TIMERES: A Turkish Benchmark For Evaluating Temporal Understanding of Large Language Models

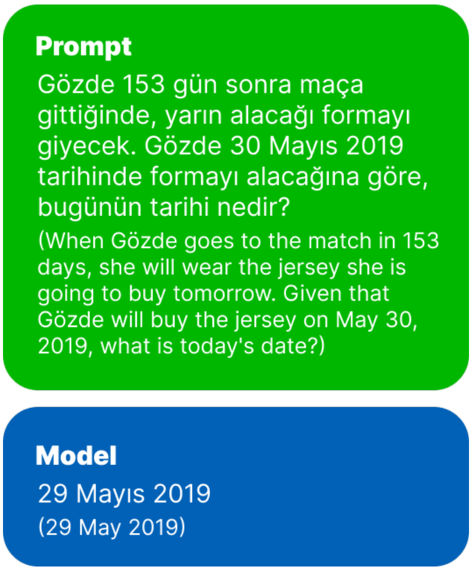
Habib Yağız Demir, Ümit Atlamaz, Susan Üsküdarlı
Bogazici University
Istanbul, Turkey
habib.demir@std.bogazici.edu.tr
umit.atlamaz@bogazici.edu.tr
suzan.uskudarli@bogazici.edu.tr

Abstract

Temporal information is an essential part of communication, and understanding language requires processing it effectively. Despite recent advances, Large Language Models (LLMs) still struggle with temporal understanding. Existing benchmarks primarily focus on English and underexplore how linguistic structure contributes to temporal meaning. As a result, temporal understanding in languages other than English remains largely understudied. In this paper, we introduce TIMERES, a Turkish benchmark for evaluating temporal understanding of LLMs. TIMERES aims to investigate comprehension of Reichenbach's temporal points and reported speech through date arithmetic. Our dataset includes 4,600 questions across 4 tasks at two levels of complexity, and presents a paired question formulation to distinguish temporal discourse understanding from temporal arithmetic capabilities. We evaluated six LLMs, and demonstrated that models struggle to resolve reported speech and fail to generalize across word order variations. Code and data are available at https://github.com/yagizdemir/timeres

1 Introduction

Natural language understanding requires anchoring events in time and establishing relationships between them through markers such as tense, aspect, and adverbs. By enabling us to build a timeline of events, temporal reasoning is a fundamental aspect of human communication. Therefore, Large Language Models (LLMs) must demonstrate robust understanding of time to ensure reliable performance in tasks such as planning, scheduling, and temporal discourse comprehension. In spite of the significant advancements in the reasoning abilities of LLMs (Xu et al., 2025; Zhang et al., 2025; Comanici et al., 2025; Jaech et al., 2024), recent benchmarks reveal that they still lack such understanding (Wang and Zhao, 2024; Zhou et al., 2021; Wei et al., 2023).



Prompt
Gözde 153 gün sonra maça gittiğinde, yarın alacağı formayı giyecek. Gözde 30 Mayıs 2019 tarihinde formayı alacağına göre, bugünün tarihi nedir?
(When Gözde goes to the match in 153 days, she will wear the jersey she is going to buy tomorrow. Given that Gözde will buy the jersey on May 30, 2019, what is today's date?)

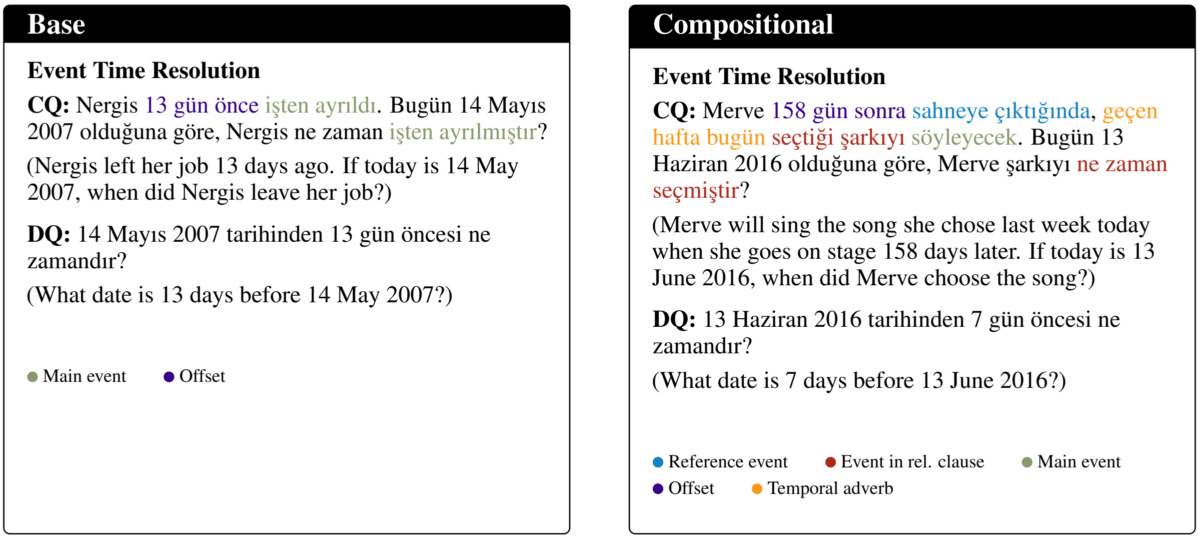
Model
29 Mayıs 2019
(29 May 2019)

Figure 1: An example prompt from TIMERES for the Speech Time Resolution task in the Compositional set.

In recent years, various benchmarks have been introduced to investigate temporal reasoning capabilities of LLMs, covering tasks spanning from event ordering to temporal arithmetic (Zhou et al., 2019; Qin et al., 2021; Ning et al., 2020; Fatemi et al., 2024). These benchmarks demonstrate that the models struggle with many aspects of temporal reasoning such as event duration, frequency and date arithmetic. However, existing datasets primarily focus on English. Consequently, LLMs' performance in underrepresented languages like Turkish remains largely understudied. Furthermore, they do not address the linguistic facet of temporal reasoning, mostly concentrating on task-level performance.

