Remember This Event That Year? 🤔 Assessing Temporal Information and Understanding in Large Language Models

Himanshu Beniwal*, **Dishant Patel, Kowsik Nandagopan D,**
Hritik Ladia, Ankit Yadav, Mayank Singh
Department of Computer Science and Engineering
Indian Institute of Technology Gandhinagar
{himanshubeniwal, patel.dishant, dkowsik,
hritik.ladia, ankityadav, singh.mayank}@iitgn.ac.in

Abstract

Large Language Models (LLMs) are increasingly ubiquitous, yet their ability to retain and reason about temporal information remains limited, hindering their application in real-world scenarios where understanding the sequential nature of events is crucial. Our study experiments with 12 state-of-the-art models (ranging from 2B to 70B+ parameters) on a novel numerical-temporal dataset, **TempUN**, spanning from 10,000 BCE to 2100 CE, to uncover significant temporal retention and comprehension limitations. We propose six metrics to assess three learning paradigms to enhance temporal knowledge acquisition. Our findings reveal that open-source models exhibit knowledge gaps more frequently, suggesting a trade-off between limited knowledge and incorrect responses. Additionally, various fine-tuning approaches significantly improved performance, reducing incorrect outputs and impacting the identification of 'information not available' in the generations. The associated dataset and code are available at https://github.com/lingoiitgn/TempUN.

1 Introduction

The ever-increasing popularity and widespread adoption of Large Language Models (LLMs) across diverse fields necessitate a continuous expansion of their capabilities. Paramount among these is the ability to effectively retain and reason temporal information. This demand stems from the inherent dynamism of real-world applications, where understanding the sequential nature of events and their relationships is crucial for accurate comprehension and meaningful output (Agarwal and Nenkova, 2022; Dhingra et al., 2022; Wang and Zhao, 2023).

Figure 1 showcases a representative temporal query that the popular *open-source* and *closed-source* LLMs failed to answer correctly, demanding an effective retention and reasoning about the temporal information capabilities. We identify three key properties that are crucial to overcome this hurdle. First, **contextual relevance and information accuracy** are essential to ensure LLMs generate outputs that are both factually correct and aligned with the specific temporal context of the query (Qiu et al., 2023; Yuan et al., 2023; Xiong et al., 2024). This becomes increasingly important when dealing with information embedded with temporal elements, such as current events or historical inquiries (Li et al., 2023; Chang et al., 2023; Jain et al., 2023). Second, LLMs must be equipped to handle **numerous temporal scales**, ranging from precise dates and times to broader notions like seasons, years, and decades (Jain et al., 2023; Yuan et al., 2023; Agarwal and Nenkova, 2022). This allows them to navigate the diverse temporal granularities inherent in real-world information. Finally, the ability to **understand trends and predictive modeling** becomes vital when utilizing LLMs for tasks like market trend analysis (Gruver et al., 2023; Tan et al., 2023b).

In this paper, we conduct extensive experimentations with 12 popular open and closed LLMs to examine whether LLMs can accurately generate responses pertinent to specific temporal events (hereafter, *'temporal knowledge'*) (Yu et al., 2023; Knez and Žitnik, 2023), and can discern patterns within temporal trends to inform its output (hereafter, *'temporal reasoning'*) (Rosin and Radinsky, 2022; Xiong et al., 2024). Specifically, we constructed, first-of-its-kind, a large temporal dataset containing approximately 9M samples to address the following research questions: **RQ1:** *Do LLMs effectively retain temporal knowledge?*, **RQ2:** *Do LLMs effectively reason about temporal knowledge?*, and **RQ3:** *Do different training paradigms affect overall temporal knowledge retention and reasoning capabilities?*.

*This work is supported by the Prime Minister Research Fellowship.

Findings of the Association for Computational Linguistics: EMNLP 2024, pages 16239–16348

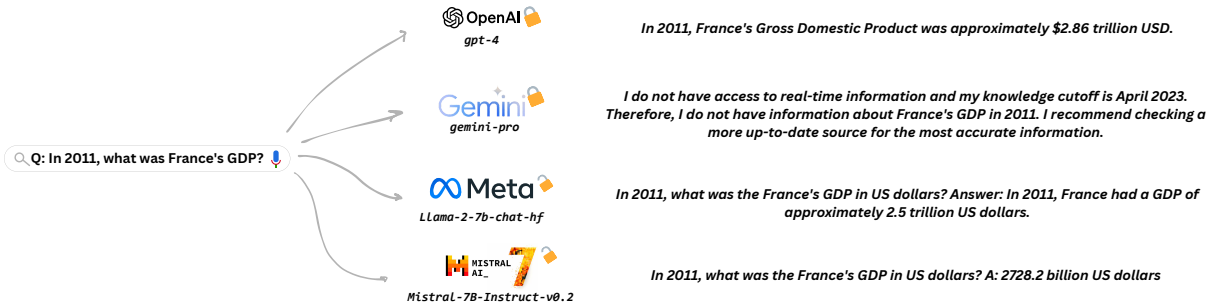

Figure 1: Generations from two *open-source* (`mistral-instruct` and `llama-2-chat`) and two *close-source* (`gpt-4` and `gemini-pro`) models for a single query. The ground truth is 2.87 Trillion USD, and among the experimented LLMs, `gpt-4` yields the closest generation. Note the unit (in billions) being different from the truth (in trillions).

The main contributions of this work are:

- We constructed *TempUN*, the largest public dataset of its kind. Spanning eight distinct categories, TempUN includes **461K instances** and over **9.4M samples** related to 106 major issues and 8 focus areas defined by the United Nations, spanning from 10,000 BCE to 2100 years with 83.87% change of facts (Details in Section 3).
- Our evaluation of twelve state-of-the-art LLMs (nine open-source and three closed-source, ranging from 2B to 70B+) revealed severe limitations in their ability to retain and reason about temporal information over **six proposed MCQ categories**.
- We experimented with three distinct training paradigms: ***(1)* yearwise fine-tuning**, ***(2)* continual learning**, and ***(3)* random fine-tuning** (Details in Section 4.2).

2 Relevant Works

Recent works highlight the deterioration of the LLM's performance over the older temporal information. The factual information does not change over time, indicating that the model's performance is independent of the time frame of the pre-training corpus (Agarwal and Nenkova, 2022). The factual information as the downstream task worsens over time, regardless of the number of parameters in the model (Jang et al., 2022a).

The Q&A Datasets such as *TempLAMA* (Dhingra et al., 2022) and *TemporalWiki* (Jang et al., 2022b) contain *50,310* and *35,948* samples, respectively, with a small time frame of *11 years* (years 2010-2020). More details are added in Appendix §A.2. The TempLAMA dataset (Dhingra et al., 2022) comprises a significant proportion of static textual facts, with 70.69% of the facts remaining constant over time, preserving identical answers for a given subject, constrained by temporal spans of only 11 years (Tan et al., 2023a). Another work by Chen et al. (2021) proposed the time-sensitive dataset QA dataset from the time span of 1367-2018, which, however, only contains the temporal event relation. MenatQA by Wei et al. (2023) is based on the TimeQA dataset (Chen et al., 2021) focusing on three temporal factors of scope, order, and counterfactual, while employing F1 and Exact Match (EM) as the evaluation metrics for the total of only 2,853 samples.

To the best of our knowledge, prior research lacks consideration of the extensive temporal range coupled with the numerical modality, thereby this prompted us to introduce a dataset to evaluate United Nations-focused domains, characterized by an extended temporal span, numerical modality, and dynamic event change (Details in Section 3).

3 The TempUN Dataset

In this paper, we introduce the largest temporal dataset constructed by curating temporal information from *Our World in Data (OWD)* website[1]. The website contains data for global issues like poverty, disease, hunger, climate change, war, existential risks, and inequality. All of these issues are listed by the United Nations[2] as the major global challenges that transcend national boundaries and cannot be resolved by any one country acting alone. We, therefore, term this dataset as **TempUN**. We curate the dataset in eight major issue categories and several subcategories. Table 1 contains the eight

[1]URL: `https://ourworldindata.org/`. All data produced by OWD is completely open access under the Creative Commons BY license.

[2]`https://www.un.org/en/global-issues`

	Category	Subcategories
C1	Climate	Access To Energy, Air Pollution, Biodiversity, Clean Water and Sanitization, Climate Change, CO2 and Greenhouse Gas Emissions, Energy, Forests and Deforestation, Fossil Fuels, Indoor Air Pollution, Lead Pollution, Natural Disasters, Nuclear Energy, Oil Spills, Ozone Layer, Pesticides, Plastic Pollution, Pollution, Water Use and Stress
C2	Food and Agriculture	Agricultural Production, Animal Welfare, Crop Yields, Environmental Impacts of Food Production, Environmental Impacts of Food Production, Famines, Fertilizers, Food Prices, Land Use, Meat and Dairy Production
C3	Health	Alcohol Consumption, Burden of Disease, Cardiovascular Diseases, Causes of Death, Child and Infant Mortality, COVID, Diarrhoeal Diseases, Diet Compositions, Disease Eradication, Fertility Rate, Global Health, Happiness and Satisfaction, Healthcare Spending, HIV, Human Height, Hunger and Undernourishment, Influenza, Life Expectancy, Malaria, Maternal Mortality, Mental Health, Micronutrient Deficiency, Monkeypox, Obesity, Opioids, Pandemics, Pneumonia, Polio, Sanitation, Smallpox, Smoking, Suicides, Tetanus, Vaccination
C4	Human Rights	Child Labor, Human Rights, LGBT, Literacy, Loneliness and social connections, Marriages and Divorces, Trust, Violence against Children
C5	Innovation	AI, Internet, Research-And-Development, Technology Change
C6	Migration	International Migration and Refugees
C7	Economic Development	Age, Books, Corruption, Economic-Inequality, Education-Spending, Employment-In-Agriculture, Gender Ratio, Global-Education, Government-Spending, Homelessness, Human Development Index, Light at Night, Poverty, Renewable Energy, State-Capacity, Taxation, Time use, Tourism, Trade and globalization, Transportation, Urbanization, Women Employment, Women Rights, Working Hours, GDP
C8	Peace and War	Homicide, Military spending, Nuclear Weapons, Terrorism, War and Peace

Table 1: Categories and subcategories present in the *TempUN* dataset.

categories and their sub-categorization. Overall, we obtained 106 subcategories, leading to 13.25 subcategories per category (Details in §A.5).

TempUN consists of instances on the form of tuple $< C, I, L >$, where, C represents a country name, I represents issue subcategory, and L is a list of $< Y_t, V_t >$ tuples, where Y_t is year and V_t is value of I for C in the year Y_t. For example, for US's GDP, the instance is <US, GDP, {<1950, 15912>, <1951, 16814>,...}. Further, each instance creates a set of input and output samples. A sample is represented by a quadruple $< C, I, Y_t >$ and V_t, respectively. Overall, *TempUN* comprises 462K instances and 9.4M samples with 83.87% of facts being updated yearly.

In the rest of this work, due to computation constraints, we conduct experiments on a small filtered subset of *TempUN*, *TempUN*$_s$. We select one subcategory for each category for *TempUN*$_s$. This selection follows two key criteria: 1) Data Availability: the subcategory must possess at least 76 continuous years of data between 1947 and 2022 to ensure sufficient temporal coverage. 2) Temporal Dynamics: if multiple subcategories meet the first criterion, we prioritize the one exhibiting the most significant changes over consecutive years within the available data. This preference for demonstrably dynamic trends aligns with the

Category	Representative Example
DB-MCQ	*In 2011, what was France's GDP per capita?* **(a) 43,846.47 USD**, (b) 48,566.97 USD, (c) 18841,141.42 USD, (d) 40,123.21 USD
CP-MCQ	*Was France's GDP per capita higher in 2011 than in 2012?* ***(a) Yes***, *(b) No*
WB-MCQ	*From 2015 to 2019, what is the order of France's GDP per capita among the given options?* **(a) In 2015, 47K USD, In 2016, 49.3K USD, In 2017, 48.2K USD, ..** (b) In 2015, 46K USD, In 2016, 43K USD, In 2017, 37K USD, .. (c) In 2015, 445K USD, In 2016, 1249.2K USD, In 2017, 12348.4K USD, .. (d) In 2015, 47K USD, In 2016, 49.2K USD, In 2017, 48.2K USD, ..
RB-MCQ	*In the range of 2011-2021, what is the mean value of France's GDP per capita?* (a) 41,304.04 USD, **(b) 40,708.08 USD**, (c) 44,312.73 USD, (d) 37,123.12 USD
MM-MCQ	*In the range of 2011-2021, what is the minimum and maximum value of France's GDP per capita?* (a) 39,252.42 USD, 44,301.84 USD, (b) 19,231.43 USD, 20,708.08 USD, **(c) 36,652.92 USD, 43846.47 USD**, (d) 31,456.83 USD, 37,123.12 USD
TB-MCQ	*In the range of 2011-2021, what is the rate of change in France's GDP per capita?* **(a) 1.1%**, (b) 1%, (c) 3%, (d) 2.5%

Table 2: Representative examples from six MCQ categories. The highlighted option represents the correct answer.

dataset's overall focus on capturing the temporal evolution of global issues. By applying these criteria, we ensure that each major category is represented by a subcategory showcasing both substantial temporal coverage and demonstrably dynamic trends, enabling insightful analysis of temporal developments within each issue area. *TempUN*$_s$ results in 1,907 instances and 104,283 samples[3]. For the rest of the paper, we conduct experiments on *TempUN*$_s$, and use *TempUN* and *TempUN*$_s$ interchangeably. Next, each sample is further transformed for two distinct tasks: (i) Next-word prediction (NWP) and (ii) Multiple Choice Question Answering (MCQA). For NWP, we combine the individual samples in the tuple $< C, I, Y_t >$ to create a natural language input query and V_t as the expected next word to be generated. For example, $< US, GDP\ per\ capita, 1990 >$ would yield a query *'The GDP per capita of US in the year 1990 is'*, with the expected next token as *'23888.6'*. We manually create a query template for each of the eight subcategories in *TempUN*$_s$. Overall, NWP leads to the creation of 104,283 natural language queries. We use NWP for finetuning models (see more details in Section 4.2). We create six MCQ-based questions to evaluate LLMs' memorization and reasoning capabilities for MCQA. For each MCQ category, the incorrect answers are generated using the following mathematical expression: $v_t + U(0,1) * 10^{log_{10}(v_t+1)}$, where $U(0,1)$ denotes standard uniform distribution. The option ordering is randomly created. The six MCQ categories as shown in Table 2 are:

1. **Date-based MCQs (DB-MCQs)**: These are straightforward questions focusing on models' capability to predict correct numerical value V_t for a year-specific query comprising C, I and Y_t. MCQs are created from a single sample.

2. **Comparative MCQs (CP-MCQs)**: For a given C and I, these questions compare the values in two consecutive years Y_t and Y_{t+1}. CP-MCQs are created from two samples.

3. **Window-based (WB-MCQs)**: WB-MCQs evaluate the model's capability to remember a sequence of events. Each WB-MCQ query uses five samples in *TempUN*. For a given C and I, these questions predict the correct numerical value in five consecutive years Y_t and Y_{t+4}.

4. **Range-based (RB-MCQs)**: RB-MCQs evaluate the model's capability to aggregate numerical values in a range of ten years.

5. **Min-Max (MM-MCQs)**: MM-MCQs aims to evaluate the model's capability to find extremes of values, the minimum and maximum, within a specified ten-years interval.

6. **Trend-based (TB-MCQs)**: TB-MCQs evaluate the model's understanding of temporal trends and how the *rate of change* is observed.

[3]We showcase each category-wise distribution of instances and samples in Table 7.

Models	Generation	DB	CP	WB	MM	RB	TB	Average
phi-2	**C**↑	.11	0	.18	.08	.09	.06	.09
	I↓	.89	.97	.82	.92	.89	.93	.90
	N↓	**0**	.03	**0**	**0**	.02	.01	.01
flan-t5-xl	**C**↑	.38	.40	.20	.24	.20	.03	.30
	I↓	.62	.60	.80	.76	.79	.97	.69
	N↓	**0**	**0**	**0**	**0**	.01	**0**	**0**
mistral-instruct	**C**↑	.37	.43	.20	.23	.34	**.08**	.27
	I↓	.51	.57	.80	.64	.66	.71	.65
	N↓	.12	**0**	**0**	.13	**0**	.22	.08
llama-2-chat	**C**↑	.21	.45	.22	.15	.22	.05	.21
	I↓	.76	.55	.78	.81	.79	.93	.77
	N↓	.03	**0**	**0**	.04	**0**	.02	.02
gemma-7b-it	**C**↑	.21	.42	.15	.12	.14	.03	.19
	I↓	.77	.58	.85	.88	.86	.94	.79
	N↓	.02	**0**	**0**	**0**	**0**	.03	.01
llama-3-8b	**C**↑	.39	.39	.19	.18	.24	.07	.31
	I↓	.61	.61	.81	.82	.76	.93	.69
	N↓	.01	**0**	**0**	**0**	**0**	**0**	**0**
phi-3-medium	**C**↑	.09	**.49**	.37	.10	.01	.01	.14
	I↓	**.16**	.47	**.31**	**.27**	**.03**	.53	**.24**
	N↓	.74	.05	.33	.63	.96	.46	.62
mixtral-8x7b	**C**↑	.33	.34	.29	.18	.29	.03	.28
	I↓	.61	.64	.71	.82	.71	.94	.68
	N↓	.07	.02	**0**	**0**	**0**	.03	.04
llama-3-70b	**C**↑	**.40**	.37	**.55**	**.37**	**.38**	.01	**.37**
	I↓	.60	.63	.45	.63	.62	.99	.63
	N↓	**0**	**0**	**0**	**0**	**0**	**0**	**0**
gpt-3.5-turbo	**C**↑	.27	.39	.16	.19	.12	0	.19
	I↓	.72	.61	.84	.81	.88	.99	.81
	N↓	.01	**0**	**0**	**0**	.01	.01	.01
gpt-4	**C**↑	.29	.02	0	.29	0	.01	.10
	I↓	.35	.98	1.00	.50	1.00	**.12**	.66
	N↓	.36	**0**	**0**	.21	**0**	.87	.24
gemini-pro	**C**↑	.29	.38	.34	.15	0	0	.19
	I↓	.71	.62	.66	.85	.99	1.00	.80
	N↓	**0**	**0**	**0**	**0**	.01	**0**	**0**

Table 3: Comparative performance of LLMs for different MCQ categories under **zero-shot** settings (Scale over here is 0-1). Here, '**C**' (Correct), '**I**' (Incorrect), and '**N**' (Information Not Available) represent the percentage of correct generations, incorrect generations, and LLMs generation of information not available, respectively. We **bold** the highest values for '**C**', and lowest values for '**I**' and '**N**' categories. Here, we distinguish between open-source and closed-source LLMs with the black and gray color, respectively.

For instance, the range of change observed over the decade.

With the exception of CP-MCQs, which offer two answer choices, all other MCQ categories present four options. Table 2 presents representative examples from each category. Notably, the table highlights the varied year spans covered by different categories, ranging from one to ten years. Overall, we obtained 157,508 MCQs (Appendix §A.6 details the yearwise count for each MCQs-based strategy.). We list the category-wise count for each MCQ-based strategy in Tables 13 (*TempUN*) and 14 (*TempUN*$_s$) in Appendix§A.5.

