World Knowledge Resolves Some Aspectual Ambiguity

Katarzyna Pruś and **Mark Steedman** and **Adam Lopez**
School of Informatics
University of Edinburgh
Edinburgh, Scotland
k.prus@ed.ac.uk, {steedman, alopez}@inf.ed.ac.uk

Abstract

Annotating event descriptions with their aspectual features is often seen as a pre-requisite to temporal reasoning. However, a recent study by Pruś et al. (2024) has shown that non-experts' annotations of the aspectual class of English verb phrases can disagree with both expert linguistic annotations and each another. They hypothesised that people use their world knowledge to tacitly conjure their own contexts, leading to disagreement between them. In this paper, we test that hypothesis by adding context to Pruś et al.'s examples and mirroring their experiment. Our results show that whilst their hypothesis explains some of the disagreement, some examples continue to yield divided responses even with the additional context. Finally, we show that outputs from GPT-4, despite to some degree capturing the aspectual class division, are not an accurate predictor of human answers.

1 Introduction

Aspect as a linguistic category refers to analysing situation and action descriptions in terms of their internal structure. For example, *to unpack a bag* differs structurally from *to play with toys* in that the former one culminates at a well defined endpoint whilst the latter one does not. Aspectual properties of a predicate are consequential to the entailments one can make regarding the action it describes. In the English language specifically, aspectual class is told to determine whether a given predicate's past progressive (PP) form entails its past simple (PS) or not. For example, we note the following:

(1) *I was playing with toys* $\rightarrow$ *I played with toys*

(2) *I was unpacking a bag* $\not\rightarrow$ *I unpacked a bag*

Many works attempted to either tackle automatic aspectual classification as a task of its own (Siegel and McKeown, 2000; Friedrich and Palmer, 2014; Friedrich and Gateva, 2017; Kober et al., 2020) or showed the relevance of aspectual features in other tasks such as event ordering (Chambers et al., 2014), image captioning (Alikhani and Stone, 2019) and natural language inference (NLI; Kober et al., 2019). Having said that, annotating datasets for aspectual features is particularly challenging because it requires the annotators to be domain specific experts (Friedrich et al., 2023). Framing the data collection as a crowd sourcing task, whilst possible, results in high level of disagreement. Pruś et al. (2024) point out that such disagreement need not be a problem, as it encourages the modelling of common-sense inferences (Pavlick and Kwiatkowski, 2019). In their study, Pruś et al. turned predicates into pairs of sentences – one in PP form and one in PS form – and then asked participants if they agree that PS can be inferred from PP. For example, for a predicate *boil an egg*, they asked whether *I was boiling an egg* being true, means that *I boiled an egg* is also true. They showed that for many predicates, despite their aspectual features pointing towards 'No' being the theoretically predicted answer, participant responses were split between 'Yes' and 'No'. They hypothesised that such common-sense inferences can be driven by participants not conceiving of scenarios where certain actions do not get completed. In practice, *boiling an egg* rarely ever is left unfinished.

In this paper we test that hypothesis. Whilst Pruś et al. kept their examples purposefully decontextualised, we expand the PP sentences to include context indicating that the action was interrupted or abandoned. For the *boil an egg* example, our PP sentence with an interruption was *I was boiling an egg, but my camping stove ran out of fuel.* We then mirror all of their study design details and compare the judgments collected between the two settings.

We show that providing the additional context led to increased number of participants answering in accordance with theoretical predictions for some

Findings of the Association for Computational Linguistics: ACL 2025, pages 13207–13220
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

— but not all — predicate examples. Whilst this supports the hypothesis of Pruś et al., it also reveals that there are more factors at play in shaping people's answers. The fact that certain predicates elicited mixed responses even in the setting with the additional context, tells us that aspectual class of certain predicates may be open to individual interpretation.

Finally, we present an experiment querying GPT-4 to generate responses for the same questions that were shown to the human participants. We observe that its response patterns might follows some general trends of human responses on the aspectual class level. However, looking at individual predicates we find that the distributions of GPT-4 outputs differ considerably from the distributions of human answers.

2 Background

Aspect is a linguistic concept used to characterise situation descriptions. Whilst related to tense, aspect is less concerned with *when* something occurred, but rather *how* it developed over time. It is worth noting that there exists two distinct categories referred to as aspect: grammatical and semantic. Here we focus on the semantic category, which is often referred to as **Aktionsart**.

Vendler (1967) divides situations into 4 main categories: **States**, **Achievements**, **Accomplishments** and **Activities**. Those categories are known as **aspectual classes**. Subsequent literature analysed these classes in terms of three different properties: stativity, durativity and telicity (Moens and Steedman, 1988; Pustejovsky, 1991).

Stativity differentiates between *stative* and *dynamic* eventualities. Broadly speaking, the term *stative* refers to situations that do not require effort to be sustained. All states are stative and they are contrasted with dynamic **Events**. For example, *to know the answer* is a state, but *to jump over the fence* is an event. Achievements, Accomplishments and Activities are all different types of events.

Durativity pertains to whether something occurs instantaneously (is *punctual*) or extends across time (is *durative*). For example, *to dance* is durative and *to land* is punctual. States are all durative, whilst events can be either durative or punctual. In fact, what distinguishes Accomplishments from Achievements is durativity.

Telicity takes on values *telic* or *atelic*. It describes whether the action is leading to a necessary endpoint, beyond which it cannot continue (Comrie, 1976, p. 45). A telic event will lead to such and endpoint, at which a change of state occurs (Rothstein, 2008). By the very definition, all States are atelic. Events, however, can be either telic or atelic.