To address these limitations in existing research, we present TIMERES, a Turkish benchmark, grounded in Reichenbach's tense framework (Reichenbach, 1947). This framework defines three temporal points to locate events from the speaker's perspective: **Event Time** (E) is the time at which

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 4: Student Research Workshop, pages 910–920
March 26, 2026 ©2026 Association for Computational Linguistics



Base

Event Time Resolution

CQ: Nergis 13 gün önce işten ayrıldı. Bugün 14 Mayıs 2007 olduğuna göre, Nergis ne zaman işten ayrılmıştır?

(Nergis left her job 13 days ago. If today is 14 May 2007, when did Nergis leave her job?)

DQ: 14 Mayıs 2007 tarihinden 13 gün öncesi ne zamandır?

(What date is 13 days before 14 May 2007?)

● Main event ● Offset

Compositional

Event Time Resolution

CQ: Merve 158 gün sonra sahneye çıktığında, geçen hafta bugün seçtiği şarkıyı söyleyecek. Bugün 13 Haziran 2016 olduğuna göre, Merve şarkıyı ne zaman seçmiştir?

(Merve will sing the song she chose last week today when she goes on stage 158 days later. If today is 13 June 2016, when did Merve choose the song?)

DQ: 13 Haziran 2016 tarihinden 7 gün öncesi ne zamandır?

(What date is 7 days before 13 June 2016?)

● Reference event ● Event in rel. clause ● Main event
● Offset ● Temporal adverb

Figure 2: Examples questions from TIMERES. **CQ**: Contextual Question, **DQ**: Direct Question.

the event occurs, **Speech Time** (S) represents the point the utterance is produced, and **Reference Time** (R) is the vantage point from which the event is perceived. We designed the tasks in TIMERES around these temporal points to focus on the linguistic foundations of temporal reasoning. To the best of our knowledge, TIMERES is the first benchmark specifically designed for temporal reasoning in Turkish and focused on such linguistic components.

We evaluated six models from Meta's Llama, Google's Gemini and OpenAI's GPT families: Llama 3.1 8B (Dubey et al., 2024), Llama 3.3 70B, Gemini 2.0 Flash (Google, 2024), Gemini 2.5 Flash (Comanici et al., 2025), ChatGPT 4.1 (OpenAI, 2024), and ChatGPT 5.1 (OpenAI, 2025). Our experiments indicate that Gemini and GPT models effectively handle temporal context in simple sentences while they struggle with compositional sentences containing embedded clauses, and exhibit a lack of generalization across different word orders. In summary, our contributions are as follows:

- We introduce TIMERES, the first benchmark dataset dedicated to the evaluation of temporal reasoning in Turkish, with the aim of addressing the lack of resources in low-resource languages.
- We ground our benchmark in a linguistically established framework, Reichenbach's tense framework, by developing tasks for resolution of E, S, R, and reported speech.
- We systematically investigate the impact of different word orders and sentence structures on the temporal reasoning capabilities of state-of-the-art LLMs.
- We propose a paired question formulation that allows us to isolate temporal understanding from arithmetic capabilities.
- We identify a common failure in reported speech across all models, and demonstrate that LLMs lack robustness to word order variations in speech time and reference time.

2 Related Work

2.1 Temporal Reasoning Benchmarks

Various benchmarks have been developed to investigate distinct aspects of temporal reasoning. A significant part of these benchmarks focuses on three core capabilities: temporal commonsense, time-scoped question answering (QA), and symbolic reasoning (Virgo et al., 2022; Chu et al., 2024; Dhingra et al., 2022; Jain et al., 2023). Temporal commonsense datasets mostly evaluate LLMs via event-based tasks such as MCTACO (Zhou et al., 2019), TRACIE (Zhou et al., 2021), TIMEDIAL (Qin et al., 2021), and CaTeRs (Mostafazadeh et al., 2016). In time-scoped QA, LLMs are expected to answer questions grounded in facts that change over time (Hu et al., 2024; Wallat et al., 2024; Chen et al., 2024).

Symbolic temporal reasoning benchmarks investigate models' abilities to perform explicit logical operations, such as temporal arithmetic (Srivastava et al., 2023; Tan et al., 2023). For example,

		Base				Compositional			
Model	**Setting**	**ETR**	**RTR**	**RSR**	**STR**	**ETR**	**RTR**	**RSR**	**STR**
Llama 3.1 8B	CQ	23.5/67.6	6.0/91.7	1.0/120.8	14.0/60.6	0.2/64.2	1.8/17.1	1.2/28.9	2.0/48.7
	DQ	29.0/27.6	23.0/22.0	0.0/86.7	19.5/22.1	98.0/0.0	1.6/53.5	2.4/31.7	96.4/0.0
Llama 3.3 70B	CQ	60.0/4.2	45.5/30.3	0.0/87.0	21.5/92.5	50.4/25.6	3.4/4.4	2.0/44.2	30.6/34.0
	DQ	60.0/7.0	48.5/10.3	0.5/63.3	48.5/9.9	99.6/0.0	10.4/6.4	8.0/7.2	**100.0**/0.0
Gemini 2.0 Flash	CQ	91.5/0.1	84.0/10.9	0.5/69.1	71.5/30.2	78.6/12.3	4.8/3.2	3.0/97.6	79.2/10.1
	DQ	90.5/0.1	89.5/0.3	6.0/47.2	88.5/0.1	**100.0**/0.0	31.6/3.2	34.8/4.1	**100.0**/0.0
Gemini 2.5 Flash	CQ	93.0/0.2	87.5/7.7	6.5/61.9	88.0/0.4	94.8/3.0	5.0/3.7	30.8/50.8	68.2/15.4
	DQ	94.5/0.1	93.5/0.1	9.5/27.9	90.5/0.1	**100.0**/0.0	53.6/1.7	46.8/1.8	**100.0**/0.0
ChatGPT 4.1	CQ	**95.5**/0.0	92.5/1.9	1.0/86.8	56.5/81.0	81.0/10.1	2.2/4.3	0.4/110.1	51.6/15.6
	DQ	96.5/0.0	94.5/0.2	10.0/40.8	96.5/0.0	99.6/0.0	45.2/4.3	54.8/2.1	**100.0**/0.0
ChatGPT 5.1	CQ	**95.5**/0.0	**95.0**/0.1	**97.0**/0.1	**93.5**/1.8	**99.8**/0.4	**60.0**/2.7	**95.0**/2.6	**95.4**/4.2
	DQ	**97.5**/0.0	**97.5**/0.0	**93.0**/0.4	**97.0**/0.0	**100.0**/0.0	**92.0**/0.2	**93.6**/0.1	**100.0**/0.0

