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 into an Airbnb more likely means staying for a few days. Such duration judgments are important for understanding human mobility, travel behavior, and social context. However, duration is rarely stated explicitly in social media posts and must instead be inferred from noisy textual, visual, and contextual cues. In this paper, we investigate whether vision-language models possess such duration commonsense. We create a new corpus of geotagged tweets co-annotated by humans and LLMs to serve as a testbed for this task. We find that state-of-the-art models struggle with duration inference, even when guided by multi-turn Chain-of-Thought prompting. Error analyses further reveal that models often fail to integrate critical contextual information, or they overly rely on prominent visual details while overlooking subtle hashtags and semantic nuances. These findings underscore the significant challenges remaining in context-aware multimodal understanding.

1 Introduction

Duration commonsense, the ability to infer how long human activities last, is fundamental to human intelligence and allows us to navigate temporal contexts. Extracting duration commonsense from social media benefits applications such as transportation planning and personalized recommendations. For example, inferring how long users stay at checked-in locations helps estimate travel demand. However, because this inference requires interpreting nuanced social, visual, and textual cues, developing computational models to replicate this human ability remains challenging.

Consider the tweets in Figure 1. The author of the left tweet will likely stay briefly in *Morro Bay*: the text recounts kayaking before heading back, and the image shows the author in a kayak with life jackets during a day trip. Conversely, the right tweet suggests a stay of months or years in *New Orleans*; the indoor setting, pet, and references to sheltering in place indicate the author is living there during the lockdown.

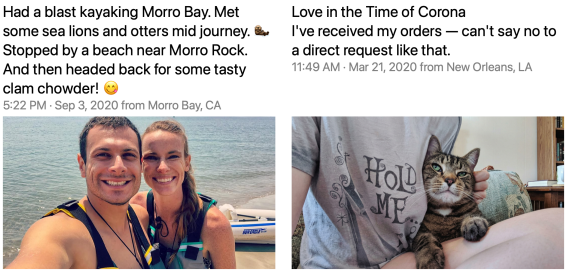

Figure 1: Examples of tweets in which the author will stay at the geotagged location for a short (left) or long (right) time. The author of the left tweet will stay at *Morro Bay* only briefly: the text recounts kayaking before heading back, and the image shows the author in a kayak during a day trip. The author of the right tweet will stay at *New Orleans* for a long time; the "Love in the Time of Corona" post, indoor setting, and pet suggest the author is sheltering at home during the COVID-19 lockdown.

Prior duration research focuses on event duration inference (Gusev et al., 2011; Leeuwenberg and Moens, 2018), predicting how long events last rather than how long individuals participate. While some work extracts human-centric duration from formal texts (Vempala and Blanco, 2016; Chinnappa et al., 2020), social media shares human experiences through noisy, multimodal posts where duration is rarely explicit. Inferring stay duration at a specific location requires reasoning across informal text, hashtags, and visual elements, a socially nuanced task that remains largely underexplored.

In this paper, we focus on this gap by investigating 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 (formerly known as Twitter). 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 images 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.

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 systematically evaluate the **pretrained language models (PTLMs), text-only LLMs, and VLMs**. Our results show that while multi-turn Chain-of-Thought prompting and advanced reasoning models significantly improve performance, they struggle with highly ambiguous and culturally nuanced cues; (d) We perform detailed error analyses, providing **insights** into the fundamental bottlenecks of multimodal duration inference. We find that complex semantic reasoning remains the primary challenge, as models frequently overlook contextual constraints or conflate disjoint visual and textual clues.

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 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

[1] Corpus and code available at https://anonymous.4open.science/r/loc_tmp3-2327/

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

Figure 2: Annotation examples. The author of the leftmost tweet is departing *SeaTac* ("Hopefully I will be safe this time" in the text and the crowded airport waiting area in the image), indicating a stay of `0-24 hours`. The middle tweet exemplifies `1-30 days`: the author visits the Gateway Arch on a #roadtrip to *St Louis*, and travelers typically spend less than one week at each stop. The author of the rightmost tweet lives in *Halawa*; the "Yep, we're eating again" text and a home-cooked meal in the image, together with hashtags such as #whatsfordinner and #onogrindz, suggest everyday residence rather than a brief visit, leading to `more than 30 days`.