Activity *to dance*	atelic	durative
Accomplishment *to build something*	telic	
Achievement *to land*		punctual

Table 1: The three Aktionsart categories studied in this paper. Explained in terms of their features and illustrated with examples.

Aktionsart should not be viewed as a property of a verb itself, but rather of an entire verb phrase (Verkuyl, 1972). Consider, for example, the verb *to swim*, and the two contexts that it is put in below:

(a) *to swim across the lake*

(b) *to swim in a lake*

The verb phrase in (a) is telic, while (b) is atelic. Because in this case we are referring not to one word, but an entire predicate, Verkuyl suggest the term **predicational aspect**.

Finally, the term Imperfective Paradox is used to describe the phenomenon that some verb phrases hold an entailment from their past progressive form to their past simple form, whilst others do not hold that entailment. For example:

I was dancing → *I danced*

I was building something ↛ *I built something*

As described by Dowty (1979) it is Activities that hold that entailment and Accomplishments that do not. Later literature extends this analysis to say that such entailment in fact holds for any atelic predicate and does not hold for any telic predicate (Lascarides, 1991; Zucchi, 2020). In fact, linguists use the imperfective paradox as one of the tests for telicity (Siegel and McKeown, 2000). It is worth noting that this analysis of imperfective paradox is inherent to the English language as it requires the existence of past simple form which is underspecified with regards to action completion.

The work which most directly motivated this paper is the study by Pruś et al. (2024). They created a collection of predicates annotated with respect to their predicational aspect class. Later, based on the imperfective paradox, they surveyed non-expert participants on whether a past progressive of a given predicate entails the past simple or not. The underlying expectation was that if telicity truly drove how people reason about event structure, a majority of participants would say that there is an entailment for the examples pre-annotated as atelic and that there isn't for the examples pre-annotated as telic. However, what they observed is that for the majority of examples the judgements of non-expert participants were split. One of the reasons for such disagreement the authors suggested, was because some participants might use their world knowledge to employ a more common-sense approach to answering the questions. For example, when asked whether *I was eating a strawberry* means that *I ate a strawberry* is also true, they are inclined to say 'Yes' because in practice the action of eating a strawberry is rarely interrupted and abandoned. By contrast, when it came to *writing a novel*, more people readily answered 'No' because it quite prevalent in real life for this action to be abandoned without completion.

In this study we want to examine that claim and ask whether, and for which examples, the interruption and subsequent abandoning of the action being 'inconceivable' was a factor in how participants answered the imperfective paradox question. To that end, we took the examples from Pruś et al., came up with contexts that explicitly signal the action being interrupted or abandoned, and presented them to non-expert participants in a setting mirroring that of Pruś et al.

3 Human Experiment

Most of the experimental design details of this study directly follow those of Pruś et al. (2024) (hereafter also referred to as *the original study*). This is to allow for a direct comparison of the results between the two studies. Much like the original study, we present the participants with pairs of sentences constructed around a common base form verb phrase. One sentence contains the verb phrase in past progressive and the other one the same phrase in past simple. The participants are then asked to judge whether there is an entailment from the past progressive to past simple. Whilst Pruś et al. kept their examples devoid of context, we purposefully add context to the past progressive sentence signalling that the action at hand has been interrupted or abandoned. Therefore, in the discussion and interpretation of the results, we will refer to the results from the original study as **decontextualised setting** and the results of this study as the **interrupted setting**.

3.1 Stimulus Design

As a starting point, we used the predicate collection from Pruś et al. and added context to each of the examples. To keep the examples consistent, the additional context is always introduced with the conjunction 'but'[1]. Ultimately, we have a template:

> *If the sentence **I was Xing, but INT** is true, does it necessarily mean that the sentence **I Xed** is also true?*

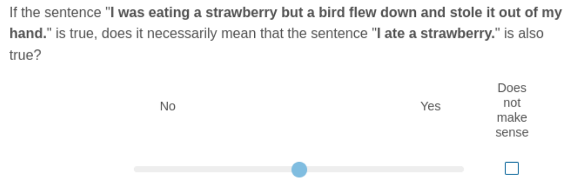
where X is the verb phrase in question and INT is the additional context. For example, for the verb phrase *eat a strawberry*, the question would look as follows:

> *If the sentence **I was eating a strawberry, but a bird flew down and stole it out of my hand** is true, does it necessarily mean that the sentence **I ate a strawberry** is also true?*

The answer 'Yes' signifies entailment (*I was Xing, but INT* → *I Xed*) and answer 'No' signifies non-entailment (*I was Xing, but INT* $\not\to$ *I Xed*).

The additional context was written manually by the authors of this paper. There were three main criteria taken into consideration at this stage. Firstly, the entire sentence has to be grammatically correct. Secondly, the introduced context has to signal that the action being described has been irreversibly interrupted or otherwise abandoned. Finally, the entire sentence has to sound plausible. This means that the interruption has to be reasonable within the context of the action described by the verb phrase in question. For example, getting injured is a reasonable interruption for walking up a mountain or catching a ball, but not a reasonable interruption for reading a book or watching television. The authors had to unanimously agree that these criteria are fulfilled for an example to be approved.

[1] We also considered the use of 'until' as a conjunction that might suggest termination of an action. Ultimately, 'but' was less restrictive and still suitable as it holds a contrastive role in discourse (Prasad et al., 2017).