Table 1: Results of model performances on the TimeRes benchmark. Scores are reported as Exact Match/Mean Absolute Error. **ETR**: Event Time Resolution, **RTR**: Reference Time Resolution, **RSR**: Reported Speech Resolution, **STR**: Speech Time Resolution, **CQ**: Contextual Question, **DQ**: Direct Question.

ChronoSense (Islakoglu and Kalo, 2025) defines 16 tasks based on Allen's interval relations and temporal arithmetic to evaluate LLMs' temporal understanding. Similarly, Test-of-Time (Fatemi et al., 2024) targets the arithmetic capabilities by employing tasks ranging from date comparison to duration calculations across time zones. In addition to these, general-purpose benchmarks like DROP (Dua et al., 2019) and BIG-Bench (Srivastava et al., 2023) also include tasks requiring date calculations. However, existing benchmarks neglect the role of linguistic devices such as tense and adverbs, which create a timeline by anchoring events to temporal points. Also, they are limited to English, leaving temporal reasoning in other languages understudied.

2.2 Turkish Reasoning Benchmarks

In recent years, the number of benchmarks in Turkish has increased, though they remain limited. Two distinct Turkish adaptations of MMLU (Hendrycks et al., 2020) exist: Yüksel et al. (2024) introduced TurkishMMLU spanning 9 subjects with over 10,000 questions from online learning materials, while Bayram et al. (2025) presented TR-MMLU, which covers 6,800 questions across 62 subjects. Cetvel (Er et al., 2025) brings together existing benchmarks and covers 23 tasks ranging from QA to Natural Language Inference. Additionally, multilingual datasets such as XCOPA (Ponti et al., 2020), XGLUE (Liang et al., 2020) and XNLI (Conneau et al., 2018) include Turkish. Although the benchmarks in Turkish are not limited to these, a dedicated temporal reasoning dataset does not exist.

We introduce TimeRes to address the lack of temporal reasoning benchmarks in Turkish, evaluating LLMs' comprehension of event time, speech time, reference time, and reported speech through arithmetic calculations in Turkish.

3 The TimeRes Dataset

TimeRes is a Turkish benchmark for evaluating LLMs on comprehension of E, R, S, and reported speech. It involves 4,600 questions across 4 tasks at two levels of complexity, presented as closed-ended questions. The dataset will be made publicly available after the review process. Example questions are shown in Figure 2. For examples of questions across all tasks and complexity levels, see Appendix A.

3.1 Dataset Settings and Task Definitions

Setup. Tasks in TimeRes require models to identify temporal points in a sentence and place them along a timeline using the offsets (i.e. the temporal distance from a reference point, such as *10 gün sonra* (10 days later)) and temporal adverbs. Given a base date that anchors one of the temporal points, models are asked to compute the date of the target temporal point.