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 use GeoCov19Tweets (Lamsal, 2020), containing approximately 420,000 tweets. We selected this dataset because its metadata provides precise geotags and it captures a wide range of stay durations—from brief essential outings to prolonged home stays. Although pandemic-era data exhibits behavioral biases, the underlying multimodal reasoning required to infer duration remains generalizable. Figure 2 illustrates annotation examples spanning all three duration labels, from a traveler leaving *SeaTac* to a #roadtrip stop in *St Louis* and everyday home life in *Halawa*. We downloaded the full metadata for the provided tweet IDs via the X Developer API.

Tweets Filtering We filtered the dataset using three criteria: First, we retained only tweets with US city-level geotags, excluding broader state-level or narrower point-of-interest tags. Second, each tweet must contain both text and images. Finally, we excluded corporate or organizational accounts using keyword matching on usernames (e.g., *daily*, *sales*) and poster-like image detection via CRAFT (Baek et al., 2019).

Additionally, we removed tweets containing explicit duration phrases (e.g., *5 days*, *one year*) to ensure the task evaluates duration inference rather than explicit extraction.

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 haiku, 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*. The instructions that were given to the annotators including the disclaimer, demographic questions, and an example task can be found in Figure 6 - 8 in the Appendix.

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. Detailed examples of the prompt can be found in Table 7.

3.3 Annotation Quality Control of Crowdworkers and LLMs

Ensuring annotation quality requires rigorous control for both human and LLM annotators. For crowdworkers, we implement two initial filters based on pilot observations: we recruit only workers with historical approval rates above 95%, and we enforce a minimum completion time of 5 seconds per Human Intelligence Task (HIT).

We collect five human annotations per HIT and filter unreliable responses to achieve *substantial* inter-annotator agreement. We quantify reliability using Krippendorff's α (Krippendorff, 2011). To ensure quality, we iteratively apply the following filtering process to each batch of human annotations until reaching $\alpha \geq 0.6$:

1. Calculate the competence score for each crowdworker using Multi-Annotator Competence Estimation (MACE) (Hovy et al., 2013).
2. Discard all annotations from the crowdworker with the lowest competence score.
3. Recompute Krippendorff's α on the remaining annotations.

This iterative process yields a final average Krippendorff's α of 0.68 across all human annotations. After filtering, we derive the aggregated human labels using MACE to weight the remaining annotators by their inferred competence, avoiding the limitations of simple majority voting.

For LLM annotators, we assess inter-rater agreement using Fleiss' Kappa (FK) (Fleiss, 1971). Table 1 reports the κ values for each model. We treat models with $\kappa < 0.2$ (Qwen 2.5 and GPT5-mini) as low-quality annotators and exclude them. We then derive the final adjudicated LLM labels by applying MACE to the predictions of the remaining four models.

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 haiku	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

To construct a reliable corpus, we must resolve conflicts between human and LLM consistency. Table 2 presents the joint distribution of human and LLM predictions. The highest agreement occurs for the `more than 30 days` category, demonstrating that both annotator types share baseline knowledge regarding prolonged quarantine stays.

LLM	Human		
	0-24 hours	1-30 days	more than 30 days
0-24 hours	**2.00**	0.73	18.36
1-30 days	4.57	**2.65**	5.40
more than 30 days	32.23	2.37	**31.69**

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

While Table 2 demonstrates baseline consensus on the diagonal, the off-diagonal cells reveal substantial discrepancies, most notably when one annotator predicts a brief stay (`0-24 hours`) while the other infers a long-term stay (`more than 30 days`). To resolve conflicts, 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

R1: COVID-19 quarantine duration commonsense	**R2: Duration not calculated at the city level**	**R3: Tweet intent unclear**	**R4: Author residency status unclear**
Quarantine Birthday for me. #lockdown #coronavirus #carrotcake #nyc #manhattan #cumplea 3:55 AM Apr 28, 2020 from Manhattan, NY	VVBS buying @tahlequahumc #vvbs #vbs2020 @ Tahlequah United Methodist Church 3:26 PM Jul 1, 2020 from Tahlequah, OK	gloom my day #boardwalk-closed #fencedin #locked #gates #cold #covid_19 #epedemic #truphotos 4:13 PM Mar 29, 2020 from Seaside Heights, NJ	Im a Barbie girl... in a Covid world #transisbeautiful #girlslikeus #transvisibility #saturdayvibes #saturdaynight #covid19 12:42 AM Sep 13, 2020 from Dallas, TX
LLM: `1-30 days` The reason is that the tweet text explicitly 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 is about buying supplies for a Vacation Bible School (#vbs2020) at the Tahlequah United Methodist Church (@tahlequahumc). Organizing and participating in such a local church event strongly suggests that the person is a resident of Tahlequah or the surrounding area, not a temporary visitor. Therefore, they are likely to be in the city for an extended period.	**LLM:** `more than 30 days` 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.	**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.
Human: `more than 30 days`	**Human:** `0-24 hours`	**Human:** `0-24 hours`	**Human:** `more than 30 days`

Table 3: 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>	8(H)	0	0	51(D)	59
<more than 30 days, 0-24 hours>	0	266(L)	87(D)	0	353
<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.