If the sentence "**I was eating a strawberry but a bird flew down and stole it out of my hand.**" is true, does it necessarily mean that the sentence "**I ate a strawberry.**" is also true?

Figure 1: Screenshot of the slider interface used to gather answers. Participants were instructed that the closer to the edge they set the slider, the more confident they are in their Yes/No answer. The 'Does not make sense' checkbox was provided in case any participant judged any of the sentences as ungrammatical.

At this stage, we had to exclude 4 of the examples. The excluded examples were *cough during his talk*, *sneeze during his talks*, *faint yesterday* and *find a parking spot*. The former two are described by the atelic+punctual feature combination. Because this feature combination is very rare and the interpretation of their past progressive forms is ambiguous (Moens and Steedman, 1988; Pustejovsky, 1991), we decided to omit that category altogether in our analysis. The latter two examples, were excluded because, in our judgement, a native English speaker wanting to signal an interruption or abandonment of these actions, would use a construction *I was about to X* rather than the past progressive *I was Xing*. As such, they would not fit in with our criteria for example construction.

To collect participants' answers, we used the same slider interface as the original study, shown in Figure 1. This allowed for capturing not only the 'Yes'/'No' judgement, but also the level of certainty a participant has in their judgement. The slider maps answers as integers from -50 (signifying a certain 'No') to 50 (signifying certain 'Yes'). Values near 0 reflect a participant's lack of confidence in either answer. Participants are not shown the exact numeric value of their judgement, as this could lead them to overthink their answers. They are provided with a 'Does not make sense' checkbox that they can use instead of the slider, in case they think any of the sentences is ungrammatical or the example is otherwise nonsensical.

With the added context, we primed the participants to realise that the action could reasonably have been abandoned. For atelic examples (Activities), the added context should not influence an answer. For telic examples (Accomplishments and Achievements), we expect to see a shift towards 'No'. Pruś et al. had also included a category of Contested examples, where the authors didn't reach immediate consensus on the phrase's belonging into one of the aspectual classes. We retain these examples, but as such, we do not have a prediction of what should happen to them in the interrupted setting.

3.2 Survey Design

After discarding the aforementioned 4 examples, we were left with a collection of 46 predicates. Following the original study, we used two verb phrases — one Activity and one Achievement — as training examples . Each participant was randomly shown one of the two on the instruction page and another one as the first example on the main part of the questionnaire. Both examples are excluded from the analysis in section 3.3.

In order to minimise the risk of survey fatigue, the main questionnaire presented any given participant with half of the remaining 44 examples. The examples were picked at random for each participant in a way that made sure to include 5 Activities, 10 Accomplishments, 8 Achievements and 3 Contested examples. We collected roughly 50 responses per example[2].

3.3 Results

The distributions of answers collected in this experiment and how they compare to the distributions from Pruś et al. are shown in Figure 2.

For every example, we calculated Welch's t-test to test how strongly, if at all, the distribution from the original study differed from the one gathered in this experiment. Furthermore, we applied the Benjamini-Yekutieli procedure to control the p-values for the false discovery rate. Where the p-value is high (we adopted $p>0.01$ threshold), then there is no statistically significant difference distribution between the two settings, and so the value of the statistic itself should be disregarded. Otherwise, if the t-statistic is positive, the distribution of answers weighs heavier towards 'Yes' answers in the interrupted setting than in the decontextualised setting. If the t-statistic is negative, the distribution of answers weighs heavier towards 'No' answers in the interrupted setting than in the decontextualised setting.

Cases where t-statistic is positive and $p<0.01$, would be interpreted as as a *shift towards 'Yes'*.

[2] Due to the the implementation details of the random example selector, there is between 48 and 51 responses for each example

However, unsurprisingly, we observe no such examples.

We refer to the cases where p<0.01 and t-statistic is negative as a *shift towards 'No'*, and the cases where p>0.01 as *no significant difference*. Amongst Accomplishments the outcome is mixed - almost half of the examples show a shift towards 'No' whilst the rest show no significant difference between the two settings. Amongst Achievements, almost all examples show a shift towards 'No'. Amongst Activities, almost all examples showed no signifiant difference. The notable exception, *cook at home*, will be discussed in section 4.

4 Discussion

In their experimental results, Pruś et al. showed that where predicates are presented to participants devoid of context, there is room for disagreement on whether for a given predicate there exists an entailment from past progressive to past simple. They suggested a number of possible explanations for that disagreement. One explanation they proposed was that people tend to use a common-sense approach to reasoning. They speculated that people tended to answer 'Yes' for some of the telic examples where the theoretically correct answer is 'No', because they simply did not conceive of a scenario where such action gets abandoned. Our study addressed this suggestion by adding such a scenario as a context to the predicate. We observed that for some examples participants' answers significantly shifted towards 'No' after adding that context. The existence of examples where that shift occurred provides empirical backing to the above-mentioned suggestion of Pruś et al.. Having said that, the shift did not occur for all of the examples. We therefore note that ambiguity is harder to resolve in some cases than in others, and so the participants were still left to make a personal judgement. In this section we will discuss selected examples to suggest further insights on what factors may have influenced participants' answers.

4.1 Activities

Our prediction was that there should not be any shift in answers for Activities. By definition, they do not lead to a culmination point, therefore the entailment from past progressive to past simple should hold regardless of any interruptions. For example, let's say Person X was listening to the radio one afternoon, when a power outage caused their radio to turn off. The sentence *Person X listened to the radio* would still be true. In other words, one can engage in an activity, but one does not complete an activity.