Complexity Levels. The tasks are presented at

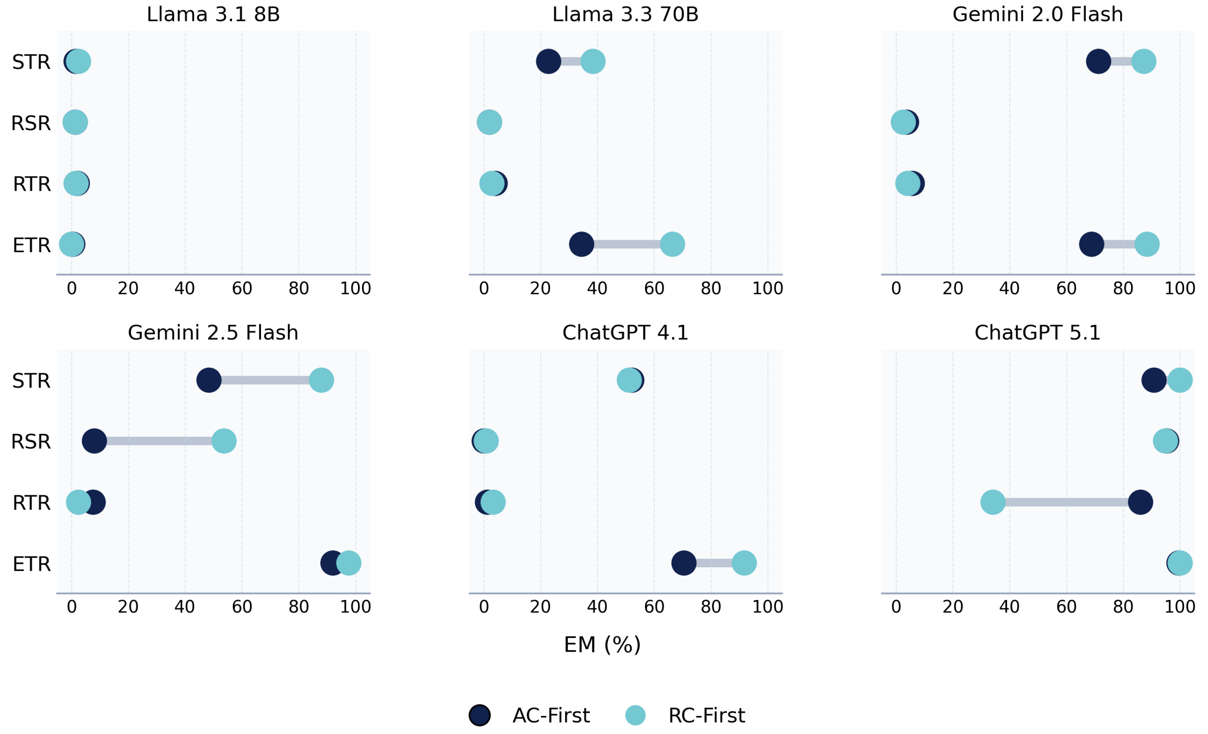


Figure 3: Performance gaps across different word orders. **ETR**: Event Time Resolution, **RTR**: Reference Time Resolution, **RSR**: Reported Speech Resolution, **STR**: Speech Time Resolution, **EM**: Exact Match

two complexity levels, which are Base and Compositional. Questions in the Base set present a simple linear timeline, whereas Compositional questions introduce a nested timeline through a temporal adverbial clause (AC) and a relative clause (RC), requiring models to resolve multiple temporal points.

Paired Questions. We use a paired question formulation to separate temporal discourse understanding from temporal arithmetic. For each question, we generated a counterpart that asks the underlying calculation explicitly without any context. We refer to the main question as Contextual Question (CQ) and the counterpart as Direct Question (DQ). The goal of this formulation is to test whether a model's performance stems from arithmetic abilities or temporal discourse understanding. For the remainder of this paper, 'questions' will denote CQs unless stated otherwise.

Linguistic Variations. The questions at each level involve different types of linguistic variation:

- **Base level:** The direction of the offsets is determined by the temporal adverbs *önce* 'before' and *sonra* 'after'. We construct an equal number of questions for each adverb to examine whether model performance varies based on the direction of the temporal offset.
- **Compositional level:** Questions have two variants with different word orders: either the RC follows the AC, or vice versa. With this, we aim to test the models' ability to generalize across word orders. Questions were distributed equally across word orders. In addition, we integrated a deictic temporal adverb (e.g., *geçen hafta bugün* (this day last week)) into the RC of questions to anchor events to S, thereby preventing potential ambiguity. These temporal adverbs are presented in Table 4

Task Definitions. We design four tasks to evaluate the models on understanding of Reichenbach's temporal points (Reichenbach, 1947) and their resolution in reported speech:

- **Event Time Resolution (ETR)** anchors S and requires models to predict the date of E by anchoring S.
- **Speech Time Resolution (STR)** flips the logic of ETR by targeting the date of the S.
- **Reference Time Resolution (RTR)** focuses on computing the date of R. Unlike the other tasks, RTR includes a reference event in Base questions. At the Compositional level, it requires an additional computation step.

Task	Set	Anchor	Target
ETR	Base	S	E
	Comp.	S	E_{RC}
STR	Base	E	S
	Comp.	E_{RC}	S
RTR	Base	E	R
	Comp.	E_{RC}	R_{AC}
RSR	Base	E in U	S
	Comp.	E_{RC} in U	S

Table 2: Anchored and target temporal points by task and complexity levels. **ETR**: Event Time Resolution, **RTR**: Reference Time Resolution, **RSR**: Reported Speech Resolution, **STR**: Speech Time Resolution, **Comp**: Compositional, $\boldsymbol{E}$: Event Time, $\boldsymbol{R}$: Reference Time, $\boldsymbol{S}$: Speech Time, **U**: Utterance, $\boldsymbol{E_{RC}}$: Event Time of the Relative Clause, $\boldsymbol{E_{AC}}$: Reference Time of the Adverbial Clause.

- **Reported Speech Resolution (RSR)** targets the date of S. RSR adds a layer of complexity by shifting the temporal center through a quoted statement. Similar to RTR, RSR requires two-step calculation at both levels.

Each task at both levels anchors and targets different temporal points. These anchors and targets are given in Table 2.

3.2 Question Generation

TIMERES employs template-based question generation. For each task and complexity level, we manually developed question templates and event pools for population. The Base set comprises 69 single events for ETR, STR, and RSR, and 79 event pairs for RTR. For Compositional questions, we curated a pool of 38 event triplets. For each question, events were randomly selected from the corresponding pool and combined with randomly generated offsets and dates. Offsets were sampled between 7 and 180 days, while the base date range is between 1950-2025. DQs were populated using the same offsets and base dates as their pair CQs.

3.3 Dataset Statistics

Our dataset consists of 4,600 questions spanning four tasks at two levels of complexity. For the Base set, we generate 100 questions per temporal adverb (*önce* (before) and *sonra* (after)), resulting in 200 questions per task. For the Compositional set, we generate 250 questions for each word order, which sums up to 500 questions per task. Since each question is paired with a DQ, the Base set contains 800 DQs. In the Compositional set, however, different word orders share the same underlying date arithmetic; therefore, we generate a single DQ per word order pair, resulting in 1,000 DQs. Dataset statistics for the TIMERES are shown in Table 3

Statistic	Base	Comp.
# of tasks	4	4
# of CQs	800	2,000
# of DQs	800	1,000
Avg. CQ length (words)	20	24
Avg. DQ length (words)	9	10.5

Table 3: Dataset statistics for the TIMERES benchmark. **Comp**: Compositional, **CQ**: Contextual Question, **DQ**: Direct Question.

4 Experiments

We evaluated the following models on TIMERES:

- Llama 3.1 8B (`Llama-3.1-8B-Instruct`)
- Llama 3.3 70B (`Llama-3.3-70B-Instruct`)
- Gemini 2.0 Flash (`gemini-2.0-flash`)
- Gemini 2.5 Flash (`gemini-2.5-flash`)
- ChatGPT 4.1 (`gpt-4.1-2025-04-14`)
- ChatGPT 5.1 (`gpt-5.1-2025-11-13`)

We accessed and prompted the Llama models via Hugging Face's Inference Providers API, and the Gemini and GPT models via their official APIs.