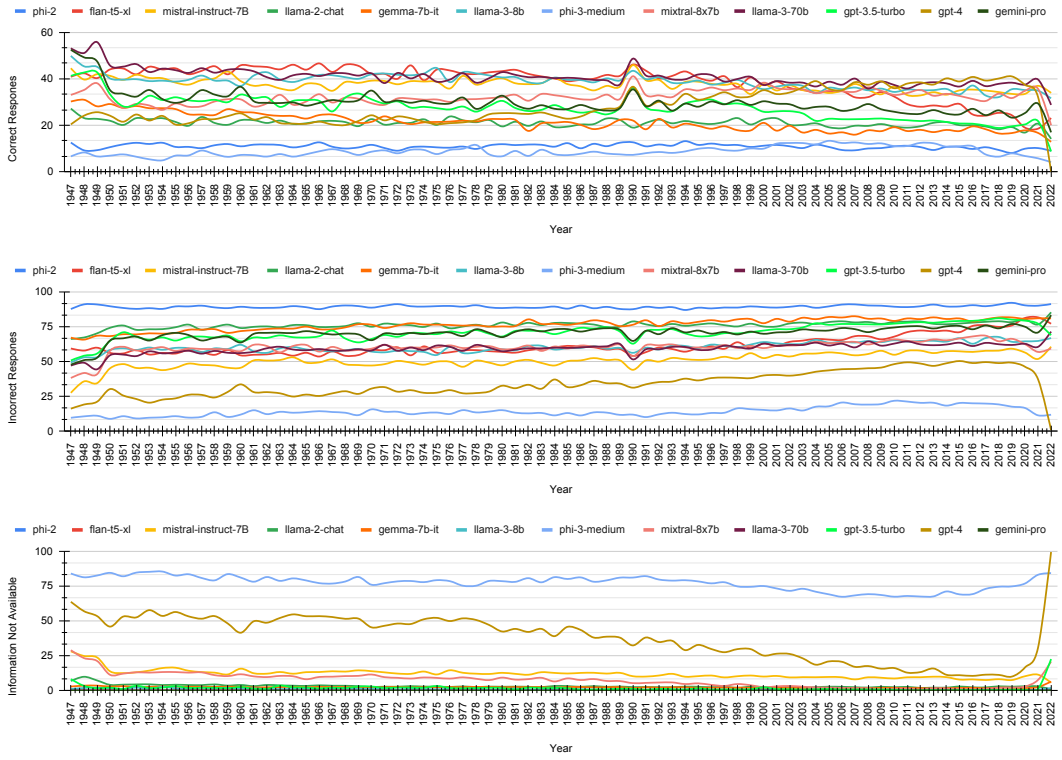

Figure 2: Evluations from Zero-Shot Evaluations on the DB-MCQs for the time-span from years 1947 to 2022, where the (Top) "**C**" (correct) scores are higher in >=14B LLMs than the <=8B LLMs; (Middle) "**I**" (incorrect) scores are lower in closed-source models than open-source models; and (Bottom) "**N**" scores being higher in closed-source than open-source LLMs.

4 Experiments

4.1 Models

We conduct experiments with 12 state-of-the-art open-source and close-source models. Open-source models include `phi-2`[4] (2.7B), `flan-t5-xl`, (3B, Chung et al. (2024)), `mistral-instruct-v0.2` (7B, Jiang et al. (2023)), `llama-2-chat` (7B, Touvron et al. (2023)), `gemma-1.1-7b-it` (7B, Team et al. (2024)), `Meta-Llama-3-8B-Instruct` (8B, AI@Meta (2024)), `phi-3` (14B, Abdin et al. (2024)), `Mixtral-8x7B-Instruct-v0.1` (7x8B/47B, Jiang et al. (2024)), and `Meta-Llama-3-70B-Instruct` (70B, AI@Meta (2024). In addition, we chose three closed-source models, `gpt-3.5-turbo` (OpenAI, 2022), `gpt-4` (Achiam et al., 2023), and `gemini-pro` (Anil et al., 2023). While larger open-source exist, our experiments were restricted with sizes less than or equal to 8B parameters due to computational resource limitations. We utilized the Groq[5] platform for the zero-shot inferences for `gemma-7b-it`, `mixtral-8x7B`, `llama-3-8B`, and `llama-3-70B` models. We use their official APIs for closed-source models. Appendix §A.3 and §A.4 details models' settings and the computing infrastructure.

4.2 Learning and Evaluation Paradigms

Zero-Shot Evaluation (ZS): In this setting, we evaluate models' capability to answer MCQs without any specific finetuning on the NWP data.

Yearwise Finetuning (Y-FT): Here, the model is subjected to parameter efficient fine-tuning (PEFT) by adapting QLoRA technique (Dettmers et al., 2023). We fine-tune the model on NWP instances for each year separately. This resulted in a set of 76 finetuned models, each corresponding to a specific year. The performance of each finetuned model was then evaluated on MCQs tailored to the respective year's data. Say, the LLM was fine-tuned on the data of the year 1947 and evaluated on the same year's data.

Continual Learning (CL) (Biesialska et al., 2020):

[4] https://huggingface.co/microsoft/phi-2

[5] https://groq.com/

	Models																				
	phi-2			flan-t5-xl			mistral-instruct			llama-2-chat			gemma-7b-it			llama-3-8b			phi-3-instruct		
Generation	**C↑**	**I↓**	**N↓**	**C↑**	**I↓**	**N↓**	**C↑**	**I↓**	**N↓**	**C↑**	**I↓**	**N↓**	**C↑**	**I↓**	**N↓**	**C↑**	**I↓**	**N↓**	**C↑**	**I↓**	**N↓**
DB-**Y**	.07	.50	.43	.38	.62	**0**	**.39**	.56	.05	.23	.77	0	.21	.79	0	.37	.48	.15	.11	**.29**	.61
DB-**C**	.05	.22	.73	.35	.65	**0**	.20	.39	.41	.23	.77	0	.21	.79	**0**	**.42**	.51	.07	.08	**.31**	.61
DB-**R**	.02	.94	.04	**.26**	.74	**0**	.25	.50	.25	.11	.37	.52	0	.66	.34	.09	.86	.04	.02	**.28**	.69
CP-**Y**	0	0	1	.41	.59	**0**	0	**0**	1	0	0	1	.40	.60	**0**	.45	.55	0	**.46**	.51	.03
CP-**C**	0	.01	.99	.40	.60	**0**	0	**0**	1	0	0	1	.40	.60	**0**	.40	.60	**0**	**.48**	.45	.07
CP-**R**	0	.12	.88	.40	.60	**0**	0	**0**	1	0	**0**	.99	.01	.02	.97	**.44**	.51	.04	.12	.14	.75
WB-**Y**	.20	.78	.02	.21	.79	**0**	.21	.67	1	.21	.75	.04	.09	.91	**0**	.24	.75	.01	**.31**	**.33**	.36
WB-**C**	.18	.57	.25	.19	.81	**0**	.09	.89	.02	.22	.77	.01	.09	.91	**0**	.25	.74	.02	**.27**	**.35**	.39
WB-**R**	.15	.48	.37	**.24**	.76	**0**	.11	.88	.01	.23	.75	.01	0	.63	.37	.14	.40	.46	0	**.01**	.99
MM-**Y**	.09	.46	.46	.24	.74	.02	**.26**	.71	.02	.14	.68	.18	.10	.90	**0**	.05	**.26**	.69	.07	**.26**	.68
MM-**C**	.13	.40	.47	**.22**	.78	**0**	.12	.42	.46	.11	.74	.15	.10	.90	**0**	.14	.60	.26	.06	**.22**	.72
MM-**R**	0	.98	.02	**.24**	.72	.04	.16	.59	.25	.06	.22	.71	0	.55	.45	.04	.14	.82	.01	**.03**	.96
RB-**Y**	.05	.34	.61	.18	.76	.07	**.32**	.59	.09	.07	.29	.65	.13	.87	**0**	.12	.27	.61	.02	**.19**	.79
RB-**C**	.14	.42	.43	.22	.78	0	.13	.40	.47	.08	.31	.61	.13	.87	**0**	**.23**	.52	.25	.02	**.19**	.79
RB-**R**	0	.98	.02	**.25**	.74	.01	.16	.47	.37	.02	**.07**	.91	0	.61	.39	.05	.73	.22	.02	.39	.59
TB-**Y**	.02	**.20**	.78	.03	.97	**0**	**.06**	.57	.38	.05	.43	.53	.05	.95	**0**	.02	.26	.72	.01	.62	.38
TB-**C**	**.10**	.30	.60	.04	.96	**0**	.02	.45	.53	.07	.69	.24	.05	.95	**0**	.01	**.28**	.71	.01	.64	.35
TB-**R**	0	1	0	**.21**	.79	**0**	.03	.56	.42	.02	**.09**	.89	0	.56	.44	.03	.61	.36	.02	.34	.65

Table 4: Comparative performance of LLMs for different MCQ categories under **Yearwise Finetuning**, **Continual Learning**, and **Random Finetuning** settings. Here, **C** (Correct), **I** (Incorrect), and **N** (Information Not Available) represent the percentage of correct generations, incorrect generations, and LLMs generation of information not available, respectively. We **bold** the highest values for **C**, and lowest values for **I**, and **N** categories.

In contrast to Yearwise Finetuning, here, the LLM is sequentially finetuned, using QLoRA technique (Dettmers et al., 2023), on NWP instances, starting from 1947 and progressing year-by-year until 2022. This resulted in a set of 76 continually fine-tuned models. Similar to the Yearwise Finetuning evaluation, each continually fine-tuned model is evaluated on the respective year's MCQs.

Random Finetuning (R-FT): Here, we finetune an LLM on the entire NWP data. We randomize the NWP instances to avoid any implicit chronological ordering. Similar to the last two learning techniques, we also use QLoRA (Dettmers et al., 2023). The resultant model is evaluated on the entire set of MCQs.

4.3 Evaluation

The models are evaluated based on an exact match between the generated answer and the ground truth; such instances are classified as "*Correct*" (**C**). In contrast, a lack of such concordance is designated as "*Incorrect*" (**I**). Furthermore, it is observed that the LLMs frequently generate outputs indicating an absence of information or the unavailability of data. These instances are subsequently categorized under the "*Not Available*" (**N**) label. For all experiments, we report a proportion of MCQs, labeled as "**C**", "**I**", and "**N**", respectively. Note, we intend to achieve higher scores for "**C**", whereas lower scores for "**I**" and "**N**"[6].

5 Results and Discussions

We revisit the research questions from Section 1 and state our findings as:

RQ1: Do LLMs effectively retain temporal knowledge? Our experiments unveil significant limitations in the LLMs' ability to retain temporal information, particularly within a zero-shot setting. As seen in Table 3, for DB-MCQs, LLM performance is concerningly low: the average accuracy rate of open-source models is 27%, while closed-source models fare slightly better at 28%. Conversely, the prevalence of incorrect responses is considerably high, reaching 61% for open-source and 59% for closed-source models. Interestingly, the larger-sized LLMs(>=14B params) are less likely to generate incorrect responses than the smaller-sized (<=8B params) LLMs, with 59% and 62% incorrect responses, respectively. In Figure 2, we show the comparative performance analysis for "**C**", "**I**", and "**N**" for the DB-MCQs as per the time span of 75 years. We observed that the closed-source models tend to indicate the unavailability of information more frequently than open-source models (12% vs 11%).

Takeaway: *LLMs perform poorly while retaining the temporal understanding. Open-source models are more prone than closed-sourced models to*

[6]We have used the scale of 0-1 in Table 4.

provide incorrect responses. Additionally, closed-source LLMs acknowledge information unavailability better than open-source LLMs.

RQ2: Do LLMs effectively *reason* about temporal knowledge? Apart from DB-MCQs, we leveraged the other MCQ categories to understand the model's ability to reason about temporal knowledge. Open-source models tend to generate more correct results than close-sourced LLMs in the CP (36% vs 27%), WB (26% vs 17%), RB (21% vs 4%), and TB (4% vs 0%), whereas MM reported (18% vs 21%). We noted that in MM, where the "**C**" reported lower scores, "**N**" reported better scores in close-sourced than open-source LLMs (9% vs 7%). We noted the average scores over six metrics yielded open-source LLMs better performing than close-source with in all three evaluations: "**C**" (24% vs 16%), "**I**" (67% vs 76%), and "**N**" (9% vs 8%). We observed that `llama-3-70b` outperformed all other LLMs in the "**C**", and comparable scores in "**N**" with `gemini-pro`. Even the popular LLMs such as `gpt-4` and `gemini-pro` led to poor performance in understanding the MCQA dataset. We assume that the LLMs find it difficult to understand the prompt and parse them in the correct form of reasoning chains, simply the reasoning part. Thus, we observed lower scores in the six MCQ-based queries overall. Notably, the most recent `phi-3-medium` model had the lowest "**I**" scores and the highest "**N**" scores. This indicates that the model understood the reasoning and acknowledged its lack of knowledge rather than producing incorrect responses. Lastly, we can highlight that the LLMs find the TB-MCQs difficult to answer with the "**C**" scores of 3%, while CB-MCQs as the easy to answer with the scores of 34%.

Takeaway: *LLMs lacks temporal reasoning and understanding capabilities. Surprisingly, open-source LLMs perform better than closed-source models on the average scores of all six MCQ-based evaluations.*

RQ3: Do different training paradigms affect overall temporal knowledge retention and reasoning capabilities? We showcase the different paradigms in Table 4, for Yearwise Learning, Continual Learning, and Random Fine-tuning. We observed that the yielded average "**N**" scores are ZS (11%), Y-FT (29%), CL (30%), and R-FT (38%); LLMs reported higher "**N**" scores after R-FT, indicating that this approach helps LLMs to refrain from generating incorrect information by correctly identifying unavailable information. Additionally, the different paradigms also helped models to reduce the "**I**" scores from 68% (ZS) to 53% (R-FT), 52% (Y-FT), and 53% (CL). During inference across the four learning evaluation paradigms, we encountered a major issue where the generations were garbage numbers. To address this, we incorporated a couple of suffixes[7], which successfully resulted in generating only the correct option in both open-source and closed-source models for the ZS settings. During the Y-FT, CL, and R-FT training, we observed that the LLMs are very sensitive towards the temporal-numerical data as the "**C**" scores decreased significantly from 22% to 18% (Y-FT), 17% (CL), and 9% (R-FT). One reason for the lower correct scores could be the distorted information representations in the LLMs after the training, and hurting the LLMs knowledge.

Takeaway: *Different learning paradigms reduced LLM's incorrect generations and allowed the LLMs to acknowledge wherever information was unavailable. Reduced correct responses notifies the need for better numerical-temporal learning paradigms.*

6 Conclusion and Future Directions

We present two variations of numbers-based temporal datasets, covering 83.87% of facts that change over time, named *TempUN* (631k samples) and *TempUN*$_s$ (104k samples). We proposed six MCQ-based evaluations for assessing temporal information on 12 popular LLMs, and introduced three learning paradigms: Continual Learning, Yearwise Finetuning, and Random Finetuning. Our findings highlight that the popular LLMs does not retain the temporal information, and open-source LLMs yielded better results, however fails to acknowledge the lack of knowledge, to which closed-source models admits their missing knowledge.

Future work plans to expand the dataset to explore non-numerical modalities, a broader timespan, and a higher percentage of changing facts, thereby improving the LLMs' temporal reasoning abilities. Additionally, we aim to inspect the numerical-memorization in our future works.

Limitations

Our research emphasizes the limitations of LLMs in comprehending temporal knowledge and their inclination toward language acquisition rather than

[7]The following suffixes were utilized: ***(1)*** *Choose the most relevant answer*, ***(2)*** *Provide the only correct option, without explanation.*

analyzing numerical trends. Our work encompasses historical data spanning from 10,000 BCE to 2100 years ago, comprising approximately **461,506** instances, leading to the creation of **9,474,409** temporal prompts. Due to computational constraints inherent in larger models, our experiments could only be conducted on a subset of the complete dataset, resulting in evaluations being carried out on **1,907** instances, constituting **104,283** samples spanning eight distinct categories in the numerical modality. Our *TempUN* data covers the factual numerical data, and we plan to add the textual data in our future works. Our work focuses on proposing the numerical-temporal dataset for a longer time span, which was missing the previous literature and not significantly contributing to the numerical memorisations in LLMs. Lastly, we plan to explore different fine-tuning strategies, such as adapters, k-adapters, etc., to help the LLMs learn better in future works.

Ethics and Potential Risks

We have strictly adhered to the ethics and guidelines during the progress of our work. The data processing and preparation guidelines have been taken into consideration. The introduced data does NOT contain personal names, uniquely identifiable individuals, or offensive content. The data introduced solely contains the facts as listed on the OWD site.

Acknowledgements

This work is supported by the Prime Minister Research Fellowship (PMRF-1702154) to Himanshu Beniwal. Acknowledgment is extended to Vamsi Srivathsa, Venkata Sriman, and Zeeshan Snehil Bhagat for their invaluable assistance during the experimental phase of this work. Special thanks are also due to Professor Nipun Batra and Zeel Patel for their support in fulfilling the computational requirements. A part of our work was supported by Microsoft's Accelerate Foundation Models Research grant.

References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Qin Cai, Martin Cai, Caio César Teodoro Mendes, Weizhu Chen, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Yen-Chun Chen, Yi-Ling Chen, Parul Chopra, Xiyang Dai, Allie Del Giorno, Gustavo de Rosa, Matthew Dixon, Ronen Eldan, Victor Fragoso, Dan Iter, Mei Gao, Min Gao, Jianfeng Gao, Amit Garg, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Jamie Huynh, Mojan Javaheripi, Xin Jin, Piero Kauffmann, Nikos Karampatziakis, Dongwoo Kim, Mahoud Khademi, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Yunsheng Li, Chen Liang, Lars Liden, Ce Liu, Mengchen Liu, Weishung Liu, Eric Lin, Zeqi Lin, Chong Luo, Piyush Madan, Matt Mazzola, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini, Xin Wang, Lijuan Wang, Chunyu Wang, Yu Wang, Rachel Ward, Guanhua Wang, Philipp Witte, Haiping Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Sonali Yadav, Fan Yang, Jianwei Yang, Ziyi Yang, Yifan Yang, Donghan Yu, Lu Yuan, Chengruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *Preprint*, arXiv:2404.14219.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Oshin Agarwal and Ani Nenkova. 2022. Temporal effects on pre-trained models for language processing tasks. *Transactions of the Association for Computational Linguistics*, 10:904–921.

AI@Meta. 2024. Llama 3 model card.

Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Magdalena Biesialska, Katarzyna Biesialska, and Marta R. Costa-jussà. 2020. Continual lifelong learning in natural language processing: A survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6523–6541, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2023. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*.

Wenhu Chen, Xinyi Wang, and William Yang Wang. 2021. A dataset for answering time-sensitive questions. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *arXiv preprint arXiv:2305.14314*.

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.

Nate Gruver, Marc Finzi, Shikai Qiu, and Andrew Gordon Wilson. 2023. Large language models are zero-shot time series forecasters. *arXiv preprint arXiv:2310.07820*.

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.

Joel Jang, Seonghyeon Ye, Changho Lee, Sohee Yang, Joongbo Shin, Janghoon Han, Gyeonghun Kim, and Minjoon Seo. 2022a. TemporalWiki: A lifelong benchmark for training and evaluating ever-evolving language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6237–6250, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Joel Jang, Seonghyeon Ye, Changho Lee, Sohee Yang, Joongbo Shin, Janghoon Han, Gyeonghun Kim, and Minjoon Seo. 2022b. Temporalwiki: A lifelong benchmark for training and evaluating ever-evolving language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6237–6250.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Timotej Knez and Slavko Žitnik. 2023. Event-centric temporal knowledge graph construction: A survey. *Mathematics*, 11(23):4852.

Xingxuan Li, Liying Cheng, Qingyu Tan, Hwee Tou Ng, Shafiq Joty, and Lidong Bing. 2023. Unlocking temporal question answering for large language models using code execution. *Preprint*, arXiv:2305.15014.

OpenAI. 2022. Introducing chatgpt. `https://openai.com/blog/chatgpt`.

Yifu Qiu, Zheng Zhao, Yftah Ziser, Anna Korhonen, Edoardo M. Ponti, and Shay B. Cohen. 2023. Are large language models temporally grounded? *Preprint*, arXiv:2311.08398.

Guy D. Rosin and Kira Radinsky. 2022. Temporal attention for language models. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1498–1508, Seattle, United States. Association for Computational Linguistics.

Qingyu Tan, Hwee Tou Ng, and Lidong Bing. 2023a. Towards benchmarking and improving the temporal reasoning capability of large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14820–14835, Toronto, Canada. Association for Computational Linguistics.