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 to some places rather than an actual visit.
- R4: Author residency status unclear. It's hard to interpret whether the author is a tourist or a resident.

The first two reasons (R1 and R2) represent the most common errors made by the LLM or human annotator, where the predictions within 30 days can be directly rejected. R3 and R4, being more ambiguous, cannot be definitively resolved, so we discarded those tweets. Examples of these cases are shown in Table 3. 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. In the rest of this paper, we work with these filtered tweets.

4 Corpus Analysis

After human–LLM alignment and filtering (Section 3), the final corpus contains 4,312 geotagged tweets with adjudicated duration labels. Figure 3 shows the label distribution. Each instance pairs a tweet with its geotagged US city; we exclude the `I cannot determine` label (<1% of instances) from all analyses and experiments because it does not reflect a usable duration judgment. Roughly 70% of tweets are labeled `more than 30 days`. Labels of the remaining tweets

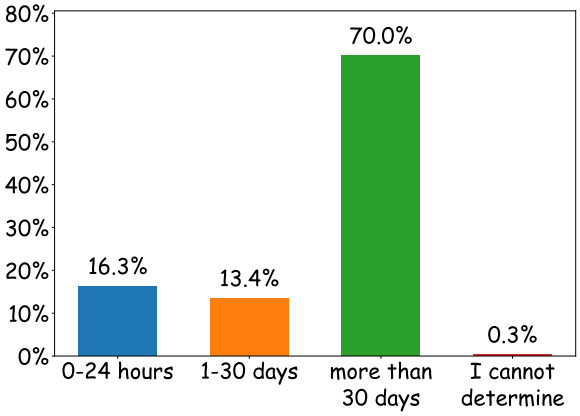

Figure 3: Label distribution.

are nearly evenly split between `1-30 days` and `0-24 hours`.

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. PTLMs are fine-tuned on the training set and then evaluated on the test set; LLMs and VLMs are evaluated on the test set without fine-tuning. The evaluation of different models is categorized into five types: **(1) Human:** The adjudicated labels by AMT. **(2) PTLM:** We evaluate pretrained encoder-based baselines using RoBERTa-Large (Liu et al., 2019) as the text encoder and ViT-Large/16 (Dosovitskiy et al., 2021) as the image encoder. We test text-only and image-only variants, as well as multimodal variants that combine text and image representations by concatenation, addition, and element-wise multiplication. The resulting representation is passed through a two-layer classifier. We also evaluate CLIP (Radford et al., 2021) as a jointly pretrained image–text baseline. **(3) LLM (Text Only):** We evaluate Qwen 2.5 (Bai et al., 2025) and Llama 3.2 (Grattafiori et al., 2024) using only the tweet text, image caption and the geotagged city. **(4) VLM (w/o deep thinking):** We use Claude 3 haiku and Gemini 1.5 Flash as proprietary VLMs on the held-out test set. **(5) VLM(w/ deep thinking) :** We evaluate GPT-5-mini (OpenAI, 2025), Grok4, Gemini 2.5 Pro (Deepmind, 2025), Qwen2.5-VL (Bai et al., 2025), Llama 3.2 Vision Instruct (Grattafiori et al., 2024), and GLM-4V(Team et al., 2025).

To bridge the gap between complex multimodal understanding and logical deduction, we incorporate Multi-turn Chain-of-Thought (CoT) prompting in our evaluations. By initially prompting the models to describe visual elements and their relation to textual cues, CoT creates an intermediate reasoning trace that explicitly grounds the final duration prediction. This motivation aligns with recent findings that breaking down complex reasoning tasks enhances LLM robustness. See more details in Appendix C and D.

5.1 Results

We benchmark the duration inference capabilities across all model families, from unimodal baselines to state-of-the-art multimodal reasoning models. Table 5 presents the comprehensive evaluation results.