4.1 Implementation Details

We evaluated all models in a few-shot setting using two examples per task (one per variation). A single example was provided for the DQs at the compositional level because they lack variations. We included a system instruction to enforce the output format, requiring dates to be generated as 'DD Month YYYY' in Turkish. The resulting outputs were parsed programmatically using Python's `datetime` library.

The `temperature` parameter was set to 0 to obtain maximally deterministic responses from the models. We evaluated ChatGPT 5.1 with its reasoning effort set to medium. When reasoning is

Adverb	Offset (days)
geçen hafta bugün (this day last week)	-7
dünden önceki gün (the day before yesterday)	-2
dün (yesterday)	-1
yarın (tomorrow)	+1
öbür gün (the day after tomorrow)	+2
haftaya bugün (this day next week)	+7

Table 4: Temporal adverbs used in questions at the compositional level, with their corresponding offsets in days.

enabled, ChatGPT 5.1 does not permit control of the `temperature` parameter, and we therefore left it at its default value.

4.2 Evaluation Metrics

We measured model performance using the Exact Match (EM) and Mean Absolute Error (MAE) metrics. EM measures the percentage of exactly correct responses, and MAE measures the average absolute error relative to the ground truth.

5 Results

Overall. The evaluation results are presented in Table 1. Our results show that models perform better on Base questions, while their performance drops substantially across nearly all tasks at the Compositional level. The only exception is ChatGPT 5.1, which achieves high EM and low MAE scores at both levels, except for RTR at the Compositional level. On the other hand, Llama 3.1 8B fails on nearly all tasks, with the exception of its >90% scores on ETR and STR DQs at the compositional level. Furthermore, we observe that model performance (except for ChatGPT 5.1) drops substantially from DQs to CQs at the compositional level. For example, ChatGPT 4.1's STR performance drops from 100% to 51.6%, while Llama 3.3 70B's ETR performance falls from 99.6% to 50.4%. This performance gap between CQs and DQs across all tasks at the Compositional level indicates that this drop stems from limited ability to interpret context in Compositional questions. These findings suggest that models struggle to resolve temporal points when embedded clauses are present.

Challenges. We observe that models struggle most with RSR at both levels, and with RTR at the Compositional level. Almost all models perform under 10% on compositional RTR CQs; even ChatGPT 5.1 drops from above 90% to 60% on the RTR task. At the Base level, models perform similarly poorly on both CQs and DQs in RSR, within the range of 0-10% for all except ChatGPT 5.1. This indicates an arithmetic limitation, as these questions require a two-step calculation involving two offsets and therefore double operation range. However, the second offset involved in the Compositional questions comes from temporal adverbs, whose maximum range is seven days. Consequently, DQ performance on RTR and RSR at the Compositional level is higher, while CQ performance lags substantially behind, with near-zero Exact Match scores, excluding Gemini 2.5 Flash on RSR and ChatGPT 5.1. This demonstrates that models struggle with multi-step reasoning and with handling a shifted temporal center in RSR. Furthermore, the substantial performance gap between CQs and DQs at the compositional level highlights the models' inability to ground 'today' within these contexts.

Model Comparison. ChatGPT 5.1 outperforms other models across nearly all tasks, while ChatGPT 4.1 achieves substantially lower performance than ChatGPT 5.1, with the smallest gap on Base ETR and RTR tasks. These results demonstrate a significant advancement in capabilities between the GPT 4 and GPT 5 generations. We observe a similar performance increase from Llama 3.1 8B to Llama 3.3 70B, likely due to increased model size; however, the 70B model still struggles with most tasks. However, we do not observe the same level of advancement from Gemini 2.0 Flash to Gemini 2.5 Flash. Interestingly, Gemini 2.0 Flash achieves higher EM and lower MAE scores on the compositional STR questions, respectively. Among the evaluated models, the open-source Llama models trail significantly behind the Gemini and GPT families, with the GPT family consistently outperforming its counterparts.

Word Orders. We analyzed models' sensitivity to word order in Compositional questions. Performance differences across word orders are illustrated in Figure 3. Our results show that model performance varies depending on word order. For the STR task, all models except ChatGPT 4.1 and Llama 3.1 8B perform better when the RC precedes the AC. For instance, ChatGPT 5.1 achieves