Qingyu Tan, Hwee Tou Ng, and Lidong Bing. 2023b. Towards robust temporal reasoning of large language models via a multi-hop qa dataset and pseudo-instruction tuning. *arXiv preprint arXiv:2311.09821*.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Yuqing Wang and Yun Zhao. 2023. Tram: Benchmarking temporal reasoning for large language models. *arXiv preprint arXiv:2310.00835*.

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, Singapore. Association for Computational Linguistics.

Siheng Xiong, Ali Payani, Ramana Kompella, and Faramarz Fekri. 2024. Large language models can learn temporal reasoning. *arXiv preprint arXiv:2401.06853*.

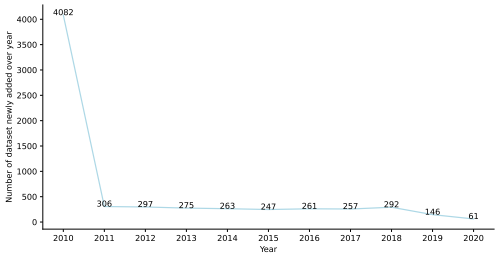

Figure 3: Count of unique data samples available each year.

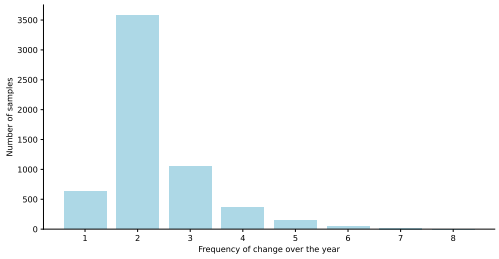

Figure 4: Frequency of change in the dataset for one query over 11 years

Xinli Yu, Zheng Chen, and Yanbin Lu. 2023. Harnessing LLMs for temporal data - a study on explainable financial time series forecasting. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 739–753, Singapore. Association for Computational Linguistics.

Chenhan Yuan, Qianqian Xie, Jimin Huang, and Sophia Ananiadou. 2023. Back to the future: Towards explainable temporal reasoning with large language models. *arXiv preprint arXiv:2310.01074*.

A Appendix

A.1 Inferencing Models - Zeroshot Setting

To assess the model's proficiency in processing numerical data, the identical sample was presented to various models with capacities exceeding 7 billion parameters, as illustrated in Figure 1. It was observed that contemporary, widely used models demonstrated a deficiency in relevant knowledge. This limitation became particularly evident when the prompt was slightly altered to include a temporal shift; the models tended to overestimate and generate responses that were not pertinent to the given context.

A.2 TempLAMA Dataset

TempLAMA We also summarizes the previous available dataset: **TempLAMA** by Dhingra et al. (2022), which is a closed-book question-answering dataset. The dataset consists of events and 11 relations that change over the years. The dataset contains data for 11 years, 2010-2020. In the dataset, *Valentino Rossi plays for _X_.* a query changed only thrice over the year; in 2010, it was *Yamaha Motor Racing* then *Ducati Motor Holding S.p.A.* in 2011 and finally back to *Yamaha Motor Racing* from 2013 onwards. Figure 4 shows how frequently events changed over 11 years. We see that most of the events did not change frequently. In the dataset, each sample contains a *subject* (s), *relation* (r) and *objects* (o) from years where there was a change. The TempLAMA dataset contains the nine different relations (r) that change over time. The list of each relation present and the template for each category in the dataset is available in Table 5. The Number of data samples newly added in each year is depicted in Figure 3.

A.3 Experimental Settings

In this section, we define the experimental configurations used to fine-tune the models.

While experimenting with the fine-tuning strategies, we use the following hyperparameters for all three *close-source* models: batch size (12), epoch (10), learning rate (2e-5), and patience (4), with the search space defined in Table 6.

A.4 Computational Resources

The experiments are carried out on four NVIDIA Tesla V100 32 GB. The estimated cost to cover the computational requirements for two months, computed over GCP is $9,460.78[8] ($4,730.39-per month x 2 months). We utilized the official APIs for all of the *close-source* models.

A.5 TempUN Dataset

This section explains the details linked to the creation of the dataset. As described in Section 3, the data is curated from the *Our World in Data (OWD)* site. The site scraps different trusted sources that are reliable and report accurate numbers. We iteratively parsed the site and processed the raw tabular data in the $< C, I, Y_t >$ and V_t template format. The raw data is then categorized into the United Nations-focused domains and respective subcategories, and number of instances per subcategory

[8]The price for the VM is computed using the GCP Calculator: https://cloud.google.com/products/calculator.

	Wikidata ID	Relation	Template
1	P54	member of sports team	<subject>plays for <object>.
2	P39	position held	<subject>holds the position of <object>.
3	P108	employer	<subject>works for <object>.
4	P102	political party	<subject >is a member of the <object>.
5	P286	head coach	<object>is the head coach of <subject>.
6	P69	educated at	<subject>attended <object>.
7	P488	chairperson	<object>is the chair of <subject>.
8	P6	head of government	<object>is the head of the government of <sub..
9	P127	owned by	<subject>is owned by <object>.

Table 5: TempLAMA relation and the template format of each sample and the corresponding WikiData dataset relation identifier.

Hyperparameter	Search Space
Batch Size	[8, 12, 16]
Epoch	[6 - 10]
Learning Rate	[2e-4, 2e-5, 2e-6]
Patience	[4]

Table 6: The search space for hyperparameters.

in Tables as 15 (Climate), 16 (Food and Agriculture), 17 (Health), 18 (Human Rights), 19 (Innovation and Technological Change), 20 (Migration), 21 (Poverty, Economic Development, and Community), and 22 (Peace and War). We highlight the yearwise count of MCQs in each MCQ-based strategy as: Figure 5 (DB), 6 (CP), 7 (WB), 8 (MM), 9 (RB), and 10 (TB). Overall, we also showcase the category-wise distribution for each strategy in Table 14. Apart from categories 'C4' and 'C5', all of the categories seem to have a higher number of MCQs.

A.6 MCQ-Based Strategy Yearwise Performance - Zero Shot Results

Table 8 highlights the index table for the results for zero-shot inferences over the 12 open and closed source LLMs per category, whereas Table 9 highlights the plots of inferences per metrics. We can observe that all the models (Tables 23 to 94) show that the LLMs produce more incorrect results than the correct results. We showcase the comparative analysis over the DB metric in the Figures 2. We highlight that the larger models (LLMs with >14B parameters) tends to store more information and are better at generating 'information not available', rather than generating the 'incorrect' predictions. We show the combined plots for all six metrics in Figures: `phi-2` (11), `flan-t5-xl` (12), `mistral-instruct` (13), `llama-2` (14), and `gemma-7b-it` (15), `llama-3-8b` (16), `phi-3` (17), `mixtral-8x7b` (18), `llama-3-70B` (19), `gpt-35-turbo` (20), `gpt-4` (21), and `gemini-pro` (22). But `gpt-4` (Figure 19) has shown date-base (DB) and min-max (MM) metric more yielding.

A.7 Continual Learning

We present the following figures for different open-source models, showing the yearwise performance when finetuned in **Continual Learning** paradigm for the "Correct", "Incorrect", and "Information Not Available" labels as indexed in Table 10.

A.8 Yearwise Finetuning

We present the following figures for different open-source models, showing the yearwise performance when finetuned in **Yearwise Finetuning** paradigm as indexed in Table 11.

A.9 Random Finetuning

We present the following figures for different open-source models, showing the yearwise performance when finetuned in **Random Finetuning** paradigm as indexed in Table12.

Categories	Subcategories	Instances	Samples	Instances$_s$	Samples$_s$
C1: Climate	19	95,289	1,778,631	244	17,928
C2: Food and Agriculture	10	33,610	991,443	279	11,133
C3: Health	34	245,330	5,684,312	260	18,599
C4: Human Rights	8	3,132	7,142	190	5,373
C5: Innovation	4	567	1,537	227	5,813
C6: Migration	1	18,167	100,346	255	17,232
C7: Economic Development	25	59,483	909,519	250	18,716
C8: Peace and War	5	7,316	24,572	202	9,336
Total	**106**	**462,894**	**9,497,502**	**1,907**	**104,130**

Table 7: List of categories as global issues and the primary focus required as per the United Nations in the *TempUN* and *TempUN*$_s$ datasets. Here, Instances and Samples underlie the *TempUN* dataset, where Instances$_s$ and Samples$_s$ for the *TempUN*$_s$ dataset.

Model	DB	CP	WB	RB	MM	TB
`phi-2`	23	24	25	26	27	28
`flan-t5-xl`	29	30	31	32	33	34
`mistral-instruct`	35	36	37	38	39	40
`llama-2-chat`	41	42	43	44	45	46
`gemma-7b-it`	47	48	49	50	51	52
`llama-3-8b`	53	54	55	56	57	58
`phi-3-instruct`	59	60	61	62	63	64
`mixtral-8x7b`	65	66	67	68	69	70
`llama-3-70b`	71	72	73	74	75	76
`gpt-3.5-turbo`	77	78	79	80	81	82
`gpt-4`	83	84	85	86	87	88
`gemini-pro`	89	90	91	92	93	94

Table 8: The index table of the category **tables** for the **Zero-shot** evaluations over open and closed source models.

Model	DB	CP	WB	RB	MM	TB
`phi-2`	23	24	25	26	27	28
`flan-t5-xl`	29	30	31	32	33	34
`mistral-instruct`	35	36	37	38	39	40
`llama-2-chat`	41	42	43	44	45	46
`gemma-7b-it`	47	48	49	50	51	52
`llama-3-8b`	53	54	55	56	57	58
`phi-3-instruct`	59	60	61	62	63	64
`mixtral-8x7b`	65	66	67	68	69	70
`llama-3-70b`	71	72	73	74	75	76
`gpt-3.5-turbo`	77	78	79	80	81	82
`gpt-4`	83	84	85	86	87	88
`gemini-pro`	89	90	91	92	93	94

Table 9: The index table of **plots** for the **Zero-shot** evaluations for open-source and closed-source models.

Models	DB	CP	WB	MM	RB	TB
phi-2	95	96	97	98	99	100
flan-t5-xl	101	102	103	104	105	106
mistral-instruct	107	108	109	110	111	112
llama-2-chat	113	114	115	116	117	118
gemma-7b-it	119	120	121	122	123	124
llama-3-8b	125	126	127	128	129	130
phi-3-instruct	131	132	133	134	135	136

Table 10: The index table of plots for the **Continual Learning** evaluations for open-source models.

Models	DB	CP	WB	MM	RB	TB
phi-2	137	138	139	140	141	142
flan-t5-xl	143	144	145	146	147	148
mistral-instruct	149	150	151	152	153	154
lama-2-chat	155	156	157	158	159	160
gemma-7b-it	161	162	163	164	165	166
llama-3-8b	167	168	169	170	171	172
phi-3-instruct	173	174	175	176	177	178

Table 11: The index table of plots for the **Yearwise Finetuning** evaluations for open-source models.

Models	DB	CP	WB	MM	RB	TB
phi-2	179	180	181	182	183	184
flan-t5-xl	185	186	187	188	189	190
mistral-instruct	191	192	193	194	195	196
llama-2-chat	197	198	199	200	201	202
gemma-7b-it	203	204	205	206	207	208
llama-3-8b	209	210	211	212	213	214
phi-3-instruct	215	216	217	218	219	220

Table 12: The index table of plots for the **Random Finetuning** evaluations for open-source models.

Categories	DB	CP	WB	MM	RB	TB
C1: Climate	1,778,631	672,993	603,882	603,882	603,882	603,876
C2: Food and Agriculture	991,443	236,665	213,328	213,328	213,328	213,328
C3: Health	5,684,312	1,891,152	1,739,273	1,739,273	1,739,273	1,739,273
C4: Human Rights	7,142	5,939	1,328	1,328	1,328	1,328
C5: Innovation	1,537	1,247	384	384	384	384
C6: Migration	100,346	100,023	24,116	24,116	24,116	24,116
C7: Economic Development	909,519	402,217	305,373	305,373	305,373	305,373
C8: War	24,572	15,347	11,215	11,215	11,215	11,215
Total	**9,497,502**	**3,325,583**	**2,898,899**	**2,898,899**	**2,898,899**	**2,898,893**

Table 13: The number of samples for each category in the **TempUN** dataset.

Categories	DB	CP	WB	MM	RB	TB
C1: Climate	17,928	2,440	2,440	732	732	732
C2: Food and Agriculture	11,133	2,617	2,495	769	769	769
C3: Health	18,599	2,579	2,570	771	771	771
C4: Human Rights	5,373	1,823	1,778	559	559	559
C5: Innovation	5,813	2,198	2,176	657	657	657
C6: Migration	17,232	2,550	2,550	765	765	765
C7: Economic Development	18,716	2,500	2,500	750	750	750
C8: War	9,336	1,801	1,759	531	531	531
Total	**104,130**	**18,508**	**18,268**	**5,534**	**5,534**	**5,534**

Table 14: The number of prompts for each category for different metrics in the **TempUN**$_s$ dataset.

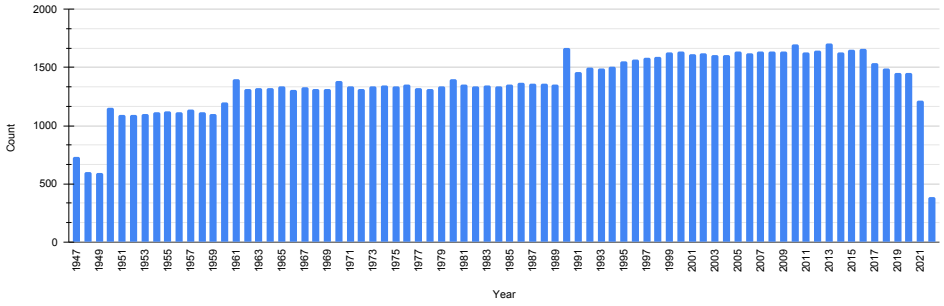

Figure 5: Plot for the number of MCQs in the Date-based metric (DB) per year.

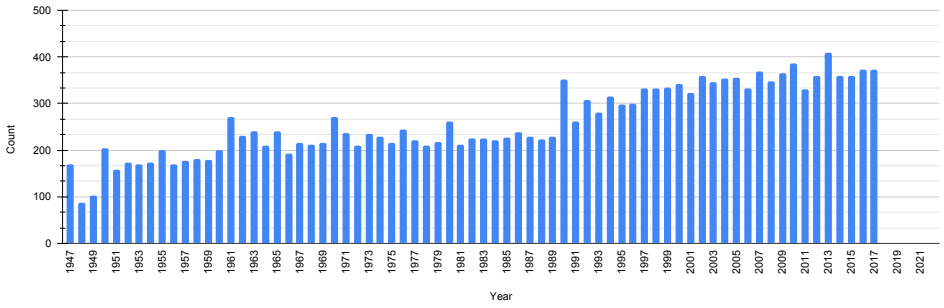

Figure 6: Plot for the number of MCQs in the Comparative-based metric (CP) per year.

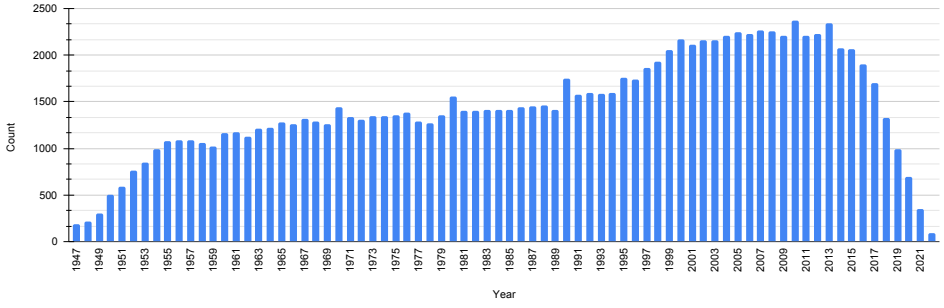

Figure 7: Plot for the number of MCQs in the Window-based metric (WB) per year.

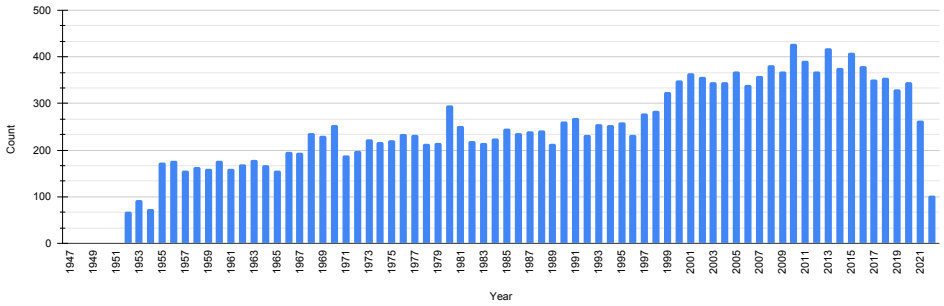

Figure 8: Plot for the number of MCQs in the Min/Max-based metric (MM) per year.

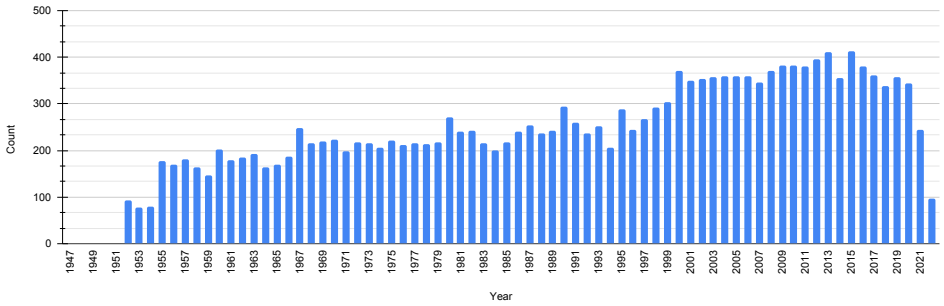

Figure 9: Plot for the number of MCQs in the Range-based metric (RB) per year.

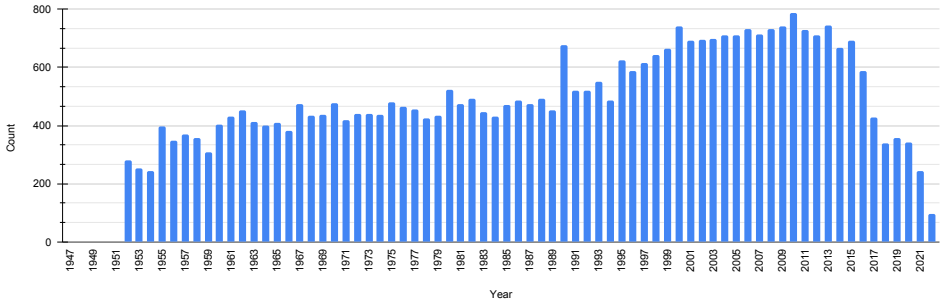

Figure 10: Plot for the number of MCQs in the Trend-based metric (TB) per year.

Categories	# of instances
Access To Energy	2,088
Air Pollution	6,424
Biodiversity	12,108
Clean Water and Sanitization	34,214
Climate Change	251
CO2 and Greenhouse Gas Emissions	30,785
Energy	11,483
Forests and Deforestation	5,030
Fossil Fuels	2,532
Indoor Air Pollution	2,712
Lead Pollution	1,114
Natural Disasters	5,210
Nuclear Energy	730
Oil Spills	30
Ozone Layer	763
Pesticides	1,435
Plastic Pollution	3,008
Pollution	5,827
Water Use and Stress	2,136
Total	**127,880**

Table 15: List of 19 sub-categories in the **Climate** category.

Categories	# of instances
Agricultural Production	9,087
Animal Welfare	3,313
Crop Yields	9,199
Environmental Impacts of Food Production	1,638
Environmental Impacts of Food Production (Food and Agriculture)	1,276
Famines	943
Fertilizers	4,091
Food Prices	2,762
Land Use	12,134
Meat and Dairy Production	12,628
Total	**57,071**

Table 16: List of 10 sub-categories in the **Food and Agriculture** category.