In the light of above, it is noteworthy that we do observe a shift towards 'No' for the predicate *cook at home*. *Cook* is normally a transitive verb, but in this example it is being used intransitively as it lends itself to the unspecified object alteration (Levin, 1993). Unspecified object means that one would expect the existence of an underlying object, but it is not expressed in the sentence because it is not relevant for the context). For example, if one asks you 'Are you hungry?', you might answer 'No, I ate'. In this example *I ate* has an unspecified object because it doesn't matter what you ate, but implicitly you ate something inferrable, such as breakfast. Likewise, if you *cooked at home*, you must have cooked *something*. Importantly, as soon as we do introduce a specific object to this verb phrase, e.g. *cooking a meal at home*, its interpretation turns from cooking as an Activity to cooking as an Accomplishment. What we are likely observing with this result, is that even without introducing an object to the sentence, the fact that the action of cooking gets interrupted drives the readers attention to the fact that there was an underlying *something* being cooked, that will now not be completed.

4.2 Accomplishments

When it comes to Accomplishments and Achievements, we predicted that a shift towards 'No' should occur after adding an interruption to the predicate's context. However, amongst Accomplishments, we only observed that shift for roughly half of the examples. Looking at individual examples, we suggest possible explanations for why the shift may have not occurred.

Consider *write a novel*. For this predicate, the answer distribution was already skewed towards 'No' in the original study. Because the action incompletion being hard to conceive was not a problem to begin with, the additional context designed to resolve this problem yielded no change.

Having said that, there are examples where the distribution from the original study was not skewed towards 'No' and it did not shift towards 'No' in the interrupted setting. Take *to read a book*, for example. It highlights an interesting property of Accomplishments: as they are durative, some Accomplishments can be coerced into Activities. The predicate *to read a book* is a common example of

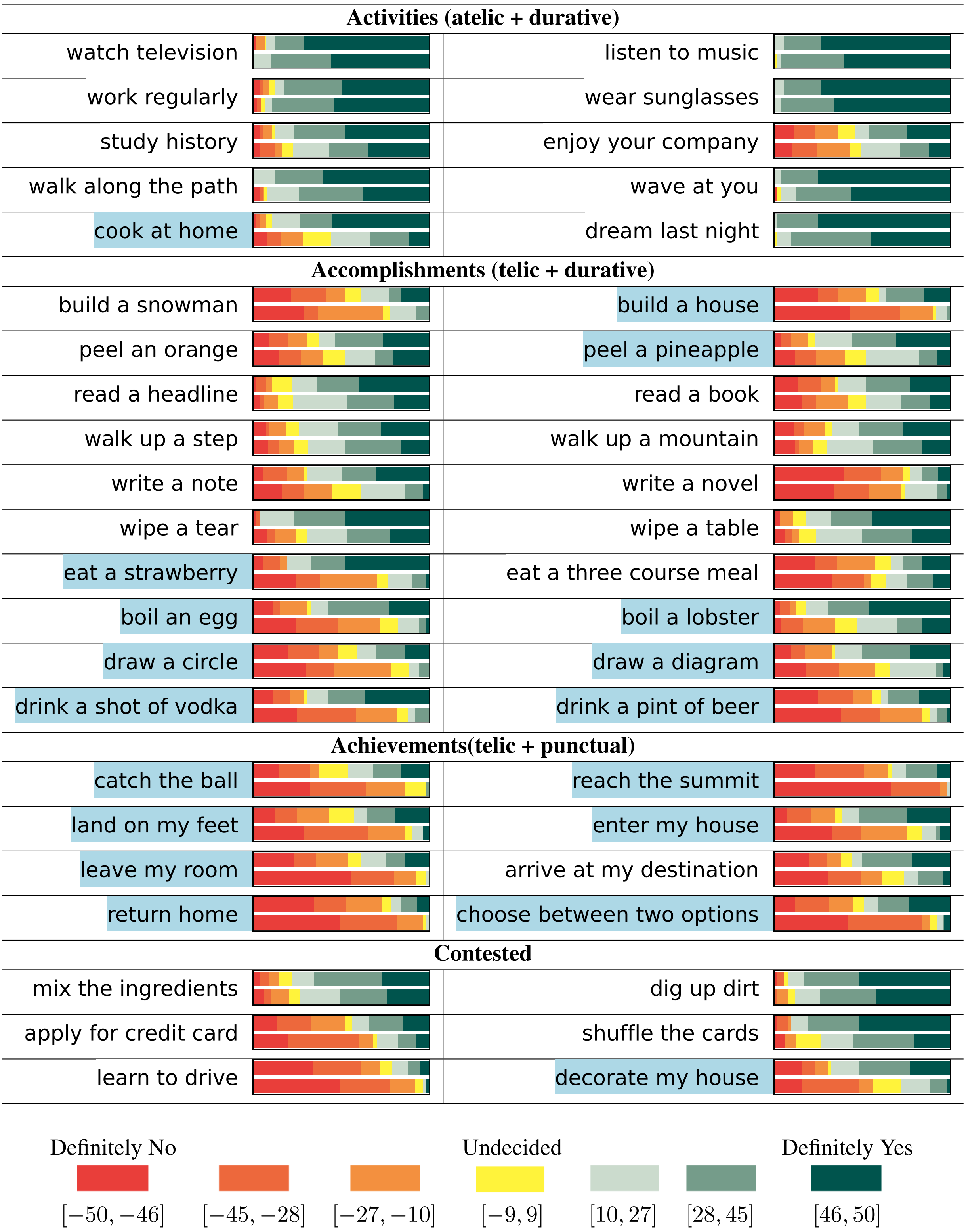


Figure 2: For each example, the lower bars present the distributions of participants' answers to our survey (interrupted setting) and the upper bar present the results as collected by Pruś et al. (2024) (decontextualised setting). The answers were encoded as integers from -50 (signifying a certain 'No') to 50 (signifying certain 'Yes'). To plot the answer distributions, we mapped the slider values into 7 intervals represented by different colours, as per the legend. Highlighted in blue are the examples for which there was a statistically significant difference between the two distributions.