Effect of multi-modality. For PTLM baselines, combining text and images improves performance. Specifically, CLIP outperforms models pretrained on only text or images, indicating that joint vision-language pretraining effectively bridges the modality gap. For LLMs, the Llama 3.2 model performs slightly better some non-deep-thinking VLMs. However, its reliance on generated image captions creates a bottleneck, preventing it from matching the accuracy of advanced deep-thinking VLMs. The reason could be it struggle to fully interpret the nuanced semantic duration information that requires direct visual grounding.

Effect of in-context learning and Multi-turn CoT. Transitioning from the zero-shot to the few-shot setting consistently improves performance across virtually all models, highlighting the importance of exemplars for complex duration reasoning. For instance, the F1 score of Gemini 1.5 Flash increases from 0.33 (zero-shot) to 0.47 (few-shot). Furthermore, Multi-turn Chain-of-Thought (CoT) prompting provides an additional performance boost. GLM-4V achieves the highest overall F1 score of 0.85 in the few-shot with Multi-turn CoT setting. Guiding the model to explicitly verbalize its visual and semantic observations before predicting the duration significantly reduces reasoning errors.

Bottleneck of deep-thinking models. Models equipped with advanced reasoning capabilities

Methods	Backbone	Input	ZS-ST	FS-ST	FS-MT
Human	Human reference	human judgment		0.37	
PTLM	RoBERTa-Large *340M*	text	0.18	0.21	–
	ViT-Large/16	image	0.05	0.07	–
	RoBERTa-ViT-Concatenation	text + image	0.22	0.24	–
	RoBERTa-ViT-Addition	text + image	0.16	0.17	–
	RoBERTa-ViT-Multiplication	text + image	0.13	0.15	–
	CLIP	text + image	0.38	0.38	–
LLM	Qwen 2.5	text	0.28	0.32	0.35
	Llama 3.2	text	0.37	0.40	0.43
VLM (w/o deep thinking)	Claude 3 Haiku	text + image	0.32	0.35	0.46
	Gemini 1.5 Flash	text + image	0.33	0.40	0.47
VLM (w/ deep thinking)	GPT5-mini	text + image	0.62	0.68	0.71
	Grok4	text + image	0.70	0.77	0.82
	Gemini 2.5 Pro	text + image	**0.72**	0.76	0.83
	Qwen 2.5VL	text + image	0.71	0.80	0.84
	Llama 3.2 Vision Instruct	text + image	0.65	0.69	0.73
	GLM-4V	text + image	0.70	**0.81**	**0.86**

Table 5: Weighted-average F1 scores of different model families on duration inference. ZS-ST, FS-ST, and FS-MT denote zero-shot single-turn, few-shot single-turn, and few-shot multi-turn settings, respectively. Text-only LLMs use tweet captions as input. The multi-turn setting separates visual analysis and duration inference into two sequential dialogue turns. The best score within each method group is underlined, and the best score in each evaluation setting is **boldfaced**. A dash indicates that the setting is not applicable. Further details are provided in Appendix D and Table 7.

(e.g., Qwen 2.5VL, GLM-4V, Grok4, GPT-5-mini, and Gemini 2.5 Pro) substantially outperform standard proprietary models (e.g., Claude 3 haiku and Gemini 1.5 Flash). Despite this, F1 scores plateau around 0.85 even with few-shot exemplars and Multi-turn CoT. Current models continue to encounter fundamental limitations when processing highly ambiguous or culturally nuanced commonsense cues, as detailed in Section 5.2.

5.2 Error Analysis

We systematically analyzed errors across all evaluated models to determine how input influences duration interpretation. Table 6 provides representative examples; we group failures into three patterns. (1) **Context:** Models ignored contextual cues, such as hashtags or geotags. In the first example, failing to recognize the hashtags which indicates local residency cause the model to underestimate duration based on the activity alone. (2) **Visual:** Models overlooked subtle visual elements. In the second example, the model predicted `0-24 hours` from the wish that #oceangallery was open, missing the beach walk in the image. (3) **Semantic:** During complex reasoning, models drifted from established facts. In the third example, even when correctly identifying a scenario (hosting an out-of-town friend), the model conflated the visitor's stay duration with the user's residency status. Table 6 provides representative examples.