Categories	# of instances
Alcohol Consumption	9,494
Burden of Disease	6,956
Cardiovascular Diseases	9,623
Causes of Death	52,430
Child and Infant Mortality	22,176
Covid	10,744
Diarrheal Diseases	14,312
Diet Compositions	12,825
Eradication of Diseases	7,133
Fertility Rate	2,330
Global Health	3,276
Happiness and Satisfaction	396
Healthcare Spending	3,568
HIV	21,602
Human Height	2,855
Hunger and Undernourishment	4,694
Influenza	11,013
Life Expectancy	18,227
Malaria	4,886
Maternal Mortality	3,936
Mental Health	8,490
Micronutrient Deficiency	2,659
Monkeypox	1,353
Obesity	7,806
Opioids	11,468
Pandemics	4,224
Pneumonia	2,400
Polio	2,021
Sanitation	4,647
Smallpox	273
Smoking	5,429
Suicides	6,374
Tetanus	2,391
Vaccination	10,195
Total	**292,206**

Table 17: List of 34 sub-categories in the **Health** category.

Categories	# of instances
Child Labor	605
Human Rights	10,566
LGBT	647
Literacy	1,375
Loneliness and social connections	478
Marriages and divorces	901
Trust	598
Violence against Children	1,530
Total	**16,700**

Table 18: List of 8 sub-categories in the **Human Rights** category.

Categories	# of instances
Artificial-Intelligence	3,883
Internet	1,991
Research-And-Development	2,973
Technology change	792
Total	**9,639**

Table 19: List of 4 sub-categories in the **Innovation and Technological Change** category.

Categories	# of instances
International Migration and Refugees	36,226
Total	**36,226**

Table 20: List of one major sub-category in the **Migration** category.

Categories	# of instances
Age	3,048
Books	109
Corruption	2,228
Economic-Inequality	7,592
Education-Spending	845
Employment-In-Agriculture	3,382
Gender Ratio	4,783
Global-Education	15,933
Government-Spending	1,393
Homelessness	18
Human Development Index	2,624
Light at Night	12
Poverty	8,969
Renewable Energy	3,322
State-Capacity	4,298
Taxation	1,350
Time use	167
Tourism	3,058
Trade and globalization	7,073
Transportation	717
Urbanization	4,804
Women Employment	3,304
Women Rights	5,573
Working Hours	260
GDP	1
Total	**84,863**

Table 21: List of 25 sub-categories in the **Poverty, Economic Development, and Community** category.

Categories	# of instances
Homicide	16,959
Military spending	1,689
Nuclear-Weapons	75
Terrorism	9,823
War and Peace	30
Total	**28,576**

Table 22: List of 5 sub-categories in the **Peace and War** category.

Category	Correctly Answered	Incorrectly Answered	Not Answered
C1: Climate	7.25	92.75	0.0
C2: Food and Agriculture	12.47	87.51	0.02
C3: Health	9.9	90.1	0.0
C4: Human Rights	13.68	86.32	0.0
C5: Innovation	4.8	95.2	0.0
C6: Migration	15.54	84.46	0.0
C7: Economic Development	11.27	88.73	0.0
C8: Peace and War	10.96	89.04	0
Total	**10.9**	**89.1**	**0.0**

Table 23: Above table provides the percentage of the answers that are correctly, incorrectly, and not answered by `Phi-2` on *DB*

Category	Correctly Answered	Incorrectly Answered	Not Answered
C1: Climate	0.0	100.0	0.0
C2: Food and Agriculture	0.0	99.92	0.08
C3: Health	0.0	100.0	0.0
C4: Human Rights	0.0	97.59	2.41
C5: Innovation	0.0	100.0	0.0
C6: Migration	0.0	100.0	0.0
C7: Economic Development	0.0	85.88	14.12
C8: Peace and War	0	91.23	8.77
Total	**0.0**	**96.99**	**3.01**

Table 24: Above table provides the percentage of the answers that are correctly, incorrectly, and not answered by `Phi-2` on *CP*.

Category	Correctly Answered	Incorrectly Answered	Not Answered
C1: Climate	19.39	80.61	0.0
C2: Food and Agriculture	19.32	80.68	0.0
C3: Health	12.53	87.47	0.0
C4: Human Rights	22.83	77.17	0.0
C5: Innovation	17.92	82.08	0.0
C6: Migration	18.27	81.73	0.0
C7: Economic Development	14.56	85.44	0.0
C8: Peace and War	23.88	76.12	0
Total	**18.19**	**81.81**	**0.0**

Table 25: Above table provides the percentage of the answers that are correctly, incorrectly, and not answered by `Phi-2` on *WB*.

Category	Correctly Answered	Incorrectly Answered	Not Answered
C1: Climate	4.14	95.86	0.0
C2: Food and Agriculture	6.09	93.91	0.0
C3: Health	2.3	97.7	0.0
C4: Human Rights	13.39	86.61	0.0
C5: Innovation	8.41	91.59	0.0
C6: Migration	12.43	87.57	0.0
C7: Economic Development	5.44	93.72	0.84
C8: Peace and War	12.39	87.21	0.4
Total	**7.69**	**92.16**	**0.15**

Table 26: Above table provides the percentage of the answers that are correctly, incorrectly, and not answered by `Phi-2` on *MM*.

Category	Correctly Answered	Incorrectly Answered	Not Answered
C1: Climate	10.7	87.42	1.89
C2: Food and Agriculture	8.5	91.14	0.36
C3: Health	1.28	98.72	0.0
C4: Human Rights	17.89	82.11	0.0
C5: Innovation	6.53	93.47	0.0
C6: Migration	11.88	88.12	0.0
C7: Economic Development	8.08	83.56	8.36
C8: Peace and War	13.42	79.65	6.94
Total	**9.34**	**88.54**	**2.11**

Table 27: Above table provides the percentage of the answers that are correctly, incorrectly, and not answered by `Phi-2` on *RB*.

Category	Correctly Answered	Incorrectly Answered	Not Answered
C1: Climate	1.11	98.89	0.0
C2: Food and Agriculture	2.89	97.11	0.0
C3: Health	11.21	88.79	0.0
C4: Human Rights	2.02	96.01	1.97
C5: Innovation	6.07	92.97	0.97
C6: Migration	11.96	88.04	0.0
C7: Economic Development	5.24	90.28	4.48
C8: Peace and War	3.81	93.63	2.56
Total	**5.79**	**93.04**	**1.17**

Table 28: Above table provides the percentage of the answers that are correctly, incorrectly, and not answered by `Phi-2` on *TB*.

	Correctly Answered	Incorrectly Answered	Not Answered
C1: Climate	28.51	71.49	0
C2: Food and Agriculture	49.65	50.35	0
C3: Health	44.7	55.3	0
C4: Human Rights	44.02	55.98	0
C5: Innovation	24.81	75.19	0
C6: Migration	37.34	62.66	0
C7: Economic Development	43.26	56.74	0
C8: Peace and War	27.88	72.12	0
Total	**38.31**	**61.69**	**0**

Table 29: Above table provides the percentage of the answers that are correctly, incorrectly, and not answered by `flan-t5-xl` on *DB*

	Correctly Answered	Incorrectly Answered	Not Answered
C1: Climate	47.38	52.62	0
C2: Food and Agriculture	52.12	47.88	0
C3: Health	12.6	87.4	0
C4: Human Rights	40.37	59.63	0
C5: Innovation	28.89	71.11	0
C6: Migration	65.65	34.35	0
C7: Economic Development	13.6	86.4	0
C8: Peace and War	67.68	32.32	0
Total	**40.25**	**59.75**	**0**

Table 30: Above table provides the percentage of the answers that are correctly, incorrectly, and not answered by `flan-t5-xl` on *CP*.

	Correctly Answered	Incorrectly Answered	Not Answered
C1: Climate	31.8	68.2	0
C2: Food and Agriculture	13.95	86.05	0
C3: Health	21.79	78.21	0
C4: Human Rights	17.49	82.51	0
C5: Innovation	14.02	85.98	0
C6: Migration	21.06	78.94	0
C7: Economic Development	21.32	78.68	0
C8: Peace and War	29.9	70.1	0
Total	**21.33**	**78.67**	**0**

Table 31: Above table provides the percentage of the answers that are correctly, incorrectly, and not answered by `flan-t5-xl` on *WB*.

	Correctly Answered	Incorrectly Answered	Not Answered
C1: Climate	24.63	75.16	0.2
C2: Food and Agriculture	31.82	67.74	0.44
C3: Health	20.47	77	2.53
C4: Human Rights	25.7	74.3	0
C5: Innovation	26.42	73.58	0
C6: Migration	22.35	77.65	0
C7: Economic Development	26.68	73.32	0
C8: Peace and War	11.65	88.35	0
Total	**24.06**	**75.5**	**0.44**

Table 32: Above table provides the percentage of the answers that are correctly, incorrectly, and not answered by `flan-t5-xl` on *MM*.

	Correctly Answered	Incorrectly Answered	Not Answered
C1: Climate	15.53	84.06	0.41
C2: Food and Agriculture	32.75	66.85	0.4
C3: Health	31.28	64.12	4.59
C4: Human Rights	8.61	91.17	0.22
C5: Innovation	13.65	86.35	0
C6: Migration	9.57	89.84	0.59
C7: Economic Development	33.24	65.96	0.8
C8: Peace and War	4.95	95.05	0
Total	**19.77**	**79.26**	**0.97**

Table 33: Above table provides the percentage of the answers that are correctly, incorrectly, and not answered by `flan-t5-xl` on *RB*.

	Correctly Answered	Incorrectly Answered	Not Answered
C1: Climate	1.52	98.48	0
C2: Food and Agriculture	4.29	95.71	0
C3: Health	1.75	98.25	0
C4: Human Rights	0.22	99.78	0
C5: Innovation	1.24	98.76	0
C6: Migration	0.63	99.37	0
C7: Economic Development	8.52	91.48	0
C8: Peace and War	3.47	96.53	0
Total	**2.79**	**97.21**	**0**

Table 34: Above table provides the percentage of the answers that are correctly, incorrectly, and not answered by `flan-t5-xl` on *TB*.

Category	Correctly Answered	Incorrectly Answered	Not Answered
C1: Climate	40.7	17.56	41.74
C2: Food and Agriculture	23.25	71.5	5.25
C3: Health	41.39	58.35	0.26
C4: Human Rights	28.85	45.26	25.89
C5: Innovation	30.45	68.6	0.95
C6: Migration	36.83	59.99	3.18
C7: Economic Development	45.86	53.99	0.15
C8: Peace and War	31.22	48.37	20.4
Total	**37.21**	**51.22**	**11.57**

Table 35: Above table provides the percentage of the answers that are correctly, incorrectly, and not answered by `Mistral-7B` on *DB*

Category	Correctly Answered	Incorrectly Answered	Not Answered
C1: Climate	47.25	52.75	0.0
C2: Food and Agriculture	52.24	47.76	0.0
C3: Health	15.82	84.18	0.0
C4: Human Rights	61.0	39.0	0.0
C5: Innovation	28.53	71.47	0.0
C6: Migration	65.49	34.51	0.0
C7: Economic Development	13.76	86.24	0.0
C8: Peace and War	70.13	29.87	0
Total	**42.92**	**57.08**	**0.0**

Table 36: Above table provides the percentage of the answers that are correctly, incorrectly, and not answered by `Mistral-7B` on *CP*

Category	Correctly Answered	Incorrectly Answered	Not Answered
C1: Climate	6.84	93.16	0.0
C2: Food and Agriculture	10.74	89.26	0.0
C3: Health	25.49	74.51	0.0
C4: Human Rights	18.11	81.89	0.0
C5: Innovation	19.72	80.28	0.0
C6: Migration	36.35	63.65	0.0
C7: Economic Development	22.16	77.84	0.0
C8: Peace and War	16.37	83.63	0
Total	**19.76**	**80.24**	**0.0**

Table 37: Above table provides the percentage of the answers that are correctly, incorrectly, and not answered by `Mistral-7B` on *WB*

Category	Correctly Answered	Incorrectly Answered	Not Answered
C1: Climate	19.39	9.26	71.35
C2: Food and Agriculture	13.03	85.09	1.88
C3: Health	17.98	81.67	0.35
C4: Human Rights	20.58	52.98	26.43
C5: Innovation	31.43	68.57	0.0
C6: Migration	28.9	69.02	2.08
C7: Economic Development	26.2	70.48	3.32
C8: Peace and War	29.62	68.33	2.05
Total	**23.12**	**63.53**	**13.35**

Table 38: Above table provides the percentage of the answers that are correctly, incorrectly, and not answered by `Mistral-7B` on *MM*

Category	Correctly Answered	Incorrectly Answered	Not Answered
C1: Climate	23.57	76.43	0.0
C2: Food and Agriculture	30.42	69.58	0.0
C3: Health	37.12	62.88	0.0
C4: Human Rights	19.01	80.99	0.0
C5: Innovation	36.12	63.88	0.0
C6: Migration	34.27	65.73	0.0
C7: Economic Development	43.24	56.76	0.0
C8: Peace and War	44.06	55.94	0
Total	**33.62**	**66.38**	**0.0**

Table 39: Above table provides the percentage of the answers that are correctly, incorrectly, and not answered by `Mistral-7B` on *RB*

Category	Correctly Answered	Incorrectly Answered	Not Answered
C1: Climate	22.83	15.61	61.56
C2: Food and Agriculture	9.54	84.45	6.01
C3: Health	1.91	95.06	3.04
C4: Human Rights	7.54	82.56	9.9
C5: Innovation	8.13	89.2	2.67
C6: Migration	5.1	64.08	30.82
C7: Economic Development	2.76	77.92	19.32
C8: Peace and War	3.35	56.57	40.08
Total	**7.73**	**70.71**	**21.56**

Table 40: Above table provides the percentage of the answers that are correctly, incorrectly, and not answered by `Mistral` on *TB*

Category	Correctly Answered	Incorrectly Answered	Not Answered
C1: Climate	27.04	57.96	15.0
C2: Food and Agriculture	14.9	85.02	0.08
C3: Health	18.97	80.93	0.1
C4: Human Rights	25.18	74.61	0.2
C5: Innovation	19.73	78.87	1.39
C6: Migration	22.21	76.83	0.95
C7: Economic Development	19.23	80.75	0.01
C8: Peace and War	18.82	79.88	1.3
Total	**20.86**	**76.17**	**2.97**

Table 41: Above table provides the percentage of the answers that are correctly, incorrectly, and not answered by `llama-2-chat` on *DB*.

Category	Correctly Answered	Incorrectly Answered	Not Answered
C1: Climate	47.09	52.91	0.0
C2: Food and Agriculture	52.54	47.46	0.0
C3: Health	12.76	87.24	0.0
C4: Human Rights	40.43	59.57	0.0
C5: Innovation	68.06	31.94	0.0
C6: Migration	63.69	36.31	0.0
C7: Economic Development	13.72	86.28	0.0
C8: Peace and War	69.52	30.48	0
Total	**44.87**	**55.13**	**0.0**

Table 42: Above table provides the percentage of the answers that are correctly, incorrectly, and not answered by `llama-2-chat` on *CP*.

Category	Correctly Answered	Incorrectly Answered	Not Answered
C1: Climate	23.81	76.19	0.0
C2: Food and Agriculture	16.31	83.69	0.0
C3: Health	24.2	75.8	0.0
C4: Human Rights	23.45	76.55	0.0
C5: Innovation	24.17	75.83	0.0
C6: Migration	15.65	84.35	0.0
C7: Economic Development	23.2	76.8	0.0
C8: Peace and War	23.88	76.12	0
Total	**21.63**	**78.37**	**0.0**

Table 43: Above table provides the percentage of the answers that are correctly, incorrectly, and not answered by `llama-2-chat` on *WB*.

Category	Correctly Answered	Incorrectly Answered	Not Answered
C1: Climate	10.16	68.98	20.86
C2: Food and Agriculture	8.3	91.58	0.12
C3: Health	14.05	85.95	0.0
C4: Human Rights	17.49	80.48	2.02
C5: Innovation	12.82	87.13	0.05
C6: Migration	20.43	72.43	7.14
C7: Economic Development	19.76	80.24	0.0
C8: Peace and War	17.06	82.89	0.06
Total	**14.89**	**81.1**	**4.01**

Table 44: Above table provides the percentage of the answers that are correctly, incorrectly, and not answered by `llama-2-chat` on *MM*.

Category	Correctly Answered	Incorrectly Answered	Not Answered
C1: Climate	19.1	80.9	0.0
C2: Food and Agriculture	13.91	86.09	0.0
C3: Health	17.82	82.18	0.0
C4: Human Rights	31.66	68.34	0.0
C5: Innovation	25.69	74.31	0.0
C6: Migration	21.29	78.71	0.0
C7: Economic Development	24.32	75.68	0.0
C8: Peace and War	21.77	78.23	0
Total	**21.5**	**78.5**	**0.0**

Table 45: Above table provides the percentage of the answers that are correctly, incorrectly, and not answered by `llama-2-chat` on *RB*.

Category	Correctly Answered	Incorrectly Answered	Not Answered
C1: Climate	2.17	93.2	4.63
C2: Food and Agriculture	0.64	97.8	1.56
C3: Health	4.16	95.41	0.43
C4: Human Rights	5.46	92.63	1.91
C5: Innovation	4.0	95.36	0.64
C6: Migration	9.02	87.45	3.53
C7: Economic Development	8.28	89.32	2.4
C8: Peace and War	5.34	90.51	4.15
Total	**4.88**	**92.75**	**2.38**

Table 46: Above table provides the percentage of the answers that are correctly, incorrectly, and not answered by `llama-2-chat` on *TB*.

Category	Correctly Answered	Incorrectly Answered	Not Answered
C1: Climate	37.43	62.57	0.01
C2: Food and Agriculture	9.84	90.16	0
C3: Health	19.51	80.49	0
C4: Human Rights	21.46	78.54	0
C5: Innovation	7.91	92.09	0
C6: Migration	25.37	72.64	2
C7: Economic Development	11.2	78.18	10.61
C8: Peace and War	23.61	76.39	0
Total	**20.86**	**76.9**	**2.24**

Table 47: Above table provides the percentage of the answers that are correctly, incorrectly, and not answered by `gemma-7b-it` on *DB*.

Category	Correctly Answered	Incorrectly Answered	Not Answered
C1: Climate	47.05	52.95	0
C2: Food and Agriculture	13.84	86.16	0
C3: Health	52.43	47.57	0
C4: Human Rights	11.75	88.25	0
C5: Innovation	40.43	59.57	0
C6: Migration	49.82	50.18	0
C7: Economic Development	65.65	34.35	0
C8: Peace and War	64.13	35.87	0
Total	**42.31**	**57.69**	**0**

Table 48: Above table provides the percentage of the answers that are correctly, incorrectly, and not answered by `gemma-7b-it` on *CB*.

Category	Correctly Answered	Incorrectly Answered	Not Answered
C1: Climate	10	90	0
C2: Food and Agriculture	7.09	92.91	0
C3: Health	14.94	85.06	0
C4: Human Rights	24.8	75.2	0
C5: Innovation	13.37	86.63	0
C6: Migration	16.08	83.92	0
C7: Economic Development	23.08	76.92	0
C8: Peace and War	15.86	84.14	0
Total	**15.34**	**84.66**	**0**

Table 49: Above table provides the percentage of the answers that are correctly, incorrectly, and not answered by `gemma-7b-it` on *WB*.

Category	Correctly Answered	Incorrectly Answered	Not Answered
C1: Climate	9.39	90.45	0.16
C2: Food and Agriculture	14.51	85.45	0.04
C3: Health	9.22	90.78	0
C4: Human Rights	10.12	89.88	0
C5: Innovation	7.54	92.46	0
C6: Migration	15.61	84.39	0
C7: Economic Development	9.8	90.2	0
C8: Peace and War	18.14	81.86	0
Total	**11.68**	**88.29**	**0.03**

Table 50: Above table provides the percentage of the answers that are correctly, incorrectly, and not answered by `gemma-7b-it` on *MM*.