an Accomplishment in literature. Yet, *I read a book for an afternoon* is a viable sentence which would be interpreted as an Activity. The way the literature suggests to treat such examples is to annotate *I read a book* as an Accomplishment and after adding the context, annotate *I read a book for an afternoon* as an Activity. Instead, we propose to look at the predicate *I read a book* as underspecified with respect to its predicational aspect. The full sentence we showed our participants was *I was reading a book but I lost all interest so I never finished the last chapter*, so it is clear that the endpoint was not reached. We observe that, much like the group who was shown *I was reading the book* without any context, their responses are fairly divided between 'Yes' and 'No'. Simply because this option to mould *I read a book* into an Activity exists, it will be seen as such by some. Even if the context along the lines of *for an afternoon* is not included in the sentence, the fact that it can be, means that the endpoint is optional.

Another example where there is a split between 'Yes' and 'No' answers and no significant difference between decontextualised and interrupted setting is *to walk up a hill*. Pruś et al. (2024) in their discussion of that example brought up that the endpoint (inherent to all telic predicates) might not viewed as necessary. In fact, our study's result showing no shift towards 'No' for the full sentence *I was walking up a hill but a sprained ankle prevented me from reaching the top* seems to confirm their analysis. They put forward that this difference in 'necessity of the endpoint' amongst Accomplishments is rooted in one's beliefs about the world. Namely, a casual hiker might find saying *I walked up a mountain but didn't reach the top* acceptable, whilst a committed mountaineer would disagree. Whilst there is definitely merit to that suggestion, we want to put forward another factor which might play a role in this example: the preposition phrase 'up a mountain' can be describing either a path to a goal or a location of a hike.

Finally, in some cases, despite the additional context, there may be some uncertainty left about the action's completion status. Consider the sentence *I was peeling an orange but I realised it was rotten so I threw it away*. Here, a possible interpretation is that one only threw the orange away once it was completely peeled. Having said that, further investigation would be required to determine to what degree this uncertainty is inherent of the predicate and to what degree it can be manipulated by different context.

4.3 Achievements

Because Achievements are punctual, unlike Accomplishments, they do not lend themselves to being coerced into Activities. Instead, putting an Achievement into a progressive form shifts the interpretation from looking at the punctual event itself to looking at its preparatory stages. The preparatory stages are still characterised as telic and the Achievement itself is a readily available endpoint. This provides a likely explanation for why all but one of the examples from the category observed a shift towards 'No'. The only exception here was *arrive at my destination*. The full sentence presented was *I was arriving at my destination but a truck pulled out in front of me and I stopped*. It is possible that, similarly to the *peel an orange* example above, the participants interpreted that the stopping occurred already at the destination. It is also worth noting that whilst this example was considered to not yield a significant shift under the p<0.01 threshold, it did meet the p<0.05 threshold.

5 GPT-4 Experiment

The result of our experiments with human participants shows that judgements on the answer to the imperfective paradox are divided even with the additional context. Those differences between judgements are at least in part driven by the differences in participants individual experiences in the real world. In this section we outline an experiment to test whether the distribution of event descriptions in the vast training dataset of a large language model (LLM) reflects the common-sense inferences of the subjects to such an extent that its responses follow a similar distribution to theirs. We chose GPT-4 (OpenAI et al., 2024) because it is one of the most widely used models at the time of writing.

In order to query the model one sends an array of messages through the API. The messages can be assigned different roles. We use the *developer* role to provide instruction on the answer format and the *user* role to ask the actual question. The complete prompt looks as follows:

Developer: *Answer the question with Yes or No.*
User: *If the sentence S1 is true, does it necessarily mean that the sentence S2 is also true?*

where S1 and S2 are the sentences derived from a given predicate. Mirroring our study with human