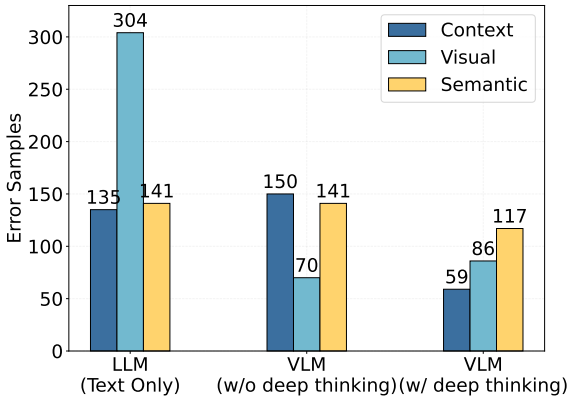

Figure 4: Distribution of error types across model categories

Figure 4 details error distributions across model architectures. Visual errors dominated LLM (Text Only) models, reflecting their lack of visual grounding. VLM (w/o deep thinking) models resolved many visual errors but primarily struggled with context errors. The reason could be that although they perceive visual features, they lack the reasoning capacity to synthesize those visuals with disjoint textual hints like geotags or specific hash-

Context	Visual	Semantic
I carried the memories of my dad on a Social Distancing (SoDi) Half Marathon for Father's Day. #MNStrong #AllOutin2020 #Minnesota #FathersDay #fitgay 3:12 AM Jun 23, 2020 from St Joseph, MN	Love me a good day in #oceancitymaryland. The weather was perfect for a walk. Most folks were #socialdistancing. Just wish #oceangallery was open. The #clouds and sky were amazing. 12:35 AM Jun 21, 2020 from Ocean City, MD	Glad to see one of my best friends Raul from Chicago and college in town! #LewisUniversity #Chicago #Flyers #Botines #Liverpool @ Corona del Mar State Beach 12:29 AM Aug 18, 2020 from Newport Beach, CA
Generated: `0-24 hours` The image and the tweet describe the author completing a half marathon run on a trail for Father's Day. This is a specific, time-bound activity that typically lasts a few hours. After finishing the run and posting the photo, it is most likely that the author would return home. The nature of the activity does not suggest a multi-day vacation or residency.	**Generated:** `0-24 hours` The tweet states, "Just wish #oceangallery was open". This implies that the author had intended to visit the Ocean Gallery during their current stay but couldn't because it was closed. The sentiment of "just wish" suggests a missed opportunity that they won't be able to fulfill on this trip, strongly indicating that their stay in Ocean City is 0 hours.	**Generated:** `1-30 days` The reason is that the tweet mentions the author is meeting his friend from Chicago who is "in town". A friend visiting from another state usually stays for a few days or a short vacation, which falls within the 1–30 day range. It is unlikely to be a visit for less than 24 hours due to the travel distance, and there is no indication of a long-term stay of over 30 days.
Gold: `more than 30 days` The reason is that the tweet indicates the author is running a half marathon in Minnesota, with hashtags like #MNStrong and #Minnesota, suggesting he is a local resident. Therefore, it is highly probable that he lives in or near St. Joseph and will remain there for an extended period.	**Gold:** `1-30 days` The reason is that the tweet mentions having "a good day in #oceancitymaryland" and wishing a local gallery was open. The image shows that weather was good for walking, most people were keeping distance from each other. This suggest that the author is on a vacation or a short trip, which usually lasts from a few days to a couple of weeks.	**Gold:** `more than 30 days` The tweet states, "Glad to see one of my best friends Raul from Chicago and college in town!" The phrase "in town" implies that Raul is visiting the city where the author is located. This suggests that the author either resides in Newport Beach/Corona del Mar or has a prolonged presence there. If the author lives in the city, their stay is continuous and indefinite, which falls under "more than 30 days".
Analysis: Fails to recognize the hashtag #MNStrong, which indicates he is a local resident.	**Analysis:** Fails to distinguish the actual beach walk, which is visible in the image, from the unfulfilled Ocean Gallery visit.	**Analysis:** Correctly identifies the scenario as the author hosting a friend from another state, but incorrectly reports the friend's duration of stay as the final answer.

Table 6: Error analysis of representative LLM predictions.

tags. Furthermore, semantic errors do not decrease with using non-deep-thinking VLMs compared to text-only LLMs. While VLM (w/ deep thinking) models minimized contextual and visual mistakes, their semantic errors remained high. Therefore, semantic reasoning remains the primary bottleneck for duration inference.

6 Conclusions

We studied duration commonsense understanding. Specifically, we introduced a task: determining how long tweet authors stay at tagged locations. To support this task, we created a corpus of 4,312 geotagged tweets. Corpus analysis shows that inferring duration requires extensive commonsense reasoning; for example, a flight departure implies a stay of hours, while visiting the Gateway Arch implies days. To the best of our knowledge, we are the first to tackle this challenging problem. Our experiments show that accurate inference requires processing highly ambiguous, culturally nuanced commonsense cues and maintaining logical consistency across disjoint visual and textual signals. Our error analysis demonstrates that, even for deep-thinking VLMs, complex semantic reasoning remains a fundamental limitation.