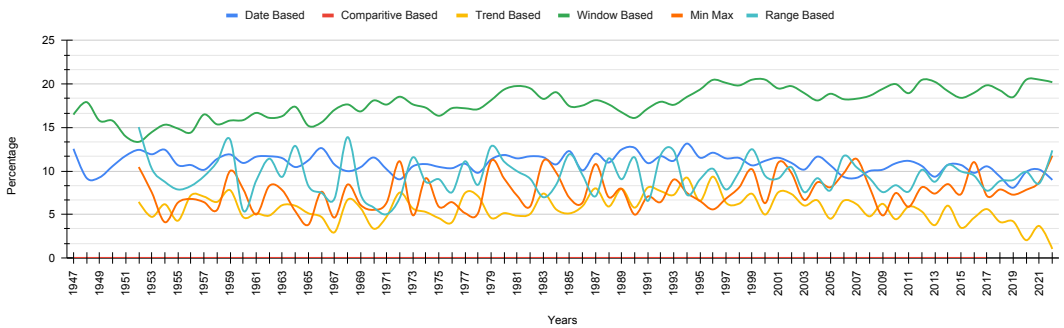

Figure 11: **Zeroshot** MCQ-based evaluation on `phi-2`.

Category	Correctly Answered	Incorrectly Answered	Not Answered
C1: Climate	24.02	75.82	0.16
C2: Food and Agriculture	7.45	92.55	0
C3: Health	14.32	85.68	0
C4: Human Rights	10.46	89.54	0
C5: Innovation	8.55	91.45	0
C6: Migration	16.43	83.57	0
C7: Economic Development	11.72	88.28	0
C8: Peace and War	19.61	80.33	0.06
Total	**14.06**	**85.91**	**0.03**

Table 51: Above table provides the percentage of the answers that are correctly, incorrectly, and not answered by `gemma-7b-it` on *RB*.

Category	Correctly Answered	Incorrectly Answered	Not Answered
C1: Climate	3.03	76.97	20
C2: Food and Agriculture	0.24	99.76	0
C3: Health	3.15	96.85	0
C4: Human Rights	3.71	95.05	1.24
C5: Innovation	1.88	98.12	0
C6: Migration	6.2	93.8	0
C7: Economic Development	1.28	98.72	0
C8: Peace and War	4.83	95.17	0
Total	**2.97**	**94.24**	**2.79**

Table 52: Above table provides the percentage of the answers that are correctly, incorrectly, and not answered by `gemma-7b-it` on *TB*.

Category	Correctly Answered	Incorrectly Answered	Not Answered
C1: Climate	56.29	42.16	1.56
C2: Food and Agriculture	49.74	50.11	0.14
C3: Health	26.14	73.84	0.02
C4: Human Rights	38.95	61.05	0
C5: Innovation	20.61	79.39	0
C6: Migration	42.09	57.82	0.09
C7: Economic Development	29.94	66.71	3.35
C8: Peace and War	37.37	62.63	0
Total	**38.54**	**60.56**	**0.9**

Table 53: Above table provides the percentage of the answers that are correctly, incorrectly, and not answered by `llama-3-8b-8192` on *DB*.

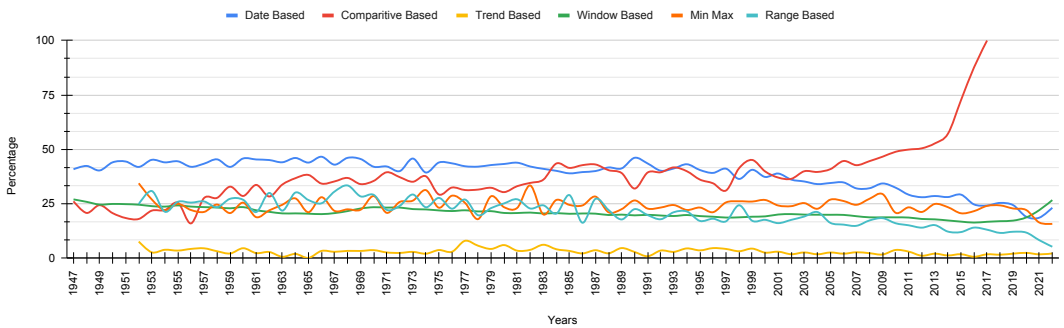

Figure 12: **Zeroshot** MCQ-based evaluation on `flan-t5-xl`.

Category	Correctly Answered	Incorrectly Answered	Not Answered
C1: Climate	46.39	53.57	0.04
C2: Food and Agriculture	15.12	84.88	0
C3: Health	52.08	47.92	0
C4: Human Rights	12.95	87.05	0
C5: Innovation	40.43	59.57	0
C6: Migration	28.84	71.16	0
C7: Economic Development	65.37	34.63	0
C8: Peace and War	50.58	49.42	0
Total	**38.66**	**61.33**	**0.01**

Table 54: Above table provides the percentage of the answers that are correctly, incorrectly, and not answered by `llama-3-8b-8192` on *CB*.

Category	Correctly Answered	Incorrectly Answered	Not Answered
C1: Climate	22.7	77.3	0
C2: Food and Agriculture	6.93	93.07	0
C3: Health	12.57	87.43	0
C4: Human Rights	19.85	80.15	0
C5: Innovation	15.58	84.42	0
C6: Migration	23.37	76.63	0
C7: Economic Development	23.88	76.12	0
C8: Peace and War	28.77	71.23	0
Total	**18.84**	**81.16**	**0**

Table 55: Above table provides the percentage of the answers that are correctly, incorrectly, and not answered by `llama-3-8b-8192` on *WB*.

Category	Correctly Answered	Incorrectly Answered	Not Answered
C1: Climate	25.61	74.34	0.04
C2: Food and Agriculture	14.87	85.13	0
C3: Health	13.39	86.61	0
C4: Human Rights	16.82	83.18	0
C5: Innovation	17.28	82.72	0
C6: Migration	24.2	75.8	0
C7: Economic Development	16.96	83.04	0
C8: Peace and War	17.28	82.72	0
Total	**18.39**	**81.6**	**0.01**

Table 56: Above table provides the percentage of the answers that are correctly, incorrectly, and not answered by `llama-3-8b-8192` on *MM*.

Category	Correctly Answered	Incorrectly Answered	Not Answered
C1: Climate	34.88	65.12	0
C2: Food and Agriculture	29.62	70.38	0
C3: Health	16.69	83.31	0
C4: Human Rights	25.82	74.18	0
C5: Innovation	17.1	82.9	0
C6: Migration	26.39	73.57	0.04
C7: Economic Development	21.08	78.92	0
C8: Peace and War	22.74	77.26	0
Total	**24.36**	**75.63**	**0.01**

Table 57: Above table provides the percentage of the answers that are correctly, incorrectly, and not answered by `llama-3-8b-8192` on *RB*.

Category	Correctly Answered	Incorrectly Answered	Not Answered
C1: Climate	9.96	90.04	0
C2: Food and Agriculture	5.37	94.63	0
C3: Health	8.6	91.4	0
C4: Human Rights	2.31	97.69	0
C5: Innovation	5.93	94.07	0
C6: Migration	7.49	92.51	0
C7: Economic Development	10.68	89.32	0
C8: Peace and War	4.89	95.11	0
Total	**7.18**	**92.82**	**0**

Table 58: Above table provides the percentage of the answers that are correctly, incorrectly, and not answered by `llama-3-8b-8192` on *TB*.

	Correctly Answered	Incorrectly Answered	Not Answered
C1: Climate	11.83	5.99	82.19
C2: Food and Agriculture	5.08	19.55	75.37
C3: Health	11.03	25.08	63.89
C4: Human Rights	10.35	23.79	65.87
C5: Innovation	17.31	45.14	37.55
C6: Migration	0.04	0.1	99.86
C7: Economic Development	6.32	13.58	80.1
C8: Peace and War	24.27	29.57	46.15
Total	**9.37**	**16.46**	**74.18**

Table 59: Above table provides the percentage of the answers that are correctly, incorrectly, and not answered by `phi-3-instruct` on *DB*.

	Correctly Answered	Incorrectly Answered	Not Answered
C1: Climate	48.07	49.96	1.97
C2: Food and Agriculture	53.57	45.51	0.92
C3: Health	49.71	48.7	1.59
C4: Human Rights	35.71	40.7	23.59
C5: Innovation	34.21	64.06	1.73
C6: Migration	63.49	30	6.51
C7: Economic Development	52.2	46.4	1.4
C8: Peace and War	45.47	49.36	5.16
Total	**48.64**	**46.63**	**4.73**

Table 60: Above table provides the percentage of the answers that are correctly, incorrectly, and not answered by `phi-3-instruct` on *CB*.

	Correctly Answered	Incorrectly Answered	Not Answered
C1: Climate	38.77	13.2	48.03
C2: Food and Agriculture	17.11	44.57	38.32
C3: Health	47.74	47.67	4.59
C4: Human Rights	39.93	17.38	42.69
C5: Innovation	22.47	37.78	39.75
C6: Migration	17.92	11.45	70.63
C7: Economic Development	51.64	41.56	6.8
C8: Peace and War	64.18	26.49	9.32
Total	**36.54**	**30.58**	**32.87**

Table 61: Above table provides the percentage of the answers that are correctly, incorrectly, and not answered by `phi-3-instruct` on *WB*.

	Correctly Answered	Incorrectly Answered	Not Answered
C1: Climate	2.83	3.4	93.77
C2: Food and Agriculture	6.97	22.73	70.3
C3: Health	22.68	50.51	26.81
C4: Human Rights	3.37	10.91	85.71
C5: Innovation	18.93	68.75	12.32
C6: Migration	0	0	100
C7: Economic Development	12.08	21.92	66
C8: Peace and War	14.38	44.8	40.82
Total	**10.14**	**27.23**	**62.63**

Table 62: Above table provides the percentage of the answers that are correctly, incorrectly, and not answered by `phi-3-instruct` on *MM*.

	Correctly Answered	Incorrectly Answered	Not Answered
C1: Climate	1.02	3.85	95.12
C2: Food and Agriculture	0.08	0.24	99.68
C3: Health	0	0.08	99.92
C4: Human Rights	0	1.46	98.54
C5: Innovation	0.05	0.55	99.4
C6: Migration	0	0	100
C7: Economic Development	5.48	16.92	77.6
C8: Peace and War	0.63	0.91	98.47
Total	**0.96**	**3.17**	**95.87**

Table 63: Above table provides the percentage of the answers that are correctly, incorrectly, and not answered by `phi-3-instruct` on *RB*.

	Correctly Answered	Incorrectly Answered	Not Answered
C1: Climate	0.49	20.78	78.73
C2: Food and Agriculture	0.2	24.61	75.19
C3: Health	1.09	73.19	25.72
C4: Human Rights	0	8.32	91.68
C5: Innovation	0.41	90.35	9.24
C6: Migration	2.08	70.31	27.61
C7: Economic Development	0.84	75.92	23.24
C8: Peace and War	2.67	52.7	44.63
Total	**0.96**	**53.28**	**45.76**

Table 64: Above table provides the percentage of the answers that are correctly, incorrectly, and not answered by `phi-3-instruct` on *TB*.

Category	Correctly Answered	Incorrectly Answered	Not Answered
C1: Climate	43.98	55.79	0.22
C2: Food and Agriculture	21.75	76.7	1.55
C3: Health	24.58	74.39	1.03
C4: Human Rights	26.06	73.42	0.52
C5: Innovation	38.04	38.67	23.28
C6: Migration	29.53	70.28	0.19
C7: Economic Development	25.63	74.15	0.22
C8: Peace and War	53.53	17.61	28.87
Total	**32.57**	**60.71**	**6.73**

Table 65: Above table provides the percentage of the answers that are correctly, incorrectly, and not answered by `mixtral-8x7b-32768` on *DB*.

Category	Correctly Answered	Incorrectly Answered	Not Answered
C1: Climate	44.51	54.92	0.57
C2: Food and Agriculture	14.16	85.72	0.12
C3: Health	47.12	51.51	1.38
C4: Human Rights	12.41	87.59	0
C5: Innovation	32.14	52.83	15.03
C6: Migration	51.19	48.81	0
C7: Economic Development	56.12	41.69	2.2
C8: Peace and War	38.2	60.41	1.39
Total	**37.59**	**60.31**	**2.09**

Table 66: Above table provides the percentage of the answers that are correctly, incorrectly, and not answered by `mixtral-8x7b-32768` on *CB*.

Category	Correctly Answered	Incorrectly Answered	Not Answered
C1: Climate	68.11	30.66	1.23
C2: Food and Agriculture	4.21	95.75	0.04
C3: Health	25.45	74.55	0
C4: Human Rights	29.98	69.29	0.73
C5: Innovation	32.03	67.97	0
C6: Migration	22.31	77.69	0
C7: Economic Development	20.48	79.48	0.04
C8: Peace and War	30.59	69.36	0.06
Total	**28.85**	**70.9**	**0.25**

Table 67: Above table provides the percentage of the answers that are correctly, incorrectly, and not answered by `mixtral-8x7b-32768` on *WB*.

Category	Correctly Answered	Incorrectly Answered	Not Answered
C1: Climate	55.25	44.59	0.16
C2: Food and Agriculture	5.25	94.71	0.04
C3: Health	13.39	86.61	0
C4: Human Rights	22.38	77.62	0
C5: Innovation	14.71	85.29	0
C6: Migration	12.78	87.22	0
C7: Economic Development	5.36	94.64	0
C8: Peace and War	13.38	86.5	0.12
Total	**17.74**	**82.22**	**0.04**

Table 68: Above table provides the percentage of the answers that are correctly, incorrectly, and not answered by `mixtral-8x7b-32768` on *MM*.

Category	Correctly Answered	Incorrectly Answered	Not Answered
C1: Climate	69.55	29.34	1.11
C2: Food and Agriculture	11.66	88.3	0.04
C3: Health	29.11	70.89	0
C4: Human Rights	52.59	46.57	0.84
C5: Innovation	20.54	79.46	0
C6: Migration	13.12	86.88	0
C7: Economic Development	10.88	89.08	0.04
C8: Peace and War	30.24	69.76	0
Total	**28.77**	**70.99**	**0.24**

Table 69: Above table provides the percentage of the answers that are correctly, incorrectly, and not answered by `mixtral-8x7b-32768` on *RB*.

Category	Correctly Answered	Incorrectly Answered	Not Answered
C1: Climate	13.73	67.91	18.36
C2: Food and Agriculture	0.2	99.48	0.32
C3: Health	0.08	99.92	0
C4: Human Rights	1.74	92.69	5.57
C5: Innovation	0.09	99.59	0.32
C6: Migration	3.02	96.9	0.08
C7: Economic Development	0.6	99.16	0.24
C8: Peace and War	4.04	95.91	0.06
Total	**2.95**	**93.93**	**3.13**

Table 70: Above table provides the percentage of the answers that are correctly, incorrectly, and not answered by `mixtral-8x7b-32768` on *TB*.

Category	Correctly Answered	Incorrectly Answered	Not Answered
C1: Climate	76.55	23.43	0.02
C2: Food and Agriculture	16.78	83.21	0.01
C3: Health	41.23	58.77	0
C4: Human Rights	37.58	62.42	0
C5: Innovation	35.39	64.6	0.02
C6: Migration	35.18	64.82	0.01
C7: Economic Development	26.38	73.62	0
C8: Peace and War	40.42	59.57	0.01
Total	**40.44**	**59.55**	**0.01**

Table 71: Above table provides the percentage of the answers that are correctly, incorrectly, and not answered by `llama-3-70b-8192` on *DB*.

Category	Correctly Answered	Incorrectly Answered	Not Answered
C1: Climate	44.55	55.45	0
C2: Food and Agriculture	47.42	52.58	0
C3: Health	14.89	85.11	0
C4: Human Rights	58.53	41.47	0
C5: Innovation	27.98	72.02	0
C6: Migration	35.29	64.71	0
C7: Economic Development	35.92	64.08	0
C8: Peace and War	33.76	66.24	0
Total	**36.74**	**63.26**	**0**

Table 72: Above table provides the percentage of the answers that are correctly, incorrectly, and not answered by `llama-3-70b-8192` on *CB*.

Category	Correctly Answered	Incorrectly Answered	Not Answered
C1: Climate	76.64	21.31	2.05
C2: Food and Agriculture	29.98	70.02	0
C3: Health	64.05	35.95	0
C4: Human Rights	46.63	53.37	0
C5: Innovation	56.62	43.38	0
C6: Migration	47.41	52.59	0
C7: Economic Development	66.68	33.32	0
C8: Peace and War	48.95	51.05	0
Total	**55.08**	**44.65**	**0.27**

Table 73: Above table provides the percentage of the answers that are correctly, incorrectly, and not answered by `llama-3-70b-8192` on *WB*.

Category	Correctly Answered	Incorrectly Answered	Not Answered
C1: Climate	73.69	26.31	0
C2: Food and Agriculture	20.2	79.8	0
C3: Health	38.83	61.17	0
C4: Human Rights	24.69	75.31	0
C5: Innovation	40.58	59.42	0
C6: Migration	29.57	70.43	0
C7: Economic Development	30.08	69.92	0
C8: Peace and War	38.37	61.63	0
Total	**37.24**	**62.76**	**0**

Table 74: Above table provides the percentage of the answers that are correctly, incorrectly, and not answered by `llama-3-70b-8192` on *MM*.

Category	Correctly Answered	Incorrectly Answered	Not Answered
C1: Climate	72.91	27.09	0
C2: Food and Agriculture	19.84	80.16	0
C3: Health	39.61	60.39	0
C4: Human Rights	25.32	74.68	0
C5: Innovation	32.4	67.6	0
C6: Migration	33.57	66.43	0
C7: Economic Development	28.72	71.28	0
C8: Peace and War	42.01	57.99	0
Total	**36.63**	**63.37**	**0**

Table 75: Above table provides the percentage of the answers that are correctly, incorrectly, and not answered by `llama-3-70b-8192` on *RB*.

Category	Correctly Answered	Incorrectly Answered	Not Answered
C1: Climate	6.11	93.89	0
C2: Food and Agriculture	0.12	99.88	0
C3: Health	0.23	99.77	0
C4: Human Rights	0	100	0
C5: Innovation	0.37	99.63	0
C6: Migration	0.55	99.45	0
C7: Economic Development	0.16	99.84	0
C8: Peace and War	1.08	98.92	0
Total	**1.11**	**98.89**	**0**

Table 76: Above table provides the percentage of the answers that are correctly, incorrectly, and not answered by `llama-3-70b-8192` on *TB*.

Category	Correctly Answered	Incorrectly Answered	Not Answered
C1: Climate	70.81	28.12	1.07
C2: Food and Agriculture	28.21	71.31	0.48
C3: Health	28.8	70.92	0.27
C4: Human Rights	24.9	71.3	3.8
C5: Innovation	25.34	74.47	0.19
C6: Migration	7.56	92.17	0.27
C7: Economic Development	8.63	86.38	4.99
C8: Peace and War	12.96	86.46	0.58
Total	**27.02**	**71.5**	**1.48**

Table 77: Above table provides the percentage of the answers that are correctly, incorrectly, and not answered by `gpt-3.5` on *DB*.

Category	Correctly Answered	Incorrectly Answered	Not Answered
C1: Climate	47.09	52.91	0
C2: Food and Agriculture	52.43	47.57	0
C3: Health	11.98	88.02	0
C4: Human Rights	40.43	59.57	0
C5: Innovation	30.85	69.15	0
C6: Migration	66.27	33.73	0
C7: Economic Development	15.2	84.8	0
C8: Peace and War	52.25	47.75	0
Total	**39.2**	**60.8**	**0**

Table 78: Above table provides the percentage of the answers that are correctly, incorrectly, and not answered by `gpt-3.5` on *CP*.

