A APPENDIX

Relation	#Qs	Template
member of sports team	9033	<sub.> played for which team in <year>?
position held	7343	<sub.> held which position in <year>?
employer	9049	<sub.> worked for which company in <year>?
political party	7324	<sub.> was a member of which party in <year>?
head coach	4886	<obj.> was the head coach of which team in <year>?
educated at	1672	<sub.> attended which university in <year>?
chairperson	4190	<obj.> was the chair of which entity in <year>?
head of government	4125	<obj.> is the head of the government of which state in <year>?
owned by	2688	<sub.> is owned by whom in <year>?

Table 6: TempLAMA reformulations to natural questions.

`alapaca-7B`, `open-llama-7B`:

Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.

Instruction:
`Answer the question.`

Input:
{**davinci prompt** (except last line) – see below}

Response:

`text-davinci-003`, `falcon-7B`, `red-pajama-3B`, `red-pajama-7B`:

I am a highly intelligent question answering bot. If you ask me a question that is rooted in truth, I will give you the answer. If you ask me a question that is nonsense, trickery, or has no clear answer, I will respond with 'Unknown'.

`Q: What is human life expectancy in the United States?`
A: 78 years

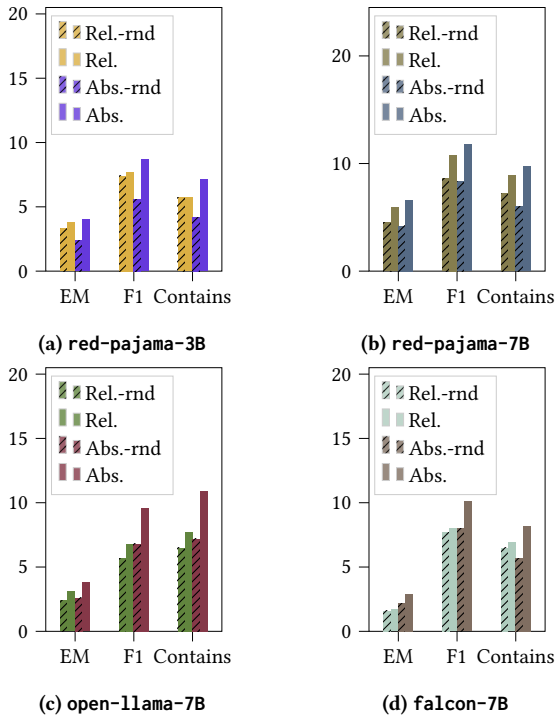
`Q: Who was president of the United States in 1955?`
A: Dwight D. Eisenhower

`Q: Which party did he belong to?`
A: Republican Party

`Q: Where were the 1992 Olympics held?`
A: Barcelona, Spain.

Q: {question}
A:

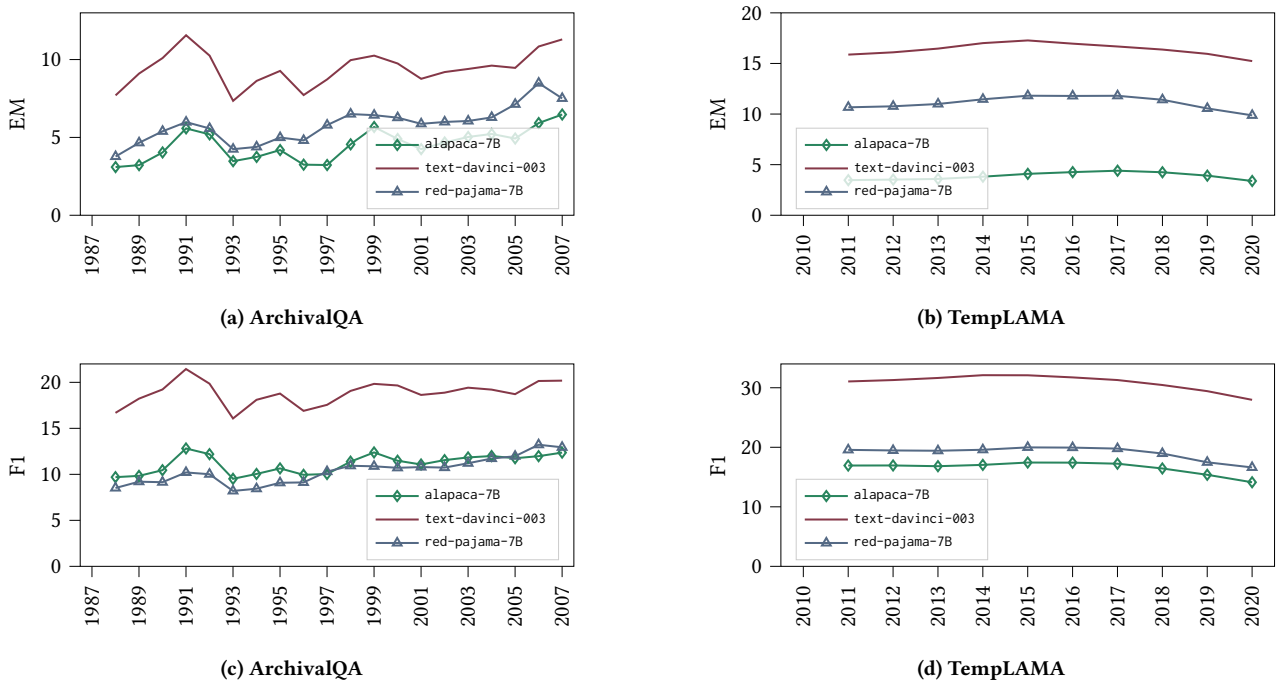
Table 7: Prompts used in this study. {question} is replaced with the current question.