Limitation

Corpus 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 is primarily collected from X and 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.

Annotations Our labels are based on inferred duration commonsense rather than directly observed future movement. Tweets often express sparse, ambiguous, or indirect intent, making fine-grained duration annotation unreliable. We therefore use three coarse labels: `0-24 hours`, `1-30 days`, and `more than 30 days`, corresponding roughly to leaving soon, temporarily staying, and long-term residence. The human–LLM co-annotation process improves consistency, but it does not eliminate annotation bias. Duration cues in tweets often depend on subjective interpretation, background knowledge, and assumptions about the author's activity or residency status. In some cases, no single label is clearly correct. Future work could investigate more objective annotation protocols, collect longitudinal evidence when available, and evaluate whether finer-grained or uncertainty-aware labels better capture duration-related commonsense.

Ethical Considerations

We take several steps to reduce privacy risks associated with using social media data. First, the released corpus will exclude direct user identifiers, including usernames, profile images, and tweet IDs. Second, we provide a takedown request process through an online form, following Mirowski et al. (2019), so that individuals can request removal of specific examples from the corpus. We also protect crowdworker privacy. Demographic information is self-reported during the qualification task only to help us understand potential sources of annotation bias. We do not release individual-level demographic information. Access to raw demographic responses is restricted to the research team, and any demographic statistics are reported only in aggregate form to avoid identifying individual workers. Crowdworkers are compensated through Amazon Mechanical Turk according to the platform's norms and requirements.

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.

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 Sangani, 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.

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.

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.

Minheng Ni, Yutao Fan, Lei Zhang, and Wangmeng Zuo. 2024. Visual-o1: Understanding ambiguous instructions via multi-modal multi-turn chain-of-thoughts reasoning.

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 Lingustics.

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.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

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.

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.

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 Annotation Details

Approximately 2,500 annotators participated in our crowdsourcing efforts. Figure 5 summarizes their demographic information. Collecting annotations from a diverse pool of crowdworkers was essential because duration interpretation varies across age groups and educational backgrounds. To ensure high-quality labels for this ambiguous task, we implemented rigorous quality control. We discarded approximately 30% of the annotations collected through Amazon Mechanical Turk (AMT) because they failed sanity checks or exhibited random clicking behavior. We compensated crowdworkers above the United States federal minimum wage of $7.25 per hour, adhering to platform guidelines. After adjudication and filtering, each tweet in our corpus retains an average of 2.5 high-quality annotations.

To ensure annotators understood the task, we provided comprehensive guidelines defining duration commonsense and the four label categories (Figure 6). Following the guidelines, workers completed a demographic questionnaire covering age and education level to help us analyze annotator diversity (Figure 7). Figure 8 illustrates the main annotation interface, which displays a tweet's text, image, and geotag alongside the multiple-choice duration question.

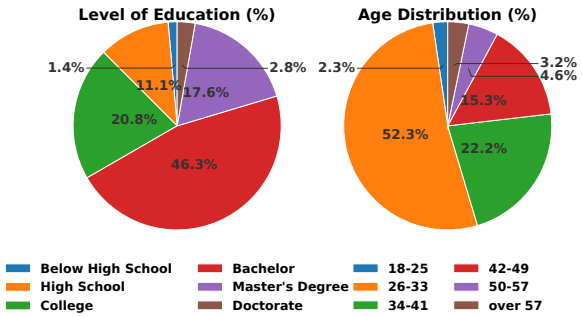

Figure 5: Annotator demographic statistics from AMT

B 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.

C Experiment setting for pretrained language models

We use RoBERTa-Large *340M* (Liu et al., 2019) as the text encoder and Vision Trans-

