: 256 and 3) to reduce the dimensionality and make the final prediction. We use dropout (Srivastava et al., 2014) in the second-to-last fully connected layer (rate: 0.2) to avoid overfitting. We use the pretrained and Vision Transformer released by HuggingFace (Wolf et al., 2020) and conduct all experiments using Pytorch (Paszke et al., 2019). We train all networks for up to 100 epochs using Adam optimizer (Kingma and Ba, 2015), categorical cross-entropy as the loss function, and a batch size of 16. We stop the training process before 100 epochs if there is no improvement for 5 consecutive epochs.

[2] We do not freeze RoBERTa and Vision Transformer during the training process. All their parameters are trainable.

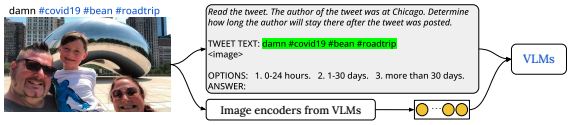

Figure 5: Example of our experiments for duration commonsense understanding. We first use image encoders from VLMs to obtain image representations. Then, we use similar prompts to include tweet text, placeholders ("<image>") indicating where image representations should be taken, and options that VLMs need to select. Lastly, with image representations and prompts, we use VLMs to perform inference.

A.4 Experiment setting for vision language models

To analyze the capability of recent large models (with at least 1 billion parameters) on duration commonsense understanding, we conduct extensive experiments with modern vision language models (VLMs). To ensure the coverage of VLMs, we choose to work with five most recent VLMs: three open models—Qwen 2.5 (Bai et al., 2025), Llama 3.2 (Grattafiori et al., 2024), Glm-4v (Team et al., 2025) and five closed (commercial) models–Claude 3, Gemini 1.5 Flash, GPT5-mini (OpenAI, 2025), Grok4 (X.AI, 2025) and Gemini 2.5 pro (Deepmind, 2025). Figure 5 shows details of an example running through the models that we have experimented with.

We perform instruction finetuning for LLMs on the training data using LoRA (Hu et al., 2022). We use AdamW (Loshchilov and Hutter, 2019) as the optimizer, with the initial learning rate, β_1, β_2, and batch size as 2e-05, 0.9, 0.999, and 1, respectively. The attention dimension and α of LoRA are 4 and 2, respectively. Additionally, we conduct experiments using Chain-of-thought prompting (Wei et al., 2022) which has been proven useful in many NLP tasks, such as arithmetic and symbolic reasoning. Specifically, to elicit LLM's duration commonsense knowledge, for each exemplar used in the few-shot setting, we use 3-5 sentences in plain English to briefly explain how the label is inferred before presenting the answer.

Example starts
Read the tweet and determine how long the author of the tweet will stay at Manhattan after posting this tweet.

Tweet text: This Corona Virus Shit has GONE TOO FAR!! They closed the Chicken Spot and the Chinese Spot. THIS IS MADNESS!!! WE'RE ALL GONNA DIE!!! #CoronaVirus #NYC #Harlem #TheHood #ChickenSpot #ChineseFood #TheWorldHasGoneMad...

Tweet image:

Options:

1. Less than 24 hours.
2. 1-30 days.
3. More than 30 days.

Answer: more than 30 days. The reason is that the tweet's hashtags include #NYC, #Harlem, and #TheHood. The author complains about "the Chicken Spot" being closed, which suggests it is a familiar, local place. This indicates the author is a resident of that neighborhood, not a tourist. Therefore, they are likely living there and will stay for more than 30 days.
Example ends

Read the tweet and determine how long the author of the tweet will stay at Reading after posting this tweet.

Tweet text: Corona virus approved activities at the Glass House! #COVID19 #backtobasics #puzzlesrock #winelife #deerfootwinery #staycationfunwiththewife @ Wyomissing Park, Reading

Tweet image:

Options:

1. Less than 24 hours.
2. 1-30 days.
3. More than 30 days.

Answer:

Figure 6: A detailed prompt we used in the VLM experiment.