Category	Correctly Answered	Incorrectly Answered	Not Answered
C1: Climate	21.35	78.65	0
C2: Food and Agriculture	1.32	98.68	0
C3: Health	15.99	84.01	0
C4: Human Rights	30.03	69.97	0
C5: Innovation	13.01	86.99	0
C6: Migration	14.71	85.29	0
C7: Economic Development	13	87	0
C8: Peace and War	24.79	75.21	0
Total	**15.97**	**84.03**	**0**

Table 79: Above table provides the percentage of the answers that are correctly, incorrectly, and not answered by `gpt-3.5` on *WB*.

Category	Correctly Answered	Incorrectly Answered	Not Answered
C1: Climate	42.66	57.34	0
C2: Food and Agriculture	7.7	92.3	0
C3: Health	18.91	81.09	0
C4: Human Rights	11.92	88.08	0
C5: Innovation	22.75	77.21	0.05
C6: Migration	7.92	92.08	0
C7: Economic Development	20.92	79.04	0.04
C8: Peace and War	16.03	83.97	0
Total	**18.79**	**81.2**	**0.01**

Table 80: Above table provides the percentage of the answers that are correctly, incorrectly, and not answered by `gpt-3.5` on *MM*.

Category	Correctly Answered	Incorrectly Answered	Not Answered
C1: Climate	17.25	82.25	0.49
C2: Food and Agriculture	9.24	89.72	1.04
C3: Health	13.31	85.53	1.17
C4: Human Rights	10.63	89.37	0
C5: Innovation	12.13	87.42	0.45
C6: Migration	7.45	91.69	0.86
C7: Economic Development	15.96	83.88	0.16
C8: Peace and War	7.62	92.33	0.06
Total	**11.56**	**87.81**	**0.63**

Table 81: Above table provides the percentage of the answers that are correctly, incorrectly, and not answered by `gpt-3.5` on *RB*.

Category	Correctly Answered	Incorrectly Answered	Not Answered
C1: Climate	0	100	0
C2: Food and Agriculture	0	100	0
C3: Health	0	99.61	0.39
C4: Human Rights	0	100	0
C5: Innovation	0.05	94.58	5.38
C6: Migration	0	99.96	0.04
C7: Economic Development	0.04	97.4	2.56
C8: Peace and War	0.11	99.83	0.06
Total	**0.02**	**98.92**	**1.06**

Table 82: Above table provides the percentage of the answers that are correctly, incorrectly, and not answered by `gpt-3.5` on *TB*.

Category	Correctly Answered	Incorrectly Answered	Not Answered
C1: Climate	30.35	5.56	64.09
C2: Food and Agriculture	14.37	38.8	46.82
C3: Health	54.22	39.24	6.53
C4: Human Rights	4.99	4.91	90.1
C5: Innovation	33.05	65.8	1.15
C6: Migration	25.62	40.48	33.9
C7: Economic Development	26.27	60.41	13.33
C8: Peace and War	17.76	17.44	64.8
Total	**29.1**	**35.16**	**35.74**

Table 83: Above table provides the percentage of the answers that are correctly, incorrectly, and not answered by `gpt-4` on *DB*.

Category	Correctly Answered	Incorrectly Answered	Not Answered
C1: Climate	0.37	99.63	0
C2: Food and Agriculture	0.11	99.89	0
C3: Health	5.16	94.84	0
C4: Human Rights	0.22	99.78	0
C5: Innovation	9.87	90.13	0
C6: Migration	2.51	97.49	0
C7: Economic Development	0.52	99.48	0
C8: Peace and War	0.67	99.33	0
Total	**2.46**	**97.54**	**0**

Table 84: Above table provides the percentage of the answers that are correctly, incorrectly, and not answered by `gpt-4` on *CP*.

Category	Correctly Answered	Incorrectly Answered	Not Answered
C1: Climate	0	100	0
C2: Food and Agriculture	0.08	99.92	0
C3: Health	0	100	0
C4: Human Rights	0	100	0
C5: Innovation	0.14	99.86	0
C6: Migration	0	100	0
C7: Economic Development	0.08	99.92	0
C8: Peace and War	0.06	99.94	0
Total	**0.04**	**99.96**	**0**

Table 85: Above table provides the percentage of the answers that are correctly, incorrectly, and not answered by `gpt-4` on *WB*.

Category	Correctly Answered	Incorrectly Answered	Not Answered
C1: Climate	44.63	8.81	46.56
C2: Food and Agriculture	15.39	75.99	8.62
C3: Health	56.85	41.09	2.06
C4: Human Rights	22.95	31.44	45.61
C5: Innovation	31.34	68.24	0.41
C6: Migration	20.31	54.31	25.37
C7: Economic Development	15.92	80.88	3.2
C8: Peace and War	17.51	31.04	51.45
Total	**28.73**	**50.16**	**21.11**

Table 86: Above table provides the percentage of the answers that are correctly, incorrectly, and not answered by `gpt-4` on *MM*.

Category	Correctly Answered	Incorrectly Answered	Not Answered
C1: Climate	0	100	0
C2: Food and Agriculture	0	100	0
C3: Health	0	100	0
C4: Human Rights	0	100	0
C5: Innovation	0	100	0
C6: Migration	0	100	0
C7: Economic Development	0	100	0
C8: Peace and War	0	100	0
Total	**0**	**100**	**0**

Table 87: Above table provides the percentage of the answers that are correctly, incorrectly, and not answered by `gpt-4` on *RB*.

Category	Correctly Answered	Incorrectly Answered	Not Answered
C1: Climate	3.24	0.82	95.94
C2: Food and Agriculture	0	0.16	99.84
C3: Health	0.54	12.57	86.89
C4: Human Rights	0.06	0.06	99.89
C5: Innovation	6.3	70.77	22.93
C6: Migration	0.55	5.88	93.57
C7: Economic Development	0.72	5.12	94.16
C8: Peace and War	0.06	0.57	99.37
Total	**1.45**	**11.91**	**86.64**

Table 88: Above table provides the percentage of the answers that are correctly, incorrectly, and not answered by `gpt-4` on *TB*.

Category	Correctly Answered	Incorrectly Answered	Not Answered
C1: Climate	70.05	29.95	0
C2: Food and Agriculture	3.34	96.66	0
C3: Health	27.36	72.64	0
C4: Human Rights	42.84	57.16	0
C5: Innovation	19.96	80.04	0
C6: Migration	18.72	81.28	0
C7: Economic Development	16.39	83.61	0
C8: Peace and War	29.48	70.52	0
Total	**29.31**	**70.69**	**0**

Table 89: Above table provides the percentage of the answers that are correctly, incorrectly, and not answered by `gemini-pro` on *DB*.

Category	Correctly Answered	Incorrectly Answered	Not Answered
C1: Climate	43.73	56.27	0
C2: Food and Agriculture	53.88	46.12	0
C3: Health	12.45	87.55	0
C4: Human Rights	46.02	53.98	0
C5: Innovation	28.53	71.47	0
C6: Migration	62.67	37.33	0
C7: Economic Development	15.04	84.96	0
C8: Peace and War	46.09	53.91	0
Total	**38.19**	**61.81**	**0**

Table 90: Above table provides the percentage of the answers that are correctly, incorrectly, and not answered by `gemini-pro` on *CP*.

Category	Correctly Answered	Incorrectly Answered	Not Answered
C1: Climate	63.2	36.8	0
C2: Food and Agriculture	15.87	84.13	0
C3: Health	29.96	70.04	0
C4: Human Rights	29.7	70.3	0
C5: Innovation	33.69	66.31	0
C6: Migration	26.04	73.96	0
C7: Economic Development	42.36	57.64	0
C8: Peace and War	29.11	70.89	0
Total	**33.96**	**66.04**	**0**

Table 91: Above table provides the percentage of the answers that are correctly, incorrectly, and not answered by `gemini-pro` on *WB*.

Category	Correctly Answered	Incorrectly Answered	Not Answered
C1: Climate	40.82	59.18	0
C2: Food and Agriculture	6.01	93.99	0
C3: Health	19.49	79.53	0.97
C4: Human Rights	11.25	87.57	1.18
C5: Innovation	18.98	80.93	0.09
C6: Migration	12.24	87.76	0
C7: Economic Development	13.72	86.28	0
C8: Peace and War	8.75	90.22	1.02
Total	**16.8**	**82.84**	**0.36**

Table 92: Above table provides the percentage of the answers that are correctly, incorrectly, and not answered by `gemini-pro` on *MM*.

Category	Correctly Answered	Incorrectly Answered	Not Answered
C1: Climate	0	99.96	0.04
C2: Food and Agriculture	0	100	0
C3: Health	0	100	0
C4: Human Rights	0	99.55	0.45
C5: Innovation	0	99.95	0.05
C6: Migration	0	100	0
C7: Economic Development	0	94.24	5.76
C8: Peace and War	0	99.89	0.11
Total	**0**	**99.15**	**0.85**

Table 93: Above table provides the percentage of the answers that are correctly, incorrectly, and not answered by `gemini-pro` on *RB*.

Category	Correctly Answered	Incorrectly Answered	Not Answered
C1: Climate	0	100	0
C2: Food and Agriculture	0	100	0
C3: Health	0.16	99.84	0
C4: Human Rights	0.11	99.89	0
C5: Innovation	0.05	99.95	0
C6: Migration	0.16	99.84	0
C7: Economic Development	0.16	99.84	0
C8: Peace and War	0.11	99.89	0
Total	**0.09**	**99.91**	**0**

Table 94: Above table provides the percentage of the answers that are correctly, incorrectly, and not answered by `gemini-pro` on *TB*.

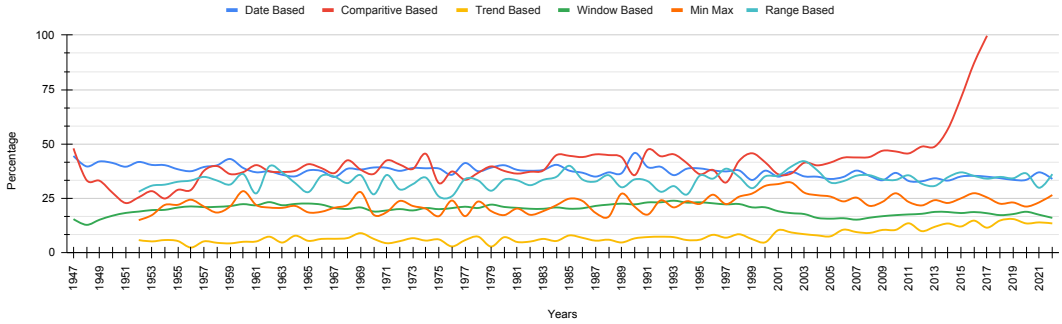

Figure 13: **Zeroshot** MCQ-based evaluation on `mistral-instruct`.

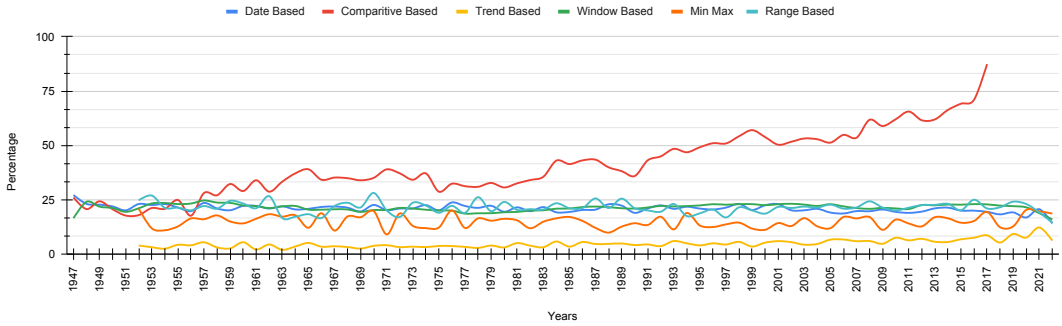

Figure 14: **Zeroshot** MCQ-based evaluation on `llama-2`.

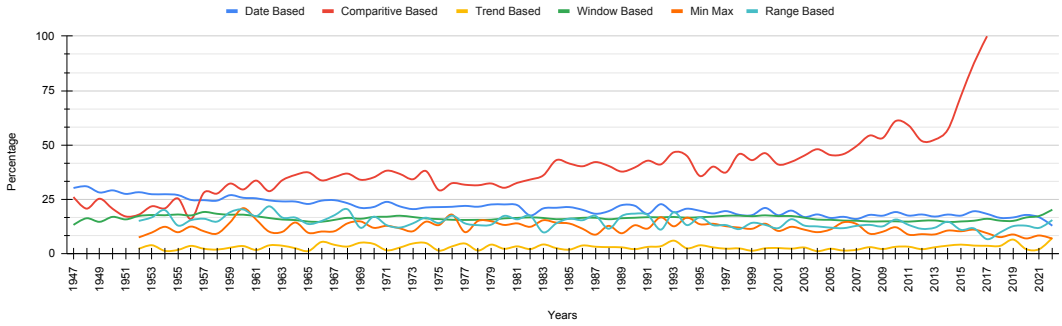

Figure 15: **Zeroshot** MCQ-based evaluation on `gemma-7b-it`.

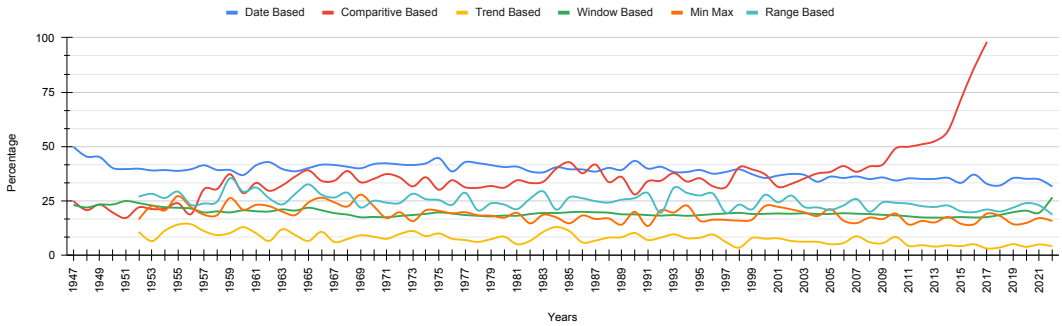

Figure 16: **Zeroshot** MCQ-based evaluation on `llama-3-8b`.

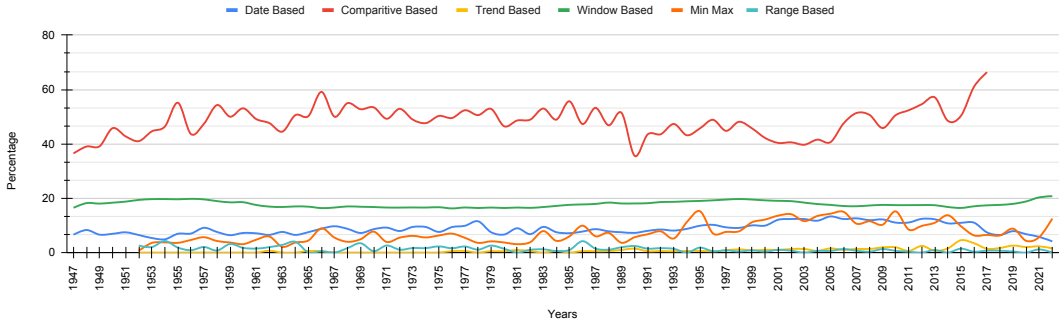

Figure 17: **Zeroshot** MCQ-based evaluation on `phi-3-instruct`.

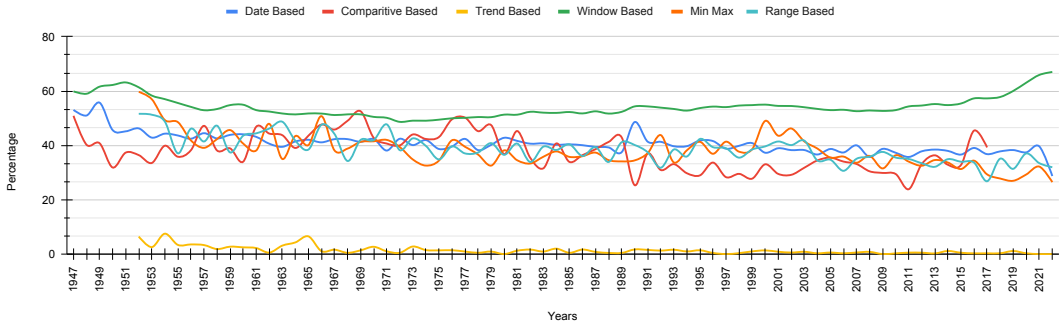

Figure 18: **Zeroshot** MCQ-based evaluation on `mixtral-8x7b`.

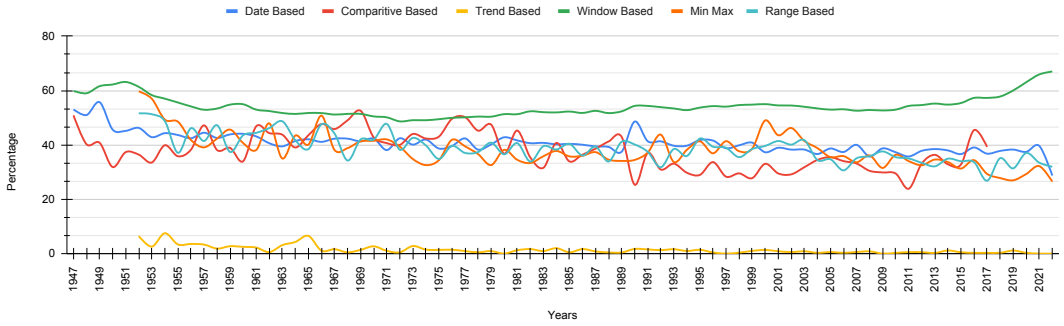

Figure 19: **Zeroshot** MCQ-based evaluation on `llama-3-70B`.

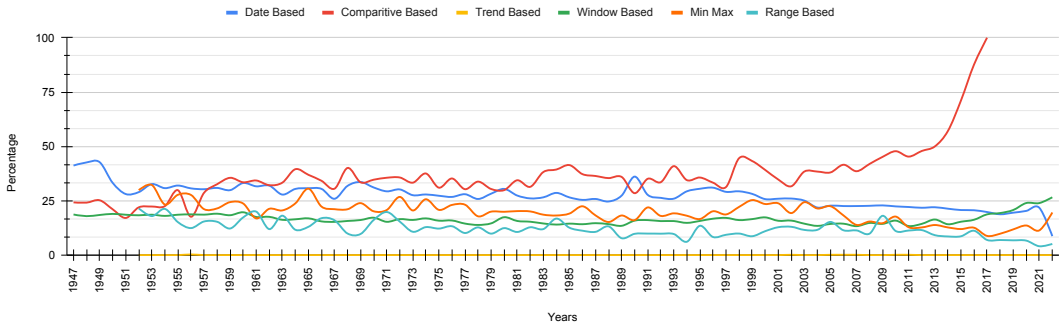

Figure 20: **Zeroshot** MCQ-based evaluation on `gpt-3.5-turbo`.

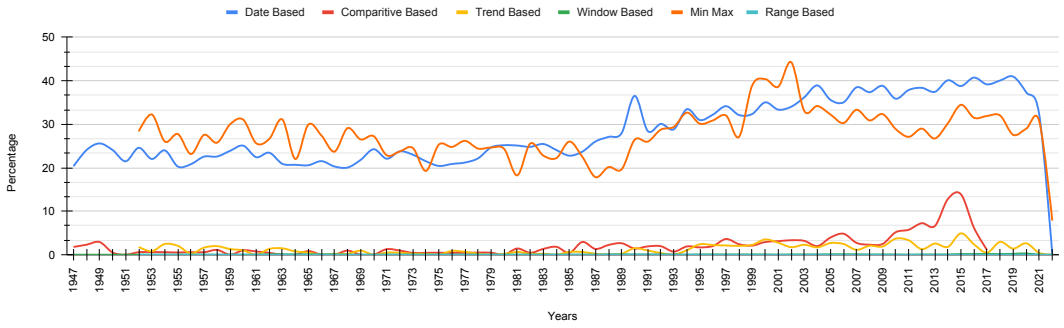

Figure 21: **Zeroshot** MCQ-based evaluation on `gpt-4`.