(a) `red-pajama-3B` (b) `red-pajama-7B`

(c) `open-llama-7B` (d) `falcon-7B`

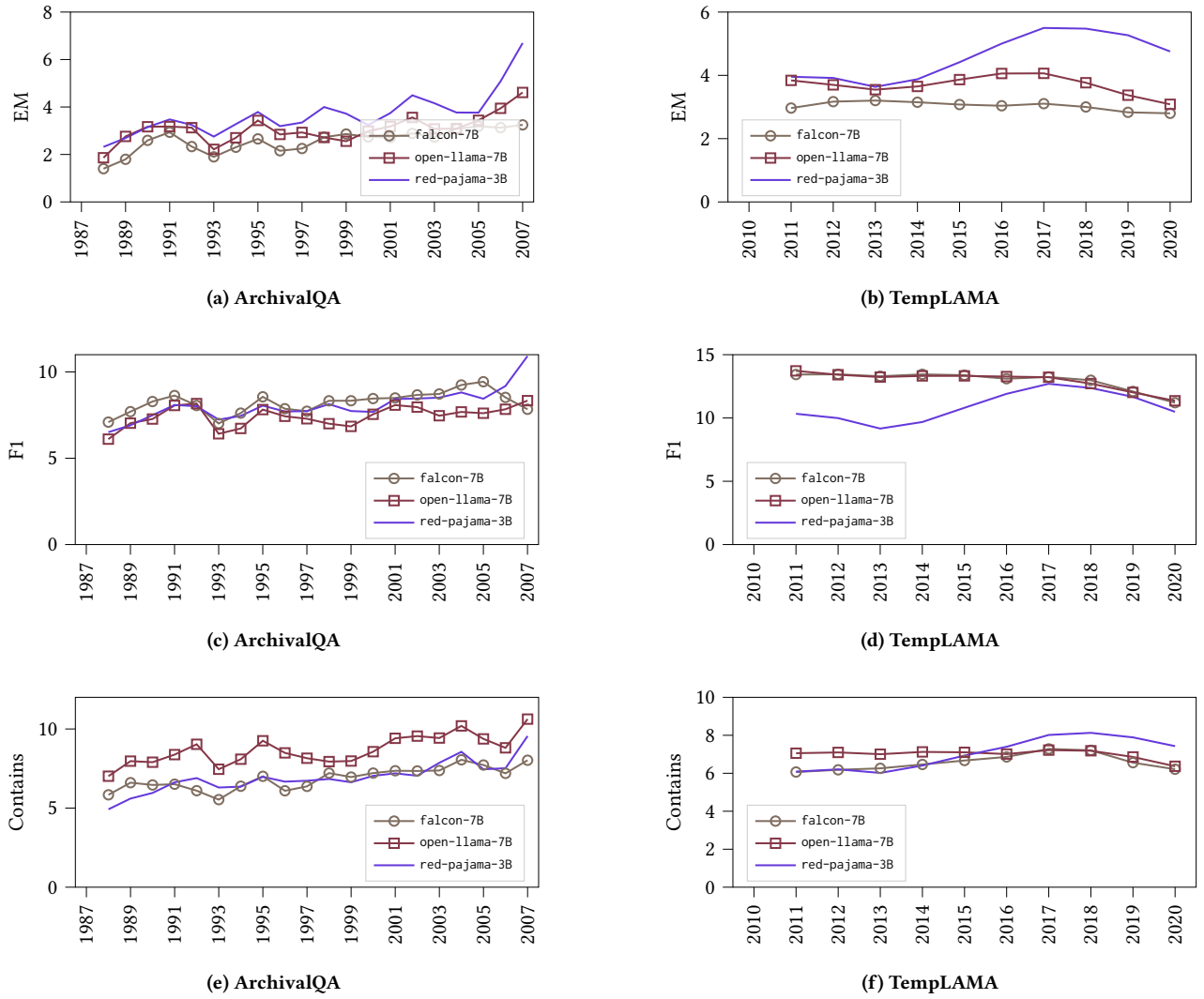
Figure 7: Effect of randomized relative and absolute time references. Textured bars show the randomized variants.



(a) ArchivalQA (b) TempLAMA

(c) ArchivalQA (d) TempLAMA

Figure 8: Stratification of the `alapaca-7B`, `red-pajama-7B`, and `text-davinci-003` models on the ArchivalQA and TempLAMA datasets. Stratified by years, the trendline is the moving average with a window of 2. We do not show plots for the TemporalQuestions dataset since the dataset is not large enough for computing individual results per year.



(a) ArchivalQA (b) TempLAMA

(c) ArchivalQA (d) TempLAMA

(e) ArchivalQA (f) TempLAMA

Figure 9: Stratification of the `falcon-7B`, `open-llama-7B`, and `red-pajama-3B` models on the ArchivalQA and TempLAMA datasets. Stratified by years, the trendline is the moving average with a window of 2. We do not show plots for the TemporalQuestions dataset since the dataset is not large enough for computing individual results per year.