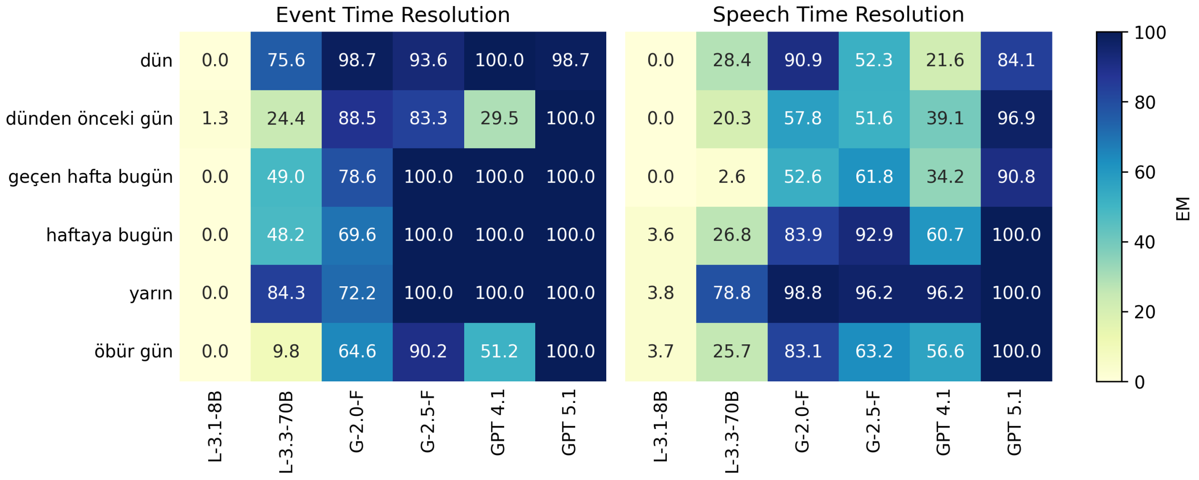


Figure 4: Performance breakdown by temporal adverbs across the Event Time Resolution and Speech Time Resolution tasks. **L-3.1-8B**: Llama 3.1 8B, **L-3.3-70B**: Llama 3.3 70B, **G-2.0-F**: Gemini 2.0 Flash, **G-2.5-F**: Gemini 2.5 Flash, **GPT 4.1**: ChatGPT 4.1, **GPT 5.1**: ChatGPT 5.1, **EM**: Exact Match

85% accuracy on AC-first sentences, compared to 100% on RC-first sentences. Gemini 2.0 Flash exhibits a similar trend, while the performance gap for Gemini 2.5 Flash exceeds 40%. This pattern is likely due to the STR task anchoring E_{RC}, which makes the target date easier to identify when the RC appears first, given the left-to-right dependency structure of autoregressive models. A similar word-order effect is observed for ETR and RSR (only for Gemini 2.5 Flash), both of which anchor E_{RC}. In particular, Llama 3.3 70B's performance on ETR illustrates a significant gap. Conversely, RTR shows better performance for AC-first questions, consistent with the fact that RTR anchors R_{AC}. However, this effect is observed only for ChatGPT 5.1, as other models achieve near-zero EM scores on RTR questions. Overall, these results indicate that current models struggle to generalize temporal reasoning across different word orders.

Temporal Adverbs. We analyze model performance across temporal adverbs, with the breakdown shown in Figure 4. In STR questions, models perform particularly poorly on adverbs indicating the past; for example, ChatGPT 4.1 drops to 20% Exact Match on dün (yesterday). Upon manual inspection, we observe that models tend to interpret these adverbs relative to the context, leading to errors, even though they are strictly anchored to S. We observe a similar trend for ETR questions, where ChatGPT 4.1 drops to 29% on *dünden önceki gün* (the day before yesterday). We also find that models perform poorly on öbür gün (the day after tomorrow). This kind of error likely stems from its relative interpretation as "the day after x", however, our questions do not logically permit this reading.

Temporal Anchoring Errors. Our manual inspection revealed that, at the compositional level of the RTR task, models frequently anchor R and E_{RC} to each other based on word order, whereas these temporal points should be anchored to S. To investigate this pattern, we calculated the hypothetical target dates resulting from this false anchoring. We found that a substantial proportion of incorrect responses stem from this specific error. For instance, this anchoring failure accounts for 87% of the errors made by Gemini 2.0 Flash. Similarly, around 60% of errors on compositional RTR questions by Gemini 2.5 Flash, Llama 3.3 70B, and ChatGPT 4.1 are due to this same pattern. However, we noted that the RC-first variation of the RTR task permits such an interpretation. Crucially, this exception is valid only for this specific format and does not extend to other tasks. We observed a similar error pattern in the compositional questions of the RSR task, where it accounts for 10–15% of errors made by the Gemini family. Although the RC-first variation of the RTR task permits such a reading, the persistence of this pattern in the RSR task and the AC-first variation of the RTR task demonstrates that these models struggle to resolve Reichenbach's temporal points.

Date Arithmetic. Gemini and GPT Models mostly achieve 90% or higher EM scores and low MAE scores on Base DQs except the RSR task. However, model performance declines sharply on DQs in RSR and RTR at the Compositional level, which require two-step calculations. This suggests that while the models can handle single-step calculations, their performance degrades substantially as the number of steps increases.

6 Conclusion

In conclusion, we introduce TIMERES, a Turkish benchmark for evaluating the temporal understanding of large language models. We define four tasks designed to assess Reichenbach's temporal points as well as reported speech. Our paired question formulation allows us to distinguish temporal discourse understanding from temporal arithmetic. We evaluate six LLMs and analyze their strengths and weaknesses in temporal reasoning in Turkish. Our results show that these models struggle with resolving reported speech, exhibit sensitivity to word order, and have difficulty interpreting sentences with embedded clauses.

Limitations

TIMERES is a synthetic dataset constructed programmatically using a template-based approach. This design results in repetitive syntactic structures and limits coverage of the variability found in naturally occurring language. Incorporating more diverse syntactic constructions could enable a more fine-grained analysis of the strengths and weaknesses of LLMs.

References

M Ali Bayram, Ali Arda Fincan, Ahmet Semih Gümüş, Banu Diri, Savaş Yıldırım, and Öner Aytaş. 2025. TR-MMLU benchmark for large language models: Performance evaluation, challenges, and opportunities for improvement. In *2025 33rd Signal Processing and Communications Applications Conference (SIU)*, pages 1–4. IEEE.

Ziyang Chen, Dongfang Li, Xiang Zhao, Baotian Hu, and Min Zhang. 2024. Temporal knowledge question answering via abstract reasoning induction. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4872–4889.

Zheng Chu, Jingchang Chen, Qianglong Chen, Weijiang Yu, Haotian Wang, Ming Liu, and Bing Qin. 2024. Timebench: A comprehensive evaluation of temporal reasoning abilities in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1204–1228.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 2475–2485.

Bhuwan Dhingra, Jeremy R Cole, Julian Martin Eisenschlos, Daniel Gillick, Jacob Eisenstein, and William W Cohen. 2022. Time-aware language models as temporal knowledge bases. *Transactions of the Association for Computational Linguistics*, 10:257–273.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. Drop: A reading comprehension benchmark requiring discrete reasoning over paragraphs. *arXiv preprint arXiv:1903.00161*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

Yakup Abrek Er, Ilker Kesen, Gözde Gül Şahin, and Aykut Erdem. 2025. Cetvel: A unified benchmark for evaluating language understanding, generation and cultural capacity of llms for turkish. *arXiv preprint arXiv:2508.16431*.