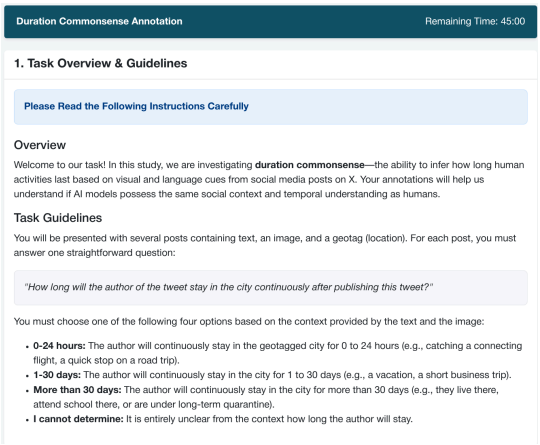

Figure 6: Task guidelines within our Amazon Mechanical Turk qualification task

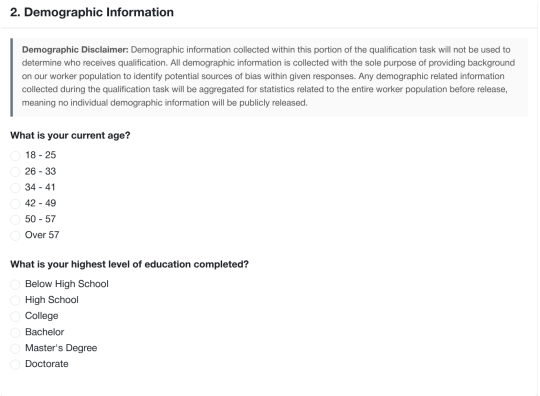

Figure 7: Demographic questions within our Amazon Mechanical Turk qualification task.

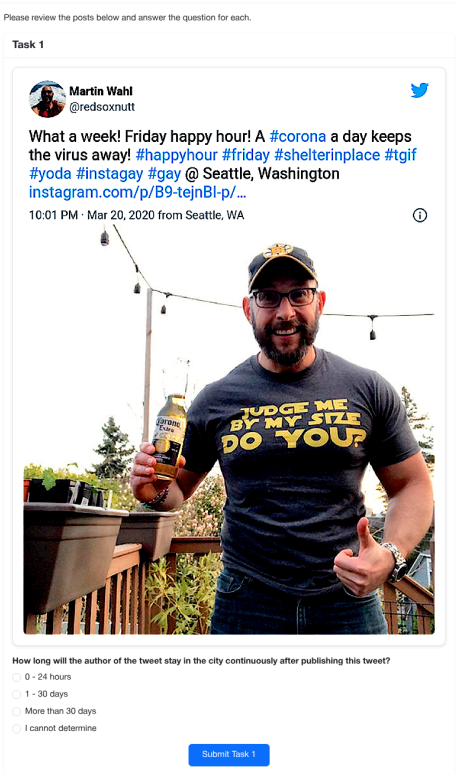

Figure 8: Example of main task for our Amazon Mechanical Turk

former (Dosovitskiy et al., 2021) as the image encoder as shown in Figure 9. [2] For multimodal variants, we encode tweet text and the first attached image with the same backbones and fuse representations by concatenation, element-wise addition, or element-wise multiplication before classification. The CLIP baseline uses `openai/clip-vit-base-patch32` (Radford et al., 2021) for both modalities. Each variant feeds the merged representation through two fully connected layers (hidden size 256, output size 3) with dropout (Srivastava et al., 2014) (rate 0.2) before the final softmax. 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.

D Experiment setting for large language models and vision language models

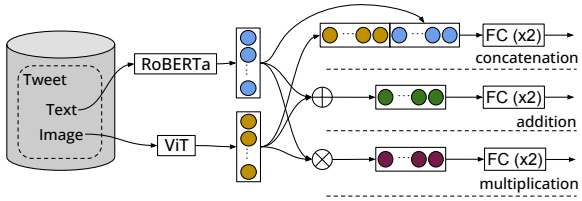

Figure 9: Architecture of PTLM baselines. Tweet text and images are encoded by pretrained RoBERTa and ViT backbones; fused representations pass through two fully connected layers for classification.

To analyze the capability of recent large models (with at least 1 billion parameters) on duration commonsense understanding, we conduct extensive experiments with modern large language models(LLMs) and vision language models (VLMs). To ensure the coverage of VLMs, we choose to work with 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 haiku, Gemini 1.5 Flash, GPT5-mini (OpenAI, 2025), Grok4 (X.AI, 2025) and Gemini 2.5 pro (Deepmind, 2025). Figure 7 shows details of an example