	No Context	Interrupted	Difference
Activities (atelic + durative)			
cook at home	**100.00**±0.00	**64.75**±2.63	**-35.25**±2.63
dream last night	**100.00**±0.00	**100.00**±0.00	**0.00**±0.00
enjoy your company	**33.50**±3.11	**98.00**±1.63	**+64.50**±3.70
listen to music	**99.75**±0.50	**100.00**±0.00	**+0.25**±0.50
study history	**25.00**±1.83	**99.75**±0.50	**+74.75**±1.71
walk along the path	**97.50**±2.38	**99.50**±0.58	**+2.00**±2.45
watch television	**99.50**±0.58	**99.75**±0.50	**+0.25**±0.50
wave at you	**99.75**±0.50	**99.50**±1.00	**-0.25**±1.26
wear sunglasses	**100.00**±0.00	**100.00**±0.00	**0.00**±0.00
work regularly	**96.75**±4.27	**99.25**±0.96	**-2.50**±3.79
Accomplishments (telic + durative)			
boil a lobster	**79.25**±5.68	**0.00**±0.00	**-79.25**±5.68
boil an egg	**25.00**±7.26	**0.00**±0.00	**-25.00**±7.26
build a house	**0.00**±0.00	**0.00**±0.00	**0.00**±0.00
build a snowman	**0.00**±0.00	**0.00**±0.00	**0.00**±0.00
draw a circle	**0.00**±0.00	**0.00**±0.00	**0.00**±0.00
draw a diagram	**27.00**±4.97	**0.00**±0.00	**-27.00**±4.97
drink a pint of beer	**93.50**±4.65	**0.00**±0.00	**-93.50**±4.65
drink a shot of vodka	**93.50**±3.32	**0.00**±0.00	**-93.50**±3.32
eat a three course meal	**4.75**±2.99	**0.00**±0.00	**-4.75**±2.99
eat a strawberry	**93.75**±5.74	**0.00**±0.00	**-93.75**±5.74
peel a pineapple	**86.00**±6.58	**0.50**±0.58	**-85.50**±6.95
peel an orange	**49.75**±6.13	**0.00**±0.00	**-49.75**±6.13
read a book	**0.00**±0.00	**84.75**±7.50	**+84.75**±7.50
read a headline	**51.50**±3.87	**0.50**±0.58	**-51.00**±4.32
walk up a mountain	**4.00**±4.83	**3.25**±2.22	**-0.75**±2.87
walk up a step	**74.00**±8.45	**4.25**±4.35	**-69.75**±12.50
wipe a table	**33.25**±8.66	**0.00**±0.00	**-33.25**±8.66
wipe a tear	**54.00**±1.41	**0.75**±0.96	**-53.25**±1.26
write a note	**0.00**±0.00	**0.00**±0.00	**0.00**±0.00
write a novel	**0.00**±0.00	**0.00**±0.00	**0.00**±0.00
Achievements (telic + punctual)			
arrive at my destination	**0.00**±0.00	**0.00**±0.00	**0.00**±0.00
catch the ball	**0.00**±0.00	**0.00**±0.00	**0.00**±0.00
choose between two options	**0.00**±0.00	**0.00**±0.00	**0.00**±0.00
land on my feet	**10.25**±6.65	**0.75**±1.50	**-9.50**±6.24
leave my room	**15.25**±10.97	**0.00**±0.00	**-15.25**±10.97
return home	**0.25**±0.50	**0.00**±0.00	**-0.25**±0.50
enter my house	**37.00**±1.83	**0.00**±0.00	**-37.00**±1.83
reach the summit	**0.00**±0.00	**0.00**±0.00	**0.00**±0.00
Contested			
apply for a credit card	**0.00**±0.00	**0.00**±0.00	**0.00**±0.00
decorate my house	**7.50**±4.73	**0.00**±0.00	**-7.50**±4.73
dig up dirt	**2.75**±2.22	**23.25**±6.18	**+20.50**±5.26
learn to drive	**0.00**±0.00	**0.00**±0.00	**0.00**±0.00
mix ingredients	**43.00**±5.16	**6.00**±4.55	**-37.00**±6.68
shuffle the cards	**3.75**±3.59	**3.50**±2.52	**-0.25**±3.59

Table 2: GPT-4 results. The model was prompted with each example 100 times. We summed up how many times it generated answer 'Yes' and confirmed that the remaining answers were 'No'. We averaged the sums over 3 different temperature setting. An average of 100 means that the model always generated 'Yes', whilst the average of 0 means that the model always answered 'No'. The final column shows the shift form the decontextualised to the interrupted setting.

participants, we considered the same two settings: decontextualised and interrupted. For example, for a predicate X, the decontextualised setting sentence S1 is the is of the *I was Xing* form; the interrupted setting sentence S1 is of the *I was Xing, but INT* (where INT is the added context) form. In either setting S2 is simply *I Xed*.

For each predicate in each setting we queried the model 100 times with the same prompt. This is to approximate the underlying probabilities the model has assigned to either of the answers ('Yes' and 'No'). The reason we used this approximation is because actual probabilities assigned to generated outputs are not available for GPT-4. We repeated this experiment for 3 different temperature values: 0.6, 0.8 and 1.0. For each predicate at each temperature setting we counted how many times the model answered 'Yes' and confirmed that the remaining answers were 'No' (i.e. there was no nonsense answers that didn't follow the developer instructions). We then calculated the arithmetic mean and standard deviation across the three temperature settings. Table 2 shows the full breakdown of GPT-4 responses per predicate.

Do the answers generated by GPT-4 then follow the similar patterns to people's answers? We observe that overall Activities collect more 'Yes' responses than Accomplishments and Achievements. This is in line with the theoretical predictions of the 'correct' answer as well as the overall picture of how people answered these questions. We also note that, similarly to human answers, *cook at home* was the only Activity where a significant shift towards 'No' was observed. Beyond that, however, our analysis revealed a number of differences between the human and the model responses.

We observe a significant shift towards 'Yes' for two examples of Activities: *enjoy your company* and *study history*. Neither of those examples elicited such a shift amongst people.

We observe a shift towards 'No' for Accomplishments and Achievements. We observed a shift in the direction amongst participants too. However, amongst humans, we observed that the shift was more pronounces for Achievements than Accomplishments and the reverse is true for the model. In general, amongst Accomplishments, human answers were often divided between 'Yes' and 'No' even in the interrupted setting. Consider *boil a lobster* as an example. Amongst people it started with a 'Yes'-skewed distribution in the decontextualised setting, whilst it did shift towards 'No', the human answered were still mixed. The model answers, however, usually followed one of two trajectories. They either started at unanimously 'No' in the decontextualised setting, in which case no shift occurred. Alternatively, they shifted to unanimously 'No' in the interrupted setting. The intriguing outlier was *read a book*, where a very significant shift towards 'Yes' occurred (from 0 to 84.75±7.5). The pattern of model answers for Achievements was similar to that of Accomplishments, with the difference that the model was less likely to assign any 'Yes' answers even in the decontextualised setting.