Bahare Fatemi, Mehran Kazemi, Anton Tsitsulin, Karishma Malkan, Jinyeong Yim, John Palowitch, Sungyong Seo, Jonathan Halcrow, and Bryan Perozzi. 2024. Test of time: A benchmark for evaluating llms on temporal reasoning. In *The Thirteenth International Conference on Learning Representations*.

Google. 2024. Introducing Gemini 2.0: our new AI model for the agentic era. Accessed: 2025-12-18.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Xuming Hu, Junzhe Chen, Xiaochuan Li, Yufei Guo, Lijie Wen, Philip S Yu, and Zhijiang Guo. 2024. Towards understanding factual knowledge of large language models. In *The twelfth international conference on learning representations*.

Duygu Sezen Islakoglu and Jan-Christoph Kalo. 2025. Chronosense: Exploring temporal understanding in large language models with time intervals of events. *arXiv preprint arXiv:2501.03040*.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, and 1 others. 2024. OpenAI o1 system card. *arXiv preprint arXiv:2412.16720*.

Raghav Jain, Daivik Sojitra, Arkadeep Acharya, Sriparna Saha, Adam Jatowt, and Sandipan Dandapat. 2023. Do language models have a common sense regarding time? revisiting temporal commonsense reasoning in the era of large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6750–6774.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, and 1 others. 2020. Xglue: A new benchmark dataset for cross-lingual pre-training, understanding and generation. *arXiv preprint arXiv:2004.01401*.

Nasrin Mostafazadeh, Alyson Grealish, Nathanael Chambers, James Allen, and Lucy Vanderwende. 2016. CaTeRS: Causal and temporal relation scheme for semantic annotation of event structures. In *Proceedings of the fourth workshop on events*, pages 51–61.

Qiang Ning, Hao Wu, Rujun Han, Nanyun Peng, Matt Gardner, and Dan Roth. 2020. TORQUE: A reading comprehension dataset of temporal ordering questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1158–1172.

OpenAI. 2024. Introducing GPT-4.1 in the API. Accessed: 2025-12-18.

OpenAI. 2025. GPT-5.1: A smarter, more conversational ChatGPT. Accessed: 2025-12-18.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. Xcopa: A multilingual dataset for causal commonsense reasoning. *arXiv preprint arXiv:2005.00333*.

Lianhui Qin, Aditya Gupta, Shyam Upadhyay, Luheng He, Yejin Choi, and Manaal Faruqui. 2021. TIMEDIAL: Temporal commonsense reasoning in dialog. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7066–7076.

Hans Reichenbach. 1947. Elements of symbolic logic.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, and 1 others. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on machine learning research*.

Qingyu Tan, Hwee Tou Ng, and Lidong Bing. 2023. Towards benchmarking and improving the temporal reasoning capability of large language models. *arXiv preprint arXiv:2306.08952*.

Felix Virgo, Fei Cheng, and Sadao Kurohashi. 2022. Improving event duration question answering by leveraging existing temporal information extraction data. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4451–4457.

Jonas Wallat, Adam Jatowt, and Avishek Anand. 2024. Temporal blind spots in large language models. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, pages 683–692.

Yuqing Wang and Yun Zhao. 2024. Tram: Benchmarking temporal reasoning for large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6389–6415.

Yifan Wei, Yisong Su, Huanhuan Ma, Xiaoyan Yu, Fangyu Lei, Yuanzhe Zhang, Jun Zhao, and Kang Liu. 2023. Menatqa: A new dataset for testing the temporal comprehension and reasoning abilities of large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1434–1447.

Fengli Xu, Qianyue Hao, Zefang Zong, Jingwei Wang, Yunke Zhang, Jingyi Wang, Xiaochong Lan, Jiahui Gong, Tianjian Ouyang, Fanjin Meng, and 1 others. 2025. Towards large reasoning models: A survey of reinforced reasoning with large language models. *arXiv preprint arXiv:2501.09686*.

Arda Yüksel, Abdullatif Köksal, Lütfi Kerem Senel, Anna Korhonen, and Hinrich Schütze. 2024. TurkishMMLU: Measuring massive multitask language understanding in turkish. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7035–7055.

Zhuosheng Zhang, Yao Yao, Aston Zhang, Xiangru Tang, Xinbei Ma, Zhiwei He, Yiming Wang, Mark Gerstein, Rui Wang, Gongshen Liu, and 1 others. 2025. Igniting language intelligence: The hitchhiker's guide from chain-of-thought reasoning to language agents. *ACM Computing Surveys*, 57(8):1–39.

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.

Ben Zhou, Kyle Richardson, Qiang Ning, Tushar Khot, Ashish Sabharwal, and Dan Roth. 2021. Temporal reasoning on implicit events from distant supervision. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1361–1371.