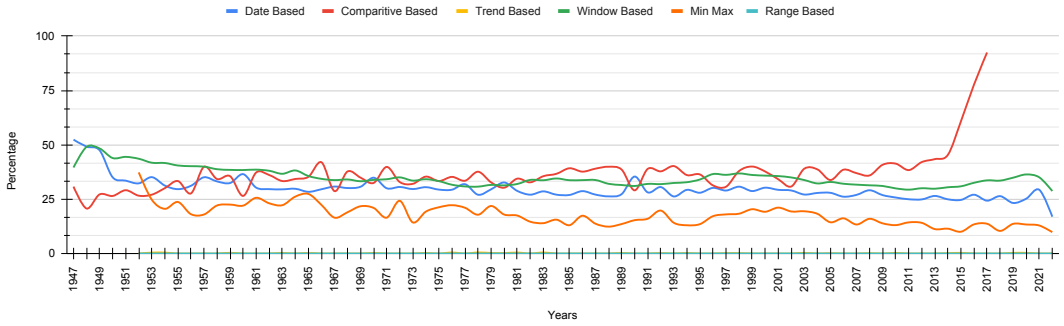

Figure 22: **Zeroshot** MCQ-based evaluation on `gemini-pro`.

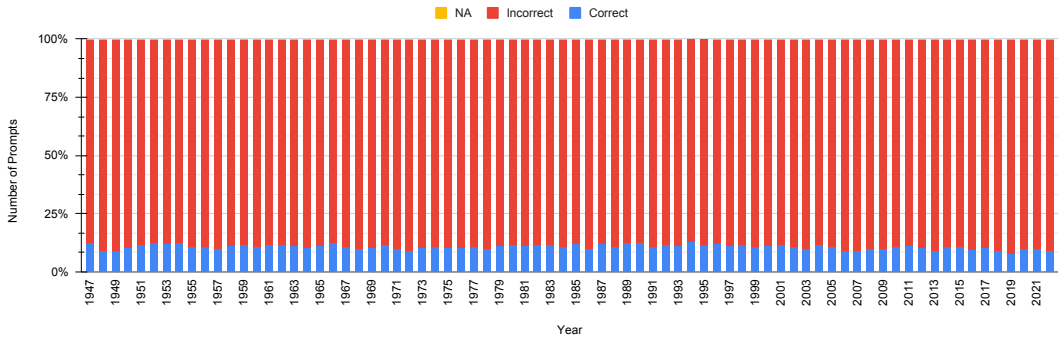

Figure 23: Plot for the Date-based metric (DB) for year-wise count for `phi-2` in **Zeroshot evaluation**.

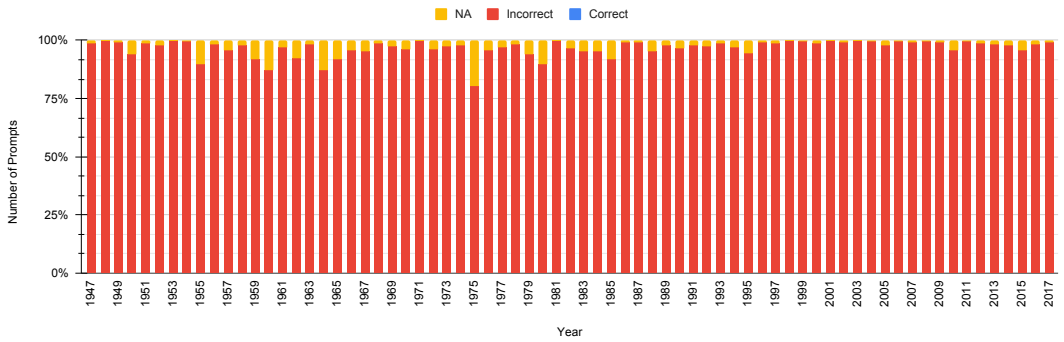

Figure 24: Plot for the Comparative-based metric (CP) for year-wise count for `phi-2` in **Zeroshot evaluation**.

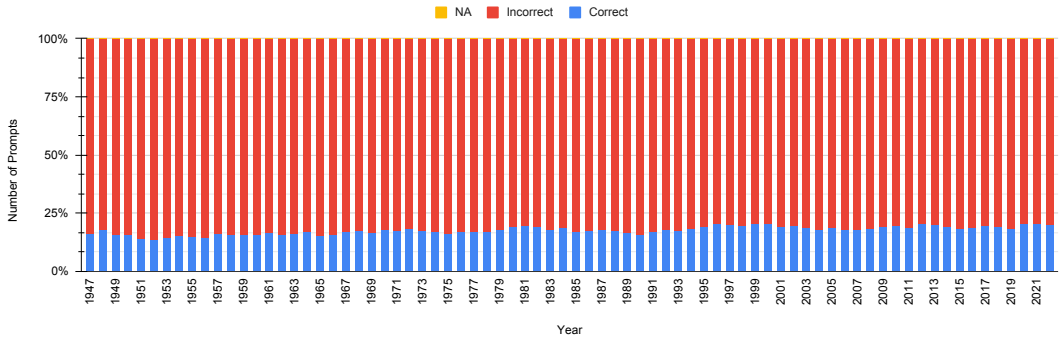

Figure 25: Plot for the Window-based metric (WB) for year-wise count for `phi-2` in **Zeroshot evaluation**.

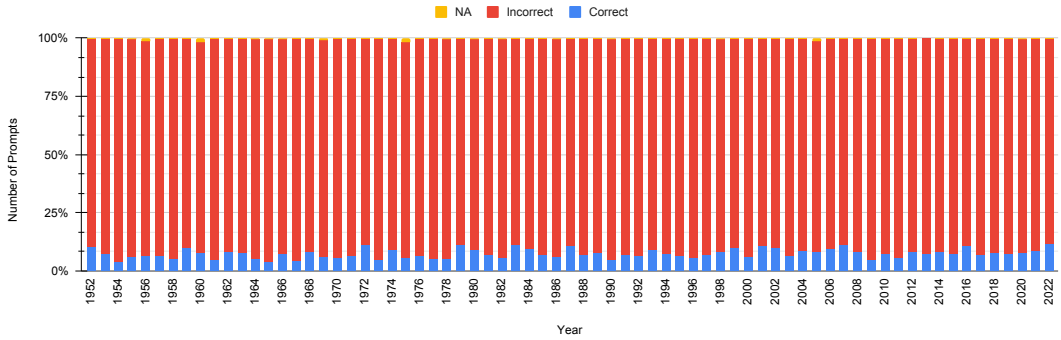

Figure 26: Plot for the Min/Max-based metric (MM) for year-wise count for `phi-2` in **Zeroshot evaluation**.

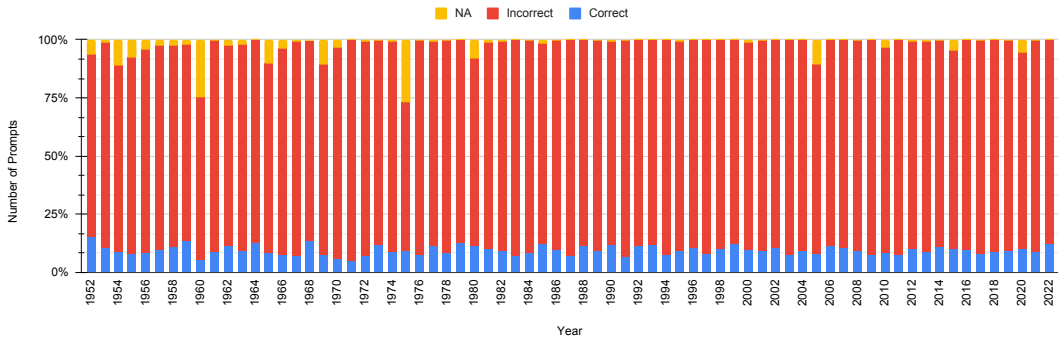

Figure 27: Plot for the Range-based metric (RB) for year-wise count for `phi-2` in **Zeroshot evaluation**.

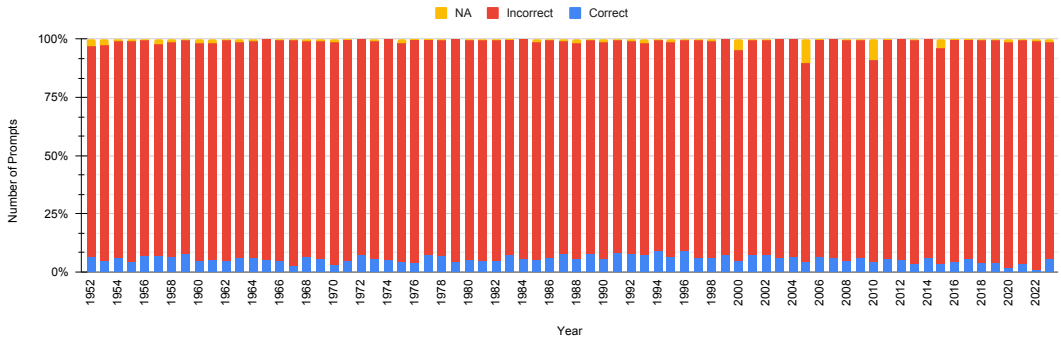

Figure 28: Plot for the Trend-based metric (TB) for year-wise count for `phi-2` in **Zeroshot evaluation**.

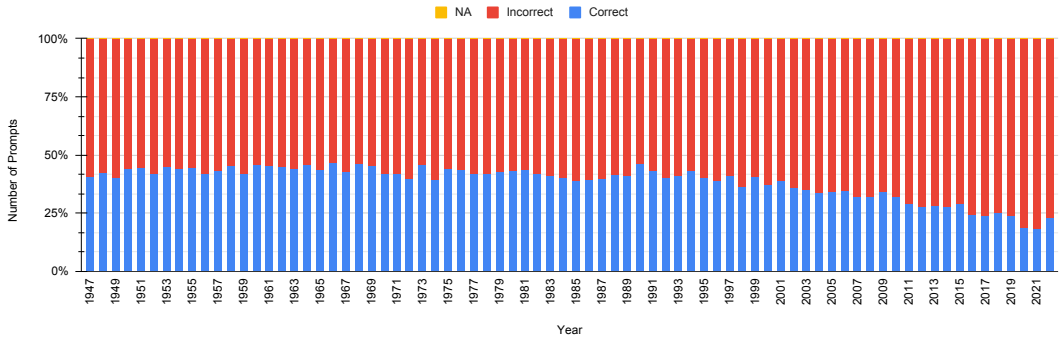

Figure 29: Plot for the Date-based metric (DB) for year-wise count for `flan-t5-xl` in **Zeroshot evaluation**.

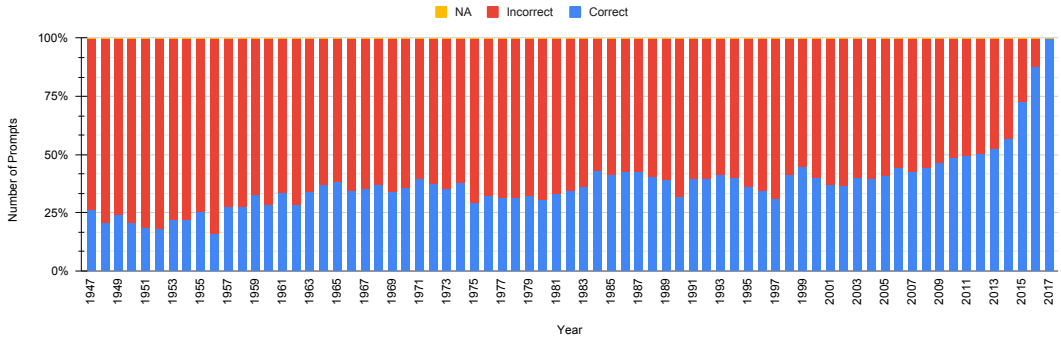

Figure 30: Plot for the Comparative-based metric (CP) for year-wise count for `flan-t5-xl` in **Zeroshot evaluation**.

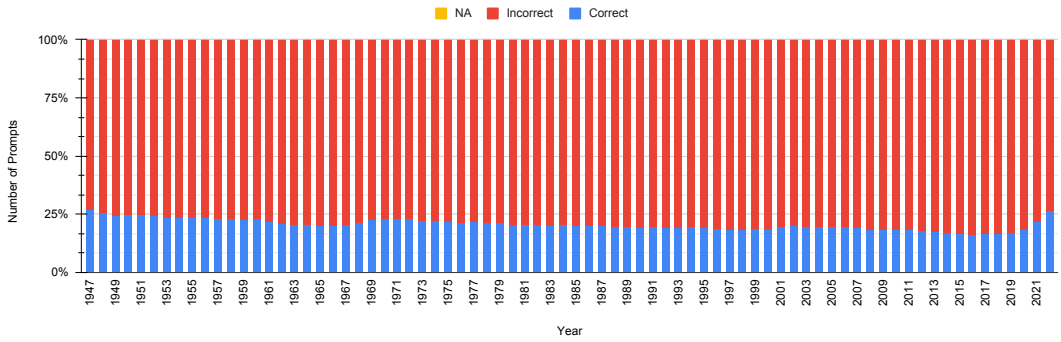

Figure 31: Plot for the Window-based metric (WB) for year-wise count for `flan-t5-xl` in **Zeroshot evaluation**.

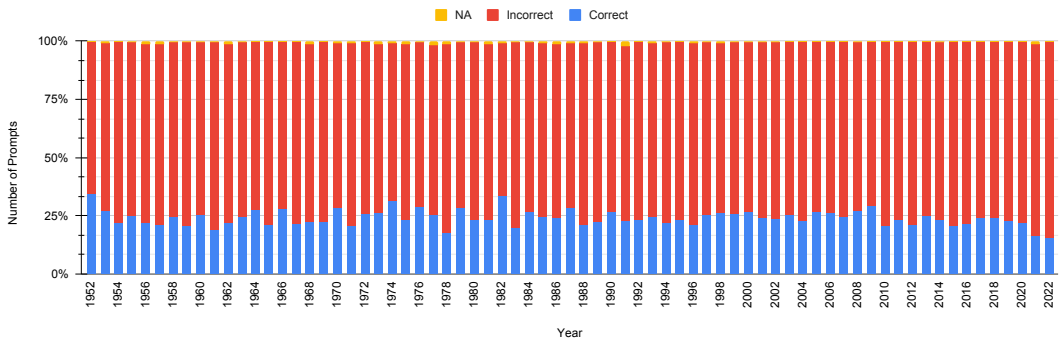

Figure 32: Plot for the Min/Max-based metric (MM) for year-wise count for `flan-t5-xl` in **Zeroshot evaluation**.

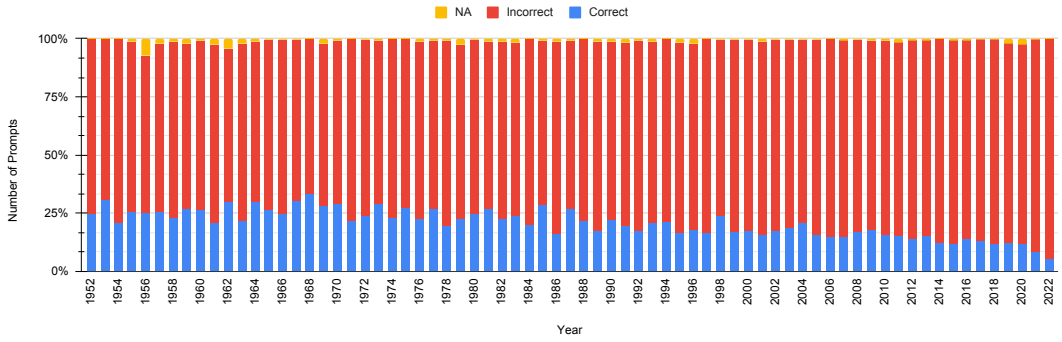

Figure 33: Plot for the Range-based metric (RB) for year-wise count for `flan-t5-xl` in **Zeroshot evaluation**.

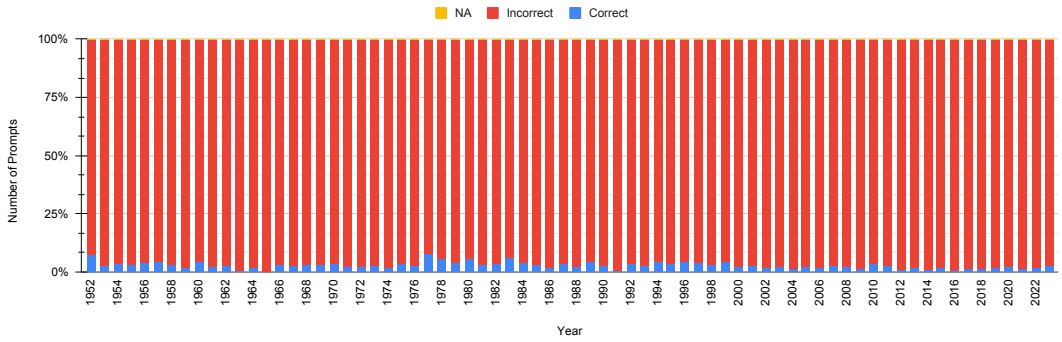

Figure 34: Plot for the Trend-based metric (TB) for year-wise count for `flan-t5-xl` in **Zeroshot evaluation**.

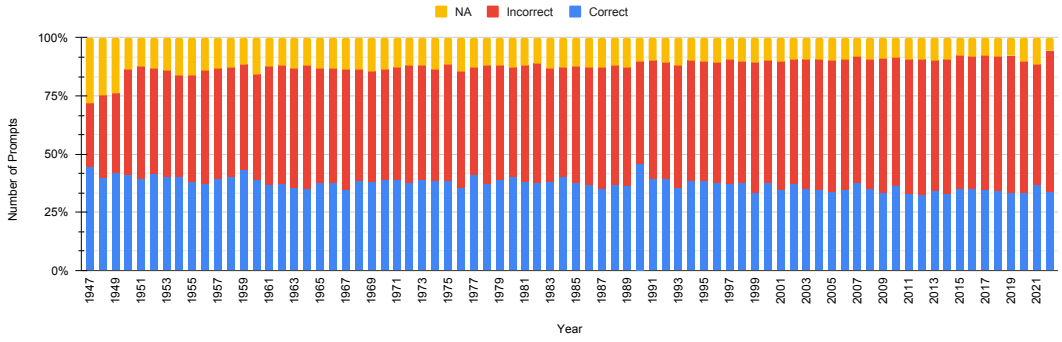

Figure 35: Plot for the Date-based metric (DB) for year-wise count for `mistral-instruct` in **Zeroshot evaluation**.

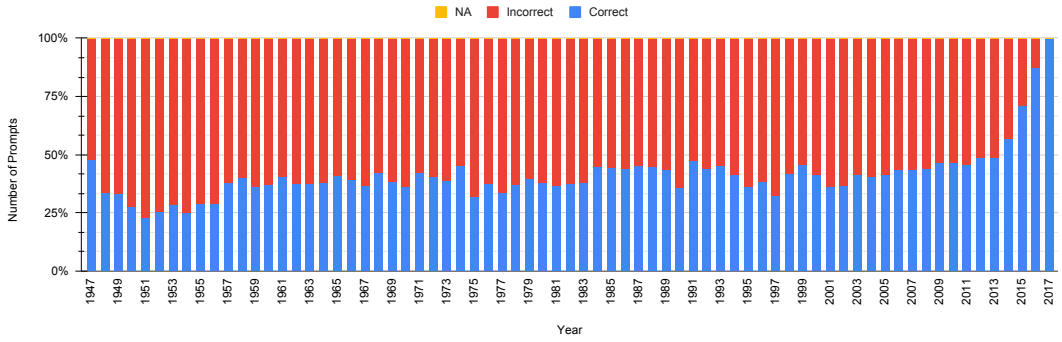

Figure 36: Plot for the Comparative-based metric (CP) for year-wise count for `mistral-instruct` in **Zeroshot evaluation**.