Overall, the model responses paint distinctions between aspectual classes similarly to how people do. However, looking at individual predicates, we see that additional context has a much more extreme effect on the model's answers than it does on the human answers. Recently, Li et al. (2024) have shown the embeddings of GPT-2 do not reflect the theoretical predictions stemming from the imperfective paradox and speculated that the context of their examples could explain that. Our results with GPT-4 are certainly in line with that observation.

6 Conclusions

Our results show that interruptions being less conceivable for some actions than others, was indeed a factor driving people to answer 'Yes' in the decontextualised setting. With some telic examples still showing a strong mix of responses even with the explicit interruption, we can say that it was likely not the only factor at play. We suggest that judgements on which actions are Activities and which are Accomplishments can depend upon individual differences in experience and understanding in the real world. Because a language model lacks such experience (Bender and Koller, 2020), the distribution of answers generated by GPT-4 differs from the distribution of human answers.

Limitations

This study analyses a set of 44 verb phrases only, which should be considered a small scale investigation. This means that whilst the study at hand points towards interesting patterns that might be influencing reasoning about actions, it might not be sufficient to make claims about predicate properties at large. Similarly, our LLM experiment involved only one model - GPT-4. Therefore, the results are not sufficient to make commentary on LLMs' capabilities overall. Moreover, the GPT-4 experiment would have been more robust if we could access the response probabilities directly instead of using prompting as a proxy (Hu and Levy, 2023). Having said that, the results still highlight interesting patterns in GPT's answers and, perhaps more importantly, interesting exceptions to these patterns. We, therefore, view our GPT-4 experiment as an invitation to further research on aspectual ambiguity and aspectual representations in LLMs.

Acknowledgements

We thank Iona Carslaw and Selma Tekir for feedback on an early draft of this paper. Katarzyna Pruś was supported by the EPSRC Centre for Doctoral Training in Data Science, funded by the UK Engineering and Physical Sciences Research Council (grant EP/L016427/1).

References

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

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.

Nathanael Chambers, Taylor Cassidy, Bill McDowell, and Steven Bethard. 2014. Dense event ordering with a multi-pass architecture. *Transactions of the Association for Computational Linguistics*, 2:273–284.

Bernard Comrie. 1976. *Aspect : an introduction to the study of verbal aspect and related problems*. Cambridge textbooks in linguistics. Cambridge University Press, Cambridge.

David R. Dowty. 1979. *Word meaning and Montague grammar: the semantics of verbs and times in generative semantics and in Montague's PTQ*. Number v. 7 in Synthese language library. D. Reidel Pub. Co, Dordrecht ; Boston.

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

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

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

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

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

Thomas Kober, Sander Bijl de Vroe, and Mark Steedman. 2019. Temporal and aspectual entailment. In *Proceedings of the 13th International Conference on Computational Semantics - Long Papers*, pages 103–119, Gothenburg, Sweden. Association for Computational Linguistics.

Alex Lascarides. 1991. The progressive and the imperfective paradox. *Synthese*, 87(6):401–447.

Beth Levin. 1993. *English verb classes and alternations : a preliminary investigation*. University of Chicago Press, Chicago.

Yuxi Li, Emmanuele Chersoni, and Yu-Yin Hsu. 2024. Investigating aspect features in contextualized embeddings with semantic scales and distributional similarity. In *Proceedings of the 13th Joint Conference on Lexical and Computational Semantics (*SEM 2024)*, pages 80–92, Mexico City, Mexico. Association for Computational Linguistics.

Marc Moens and Mark Steedman. 1988. Temporal ontology and temporal reference. *Computational Linguistics*, 14(2):15–28.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke

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

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694.

Rashmi Prasad, Bonnie Webber, and Aravind Joshi. 2017. The penn discourse treebank: An annotated corpus of discourse relations. In Nancy Ide and James Pustejovsky, editors, *Handbook of Linguistic Annotation*, pages 1197–1217. Springer Netherlands, Dordrecht.

Katarzyna Pruś, Mark Steedman, and Adam Lopez. 2024. Human temporal inferences go beyond aspectual class. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1913–1923, St. Julian's, Malta. Association for Computational Linguistics.

James Pustejovsky. 1991. The generative lexicon. *Computational Linguistics*, 17:409–441.

Susan Rothstein. 2008. *Structuring events: A study in the semantics of lexical aspect*, volume 5. John Wiley & Sons.

Eric V. Siegel and Kathleen R. McKeown. 2000. Learning Methods to Combine Linguistic Indicators: Improving Aspectual Classification and Revealing Linguistic Insights. *Computational Linguistics*, 26(4):595–628.

Zeno Vendler. 1967. *Linguistics in Philosophy*. Cornell University Press.

H. J. Verkuyl. 1972. *On the Compositional Nature of the Aspects*. Dordrecht, Netherlands: D.Reidel Publishing Company.

Sandro Zucchi. 2020. Progressive: The imperfective paradox. *The Wiley Blackwell Companion to Semantics*, pages 1–32.