A Question Examples

A.1 Base Question Examples

Task	Modifier	Questions	Answer
ETR	sonra	**CQ:** Esra 136 gün sonra sergiye gidecek. Bugün 25 Kasım 2006 olduğuna göre, Esra ne zaman sergiye gidecek? **DQ:** 25 Kasım 2006 tarihinden 136 gün sonrası ne zamandır?	10 Nisan 2007
ETR	önce	**CQ:** Nergis 13 gün önce işten ayrıldı. Bugün 14 Mayıs 2007 olduğuna göre, Nergis ne zaman işten ayrılmıştır? **DQ:** 14 Mayıs 2007 tarihinden 13 gün öncesi ne zamandır?	1 Mayıs 2007
RTR	sonra	**CQ:** Erdem, yurtdışına çıktıktan 14 gün sonra pasaport çıkardı. Erdem 21 Ağustos 1983 tarihinde pasaport çıkardığına göre, ne zaman yurtdışına çıktı? **DQ:** 21 Ağustos 1983 tarihinden 14 gün öncesi ne zamandır?	7 Ağustos 1983
RTR	önce	**CQ:** Cansu, sözlenmeden 177 gün önce evlendi. Cansu 9 Kasım 1986 tarihinde evlendiğine göre, ne zaman sözlendi? **DQ:** 9 Kasım 1986 tarihinden 177 gün sonrası ne zamandır?	5 Mayıs 1987
RSR	sonra	**CQ:** Barış 64 gün önce "128 gün sonra kontrole gideceğim" dedi. Barış 6 Eylül 2017 tarihinde kontrole gitti ise, bugünün tarihi nedir? **DQ:** 6 Eylül 2017 tarihinden 128 gün öncesinin 64 sonrası ne zamandır?	4 Temmuz 2017
RSR	önce	**CQ:** Ebru 138 gün önce "47 gün önce şehir dışına gittim" dedi. Ebru 21 Ocak 1978 tarihinde şehir dışına gitti ise, bugünün tarihi nedir? **DQ:** 21 Ocak 1978 tarihinden 47 gün sonrasının 138 sonrası ne zamandır?	25 Temmuz 1978
STR	sonra	**CQ:** Zeynep 26 gün sonra yurtdışına çıkacak. Zeynep 28 Eylül 2008 tarihinde yurtdışına çıkacağına göre, bugünün tarihi nedir? **DQ:** 28 Eylül 2008 tarihinden 26 gün öncesi ne zamandır?	2 Eylül 2008
STR	önce	**CQ:** Nur 158 gün önce fotoğraf sergisi açtı. Nur 17 Ekim 1967 tarihinde fotoğraf sergisi açtığına göre, bugünün tarihi nedir? **DQ:** 17 Ekim 1967 tarihinden 158 gün sonrası ne zamandır?	23 Mart 1968

Table 5: Base question examples by task and modifier from the TIMERES benchmark. **ETR**: Event Time Resolution, **STR**: Speech Time Resolution, **RTR**: Reference Time Resolution, **RSR**: Reported Speech Resolution, **CQ**: Contextual Question, **DQ**: Direct Question.

A.2 Compositional Question Examples

Task	Order	Questions (CQ / DQ)	Answer
ETR	AC-First	**CQ:** Bora 77 gün sonra arkadaşlarıyla buluştuğunda, geçen hafta bugün aldığı hediyeyi verecek. Bugün 14 Mayıs 2007 olduğuna göre, Bora hediyeyi ne zaman almıştır? **DQ:** 14 Mayıs 2007 tarihinden 7 gün öncesi ne zamandır?	7 Mayıs 2007
ETR	RC-First	**CQ:** Bora geçen hafta bugün aldığı hediyeyi 77 gün sonra arkadaşlarıyla buluştuğunda verecek. Bugün 14 Mayıs 2007 olduğuna göre, Bora hediyeyi ne zaman almıştır? **DQ:** 14 Mayıs 2007 tarihinden 7 gün öncesi ne zamandır?	7 Mayıs 2007
RTR	AC-First	**CQ:** Buket 50 gün sonra fuara katıldığında, yarın bastıracağı broşürleri elden verecek. Buket 10 Ağustos 1954 tarihinde broşürleri bastıracağına göre, fuara katıldığı tarih nedir? **DQ:** 10 Ağustos 1954 tarihinden 1 gün öncesinin 50 gün sonrası ne zamandır?	28 Eylül 1954
RTR	RC-First	**CQ:** Buket yarın bastıracağı broşürleri 50 gün sonra fuara katıldığında elden verecek. Buket 10 Ağustos 1954 tarihinde broşürleri bastıracağına göre, fuara katıldığı tarih nedir? **DQ:** 10 Ağustos 1954 tarihinden 1 gün öncesinin 50 gün sonrası ne zamandır?	28 Eylül 1954
RSR	AC-First	**CQ:** Büşra 160 gün önce "148 gün sonra spora başladığımda, dünden önceki gün hazırladığım programı uygulayacağım" dedi. Büşra 31 Ekim 2007 tarihinde programı hazırladığına göre, bugünün tarihi nedir? **DQ:** 31 Ekim 2007 tarihinden 2 gün sonrasının 160 gün sonrası ne zamandır?	10 Nisan 2008
RSR	RC-First	**CQ:** Büşra 160 gün önce "dünden önceki gün hazırladığım programı 148 gün sonra spora başladığımda uygulayacağım" dedi. Büşra 31 Ekim 2007 tarihinde programı hazırladığına göre, bugünün tarihi nedir? **DQ:** 31 Ekim 2007 tarihinden 2 gün sonrasının 160 gün sonrası ne zamandır?	10 Nisan 2008
STR	AC-First	**CQ:** Ramazan 51 gün sonra projeyi teslim ettiğinde, dünden önceki gün hazırladığı sunumu yapacak. Ramazan 6 Şubat 2021 tarihinde sunumu hazırladığına göre, bugünün tarihi nedir? **DQ:** 6 Şubat 2021 tarihinden 2 gün sonrası ne zamandır?	8 Şubat 2021
STR	RC-First	**CQ:** Ramazan dünden önceki gün hazırladığı sunumu 51 gün sonra projeyi teslim ettiğinde yapacak. Ramazan 6 Şubat 2021 tarihinde sunumu hazırladığına göre, bugünün tarihi nedir? **DQ:** 6 Şubat 2021 tarihinden 2 gün sonrası ne zamandır?	8 Şubat 2021

Table 6: Compositional question examples by task and word order from the TIMERES benchmark. **ETR**: Event Time Resolution, **STR**: Speech Time Resolution, **RTR**: Reference Time Resolution, **RSR**: Reported Speech Resolution, **CQ**: Contextual Question, **DQ**: Direct Question, **AC**: Adverbial clause, **RC**: Relative Clause.