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

Model	Content	Single-turn prompt	Multiturn prompt
LLM (Text Only)	**Tweet text:** Corona virus approved activities at the Glass House! #COVID19 #backtobasics #puzzlesrock #winelife #deerfootwinery #staycationfunwiththewife @ Wyomissing Park, Reading **Tweet image description:** an indoor home setting with a jigsaw puzzle spread across the floor, wine glasses, and a decorated Christmas tree	*System / user prompt:* You are a helpful assistant for inferring the duration of human activities. Please read the tweet and determine how long the author of the tweet will stay in **Reading** after posting this tweet. **Options:** 1. 0–24 hours 2. 1–30 days 3. more than 30 days **Answer: 2. 1–30 days.** The hashtag #staycationfunwiththewife suggests a temporary leisure stay rather than permanent residence. The tweet describes home-based activities such as doing puzzles and drinking wine, which are compatible with a short staycation lasting multiple days but do not by themselves imply long-term residence.	*Round 1 prompt:* You are a helpful assistant for inferring the duration of human activities. Analyze how the tweet's image description relates to its text and hashtags. **Answer:** The image description validates the tweet's narrative: a pandemic-safe, low-tech, couple-focused staycation. The immediate scene (puzzle + wine) likely represents a 2–4 hour evening activity, but the hashtag #staycationfunwiththewife suggests this isn't a one-off event but a multi-day to week-long pattern of home-based leisure *Round 2 prompt:* Please read the tweet and determine how long the author of the tweet will stay in **Reading** after posting this tweet. **Options:** 1. 0–24 hours 2. 1–30 days 3. more than 30 days **Answer: 2. 1–30 days.** The combination of the #staycation hashtag, pandemic-era context, and home-based leisure activities strongly implies the author is settled in Reading for a short-to-medium-term local break. The most evidence-supported inference is a stay lasting between 1 and 30 days after posting.
VLM	**Tweet text:** Corona virus approved activities at the Glass House! #COVID19 #backtobasics #puzzlesrock #winelife #deerfootwinery #staycationfunwiththewife @ Wyomissing Park, Reading **Tweet image:** 	*System / user prompt:* You are a helpful assistant for inferring the duration of human activities. Please read the tweet and determine how long the author of the tweet will stay in **Reading** after posting this tweet. **Options:** 1. 0–24 hours 2. 1–30 days 3. more than 30 days **Answer: 3. more than 30 days.** The tweet refers to "Corona virus approved activities" and includes #COVID19, suggesting stay-at-home behavior during the pandemic. The image shows an indoor domestic setting, and #staycationfunwiththewife further implies that the author is spending time at or near home. Taken together, these cues suggest that the author is likely a resident of Reading and will remain there for an extended period.	*Round 1 prompt:* You are a helpful assistant for inferring the duration of human activities. Describe the tweet image in detail, and analyze its relationship to the tweet text and hashtags. **Answer:** The image shows an indoor home environment with wooden flooring. A large jigsaw puzzle is spread across the floor, and wine glasses are visible nearby. These visual details align with #puzzlesrock and #winelife. The hashtags #COVID19 and #staycationfunwiththewife suggest home-based leisure during the pandemic, while #backtobasics indicates simple indoor activities. *Round 2 prompt:* Please read the tweet and determine how long the author of the tweet will stay in **Reading** after posting this tweet. **Options:** 1. 0–24 hours 2. 1–30 days 3. more than 30 days **Answer: 3. more than 30 days.** The tweet and image together indicate a home-based staycation during COVID-19. Because the activity appears to take place in the author's home environment, the author is likely staying in Reading long-term rather than visiting briefly.

Table 7: Representative prompts used for different model settings. For text-only models, the image is converted into a textual description. For VLM models, the original image is provided directly. In the multiturn CoT setting, the model first interprets the image and its relation to the tweet, and then predicts the duration label.

running through the models that we have experimented with.

Additionally, to address the challenge of complex visual-linguistic alignment, we conduct experiments using Multiturn Chain-of-thought (CoT) prompting (Ni et al., 2024). This technique has been proven highly effective in bridging reasoning gaps in NLP and multimodal tasks. Our motivation is that jumping directly to a duration label forces the model to implicitly resolve ambiguity in one step. Instead, by first explicitly eliciting a detailed description of the tweet's image and its relationship with the text, Multiturn CoT grounds the model's reasoning. For each exemplar used in the experiment, we provide prompt and response that systematically walk through how the visual and textual hints lead to the final duration inference.

The first round of dialogue
Describe the tweet image in detail, and analyze its relationship with the tweet's text and hashtags.

The second round of dialogue
{question with user prompt}

E Use of AI Statement

We used generative AI tools to assist implementation and manuscript writing. Specifically, we used Cursor, an AI-assisted development environment, and ChatGPT (GPT-5.3) to help implement evalua-

tion tasks. We also used ChatGPT to help format portions of the manuscript and draft LaTeX tables. We reviewed, edited, and verified all AI-assisted output and take full responsibility for the accuracy, originality, and integrity of the work.