A Participant Recruitment

It was key for the participants to be native English speakers. Therefore, the platform used for participant recruitment was Prolific.co, which allows to pre-screen participants on that criterion. All of the participants confirmed that English is the language they grew up with and use regularly in their life at present. All of the participants have passed the attention checks included in the survey. The median survey completion time was 8 minutes and the participants were compensated 2.50 GBP for taking part in the study. The survey was ran with approval from a relevant ethics committee. All of the participants have provided their consent for their anonymised responses to be made publicly available and used in academic publications.

B Participant Instructions

Below is the text provided as instructions to the participants. Please note that for half we used a variant of instructions , where the following example was used instead:

> If the sentence **"I was winning the race but then I stumbled and fell."** is true, does it necessarily mean that the sentence **"I won the race."** is also true?

INSTRUCTIONS:

Please read the instructions now - they will not be repeated on further pages and there will not be an option to come back to this page.

In this survey you will be presented with pairs of sentences. For each pair you will be asked to **assume that the first sentence is true**. Using your best judgement, we ask you to **indicate whether the second sentence is therefore also true**. Use the slider to indicate the confidence in your judgement - **the further away from the middle you place the slider, the more confident you are in your judgement**.

PRACTICE EXAMPLE:

Please answer this question:

If the sentence **"I was playing at the park but you said it's time to go home."** is true, does it necessarily mean that the sentence **"I played at the park."** is also true?

(Here is a slider as illustrated in Figure 1)

BEAR IN MIND:

If either sentence is not interpretable or either sentence is grammatically incorrect - tick the "Does Not Make Sense" box.

If both sentences are sensible and correct, please provide an answer with the slider. Please note, that even if you are confident that you want to leave the slider in the middle - you will have to move it slightly and ultimately put it back in the middle, before you'll be able to press "Next".

There will be **three attention checking questions** in this survey - they will vary in structure from the description above.

C GPT Prompts

The following template was used to build the prompts for the GPT experiment.

> ***Developer:*** *Answer the question with Yes or No.*
> ***User:*** *If the sentence S1 is true, does it necessarily mean that the sentence S2 is also true?*

S1 is a sentence in past progressive. S2 is a sentence in past simple.

D Added Context

Tables 3 and 4 show all of the past progressive forms of the example verb phrases together with the additional context that has been crafted for the interrupted setting of the experiment.

Past Progressive	**Added Context**
Activities (atelic + durative)	
I was cooking at home	but the stove stopped working.
I was dreaming last night	but a sudden noise woke me up.
I was enjoying your company	but you said something that really hurt my feelings.
I was listening to music	but the bell rang and I got up to answer the door.
I was studying history	but I found it boring and regretfully dropped out.
I was walking along the path	but I spotted a bear and so I turned around.
I was watching television	but suddenly the power went out.
I was waving at you	but my arm went numb and you still didn't notice me.
I was wearing sunglasses	but I took them off when I walked into the building.
I was working regularly	but an accident disrupted my routine.
Accomplishments (telic + durative)	
I was boiling a lobster	but I knocked over the pot, spilling the water and extinguishing the fire.
I was boiling an egg	but my camping stove ran out of fuel.
I was building a house	but I ran out of money to finish the construction.
I was building a snowman	but a sudden storm broke out and I ran home.
I was drawing a circle	but my pencil broke and I went off in search of a sharpener.
I was drawing a diagram	but my pen ran out of ink.
I was drinking a pint of beer	but you knocked the glass out of my hand.
I was drinking a shot of vodka	but my hands shook so I spilled it all over my jumper.
I was eating a three course meal	but I couldn't finish the last course.
I was eating a strawberry	but a bird flew down and stole it out of my hand.
I was peeling a pineapple	but I injured myself with the knife and had to go to a hospital.
I was peeling an orange	but I realised it was rotten and so I threw it away.
I was reading a book	but I lost all interest and so never finished the last chapter.
I was reading a headline	but I heard a worrying noise, so I went downstairs to investigate.
I was walking up a mountain	but a sprained ankle prevented me from reaching the top.
I was walking up a step	but I tripped up and fell down.
I was wiping a table	but you walked in and distracted me before I finished.
I was wiping a tear	but you grabbed my hand and pulled me to the side.
I was writing a note	but you walked in, so I could relay the message in person instead.
I was writing a novel	but I lacked ideas for an ending.

Table 3: The Past Progressive forms of all the predicates considered, together with the context added in this study. Continued in table 4.

Past Progressive	**Added Context**
Achievements (telic + punctual)	
I was arriving at my destination	but a truck pulled out in front of us and we stopped.
I was catching the ball	but suddenly my hand cramped.
I was choosing between two options	but it took me too long so they asked someone else to choose for me.
I was landing on my feet	but a gust of wind swiped me to the side.
I was leaving my room	but I turned around and walked back to my desk.
I was returning home	but you called asking me to come over to your place and I couldn't refuse.
I was entering my house	but suddenly the wind blew the door shut.
I was reaching the summit	but the weather conditions suddenly worsened forcing me to turn around.
Contested	
I was applying for a credit card	but I gave up when I found out my credit score was too low.
I was decorating my house	but I fell off the ladder and broke my arm.
I was digging up dirt	but you told me not to.
I was learning to drive	but my instructor quit and I gave up on ever passing the exam.
I was mixing ingredients	but I dropped the bowl on the floor and it shattered to pieces.
I was shuffling the cards	but a fight broke out and I decided to leave the casino.

Table 4: Continuation of table 3. The Past Progressive forms of all the predicates considered, together with the context added to them in this study.