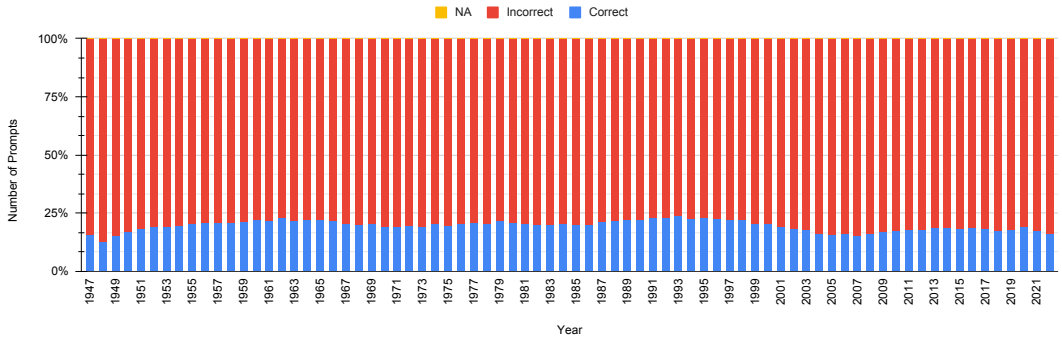

Figure 37: Plot for the Window-based metric (WB) for year-wise count for `mistral-instruct` in **Zeroshot evaluation**.

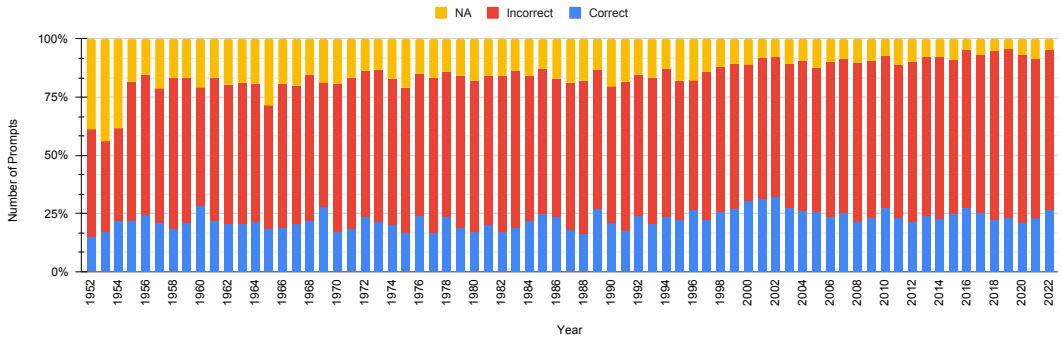

Figure 38: Plot for the Min/Max-based metric (MM) for year-wise count for `mistral-instruct` in **Zeroshot evaluation**.

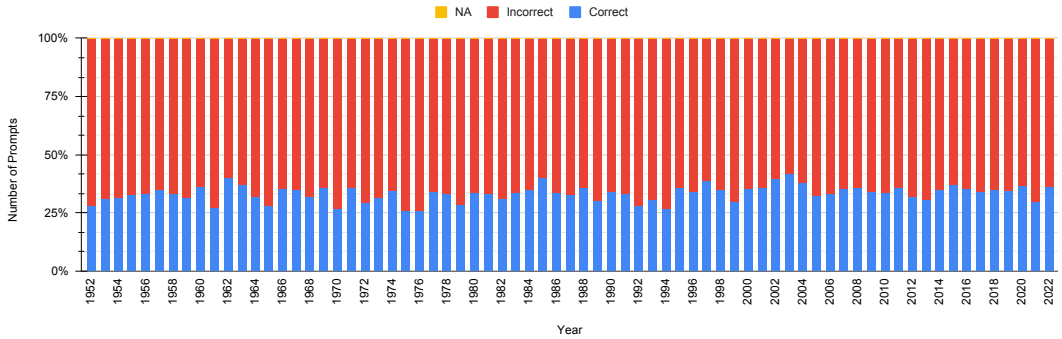

Figure 39: Plot for the Range-based metric (RB) for year-wise count for `mistral-instruct` in **Zeroshot evaluation**.

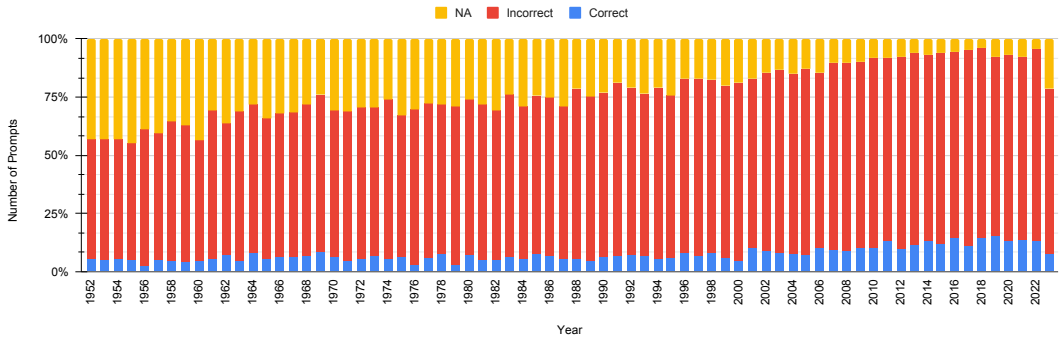

Figure 40: Plot for the Trend-based metric (TB) for year-wise count for `mistral-instruct` in **Zeroshot evaluation**.

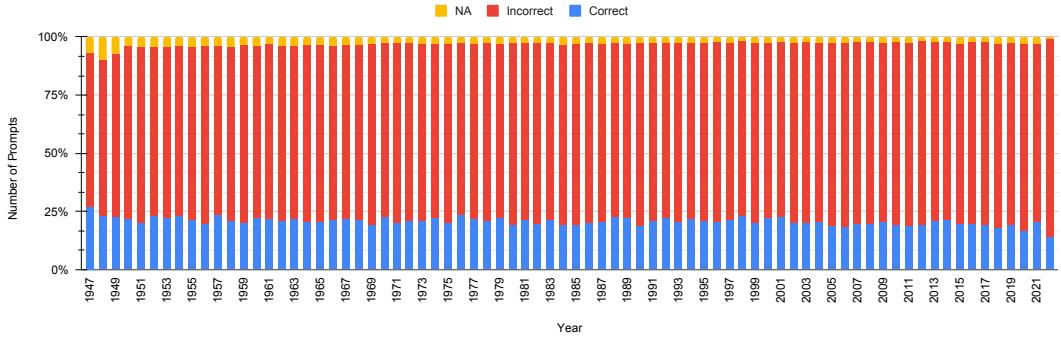

Figure 41: Plot for the Date-based metric (DB) for year-wise count for `llama-2-chat` in **Zeroshot evaluation**.

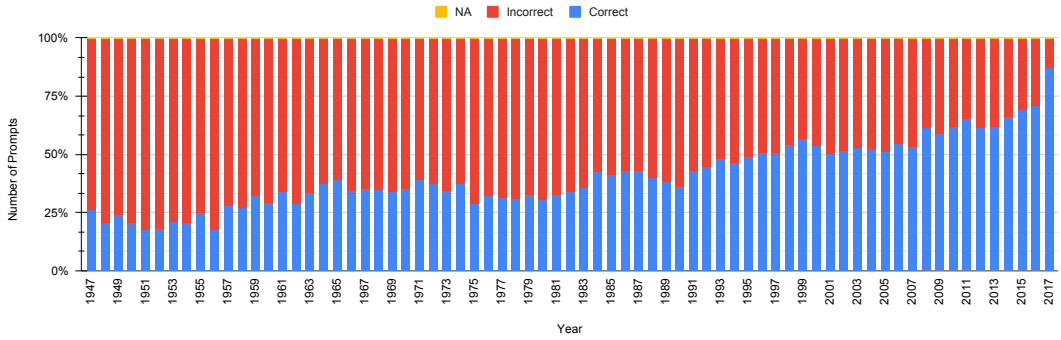

Figure 42: Plot for the Comparative-based metric (CP) for year-wise count for `llama-2-chat` in **Zeroshot evaluation**.

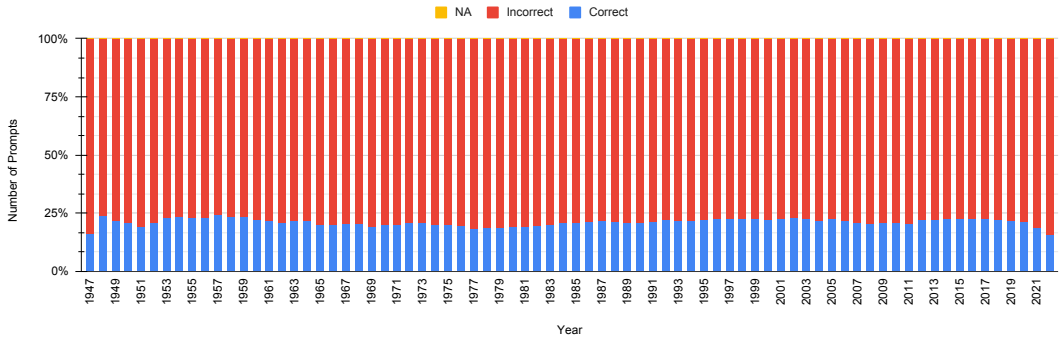

Figure 43: Plot for the Window-based metric (WB) for year-wise count for `llama-2-chat` in **Zeroshot evaluation**.

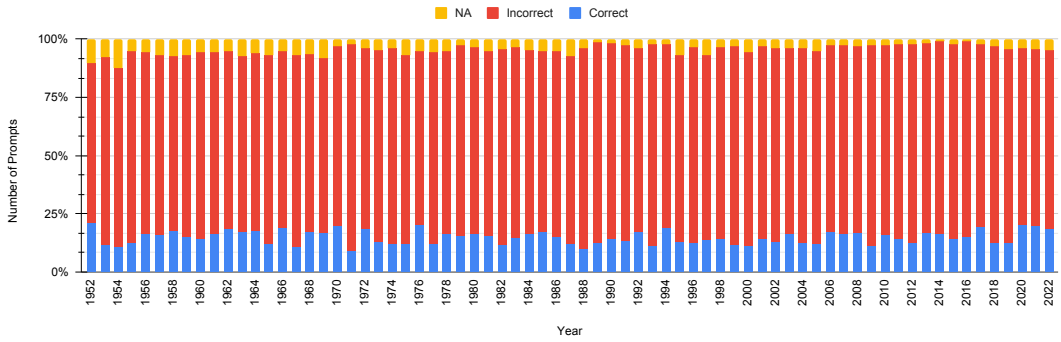

Figure 44: Plot for the Min/Max-based metric (MM) for year-wise count for `llama-2-chat` in **Zeroshot evaluation**.

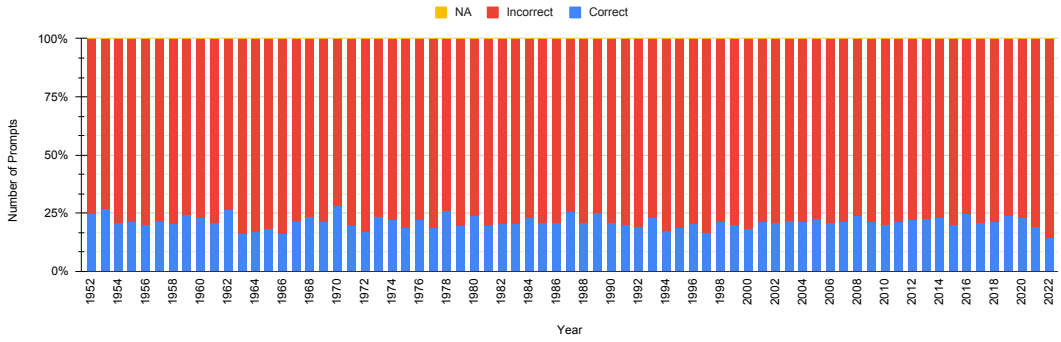

Figure 45: Plot for the Range-based metric (RB) for year-wise count for `llama-2-chat` in **Zeroshot evaluation**.

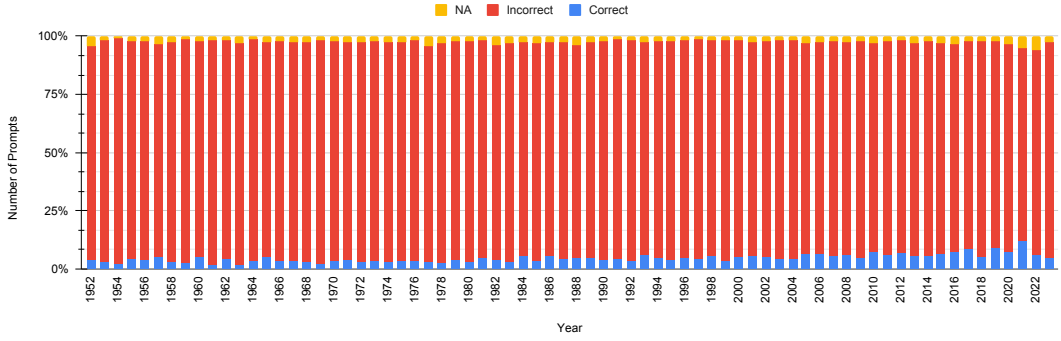

Figure 46: Plot for the Trend-based metric (TB) for year-wise count for `llama-2-chat` in **Zeroshot evaluation**.

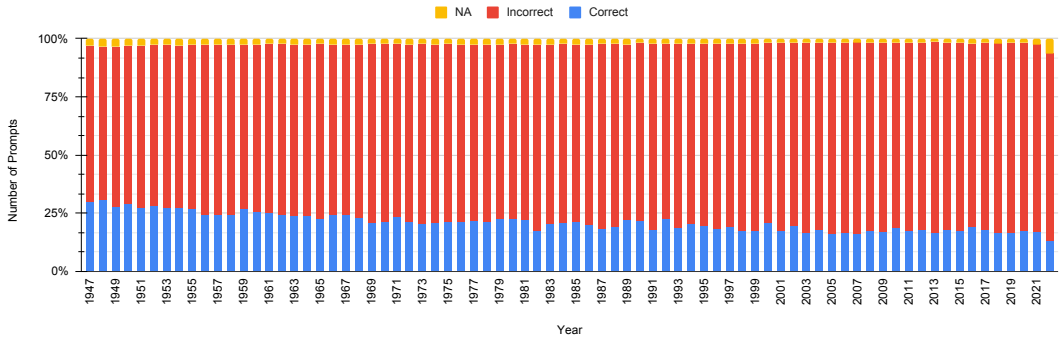

Figure 47: Plot for the Date-based metric (DB) for year-wise count for `gemma-7b-it` in **Zeroshot evaluation**.

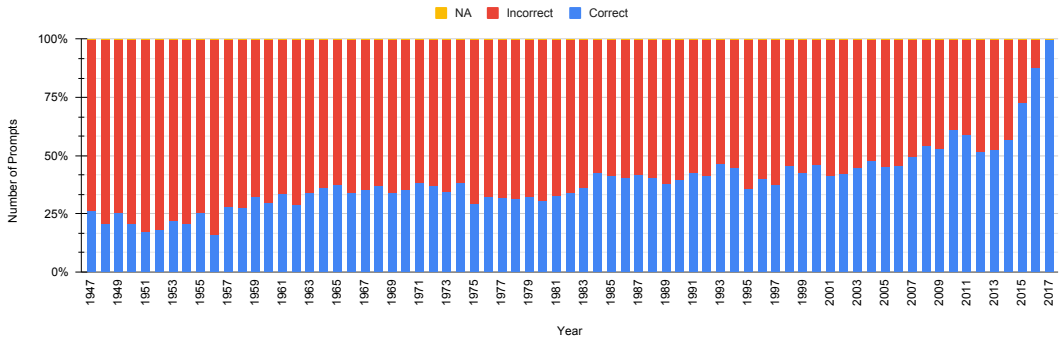

Figure 48: Plot for the Comparative-based metric (CP) for year-wise count for `gemma-7b-it` in **Zeroshot evaluation**.

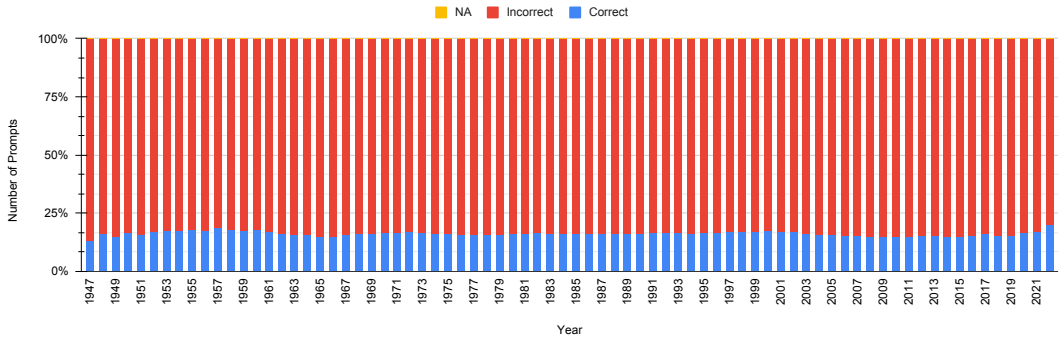

Figure 49: Plot for the Window-based metric (WB) for year-wise count for `gemma-7b-it` in **Zeroshot evaluation**.

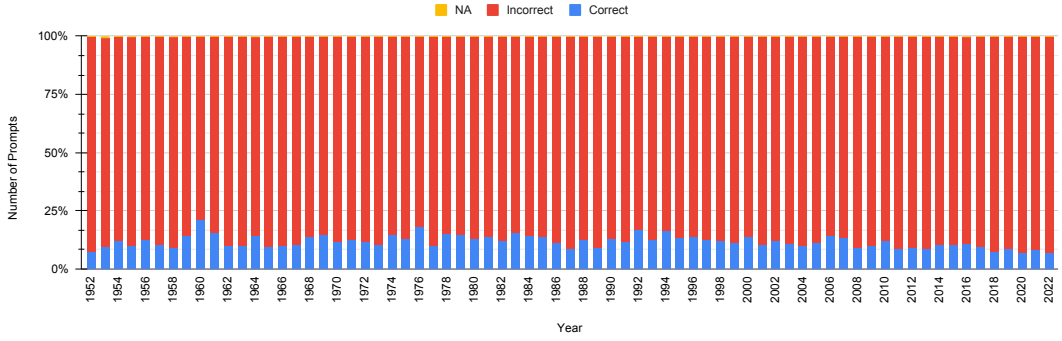

Figure 50: Plot for the Min/Max-based metric (MM) for year-wise count for `gemma-7b-it` in **Zeroshot evaluation**.

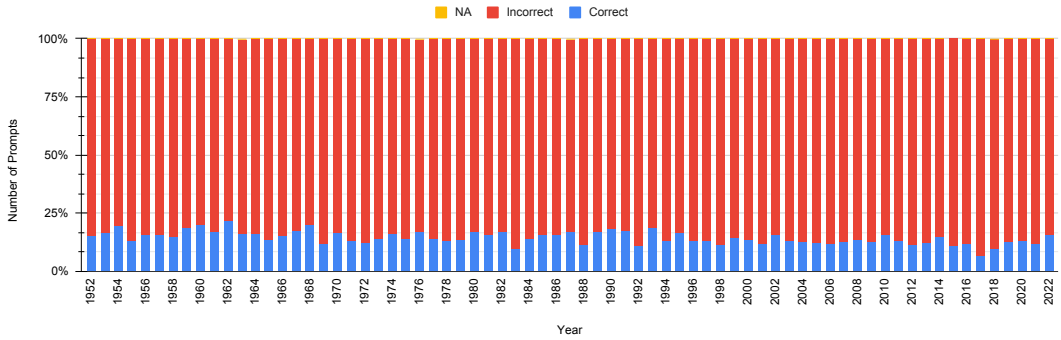

Figure 51: Plot for the Range-based metric (RB) for year-wise count for `gemma-7b-it` in **Zeroshot evaluation**.

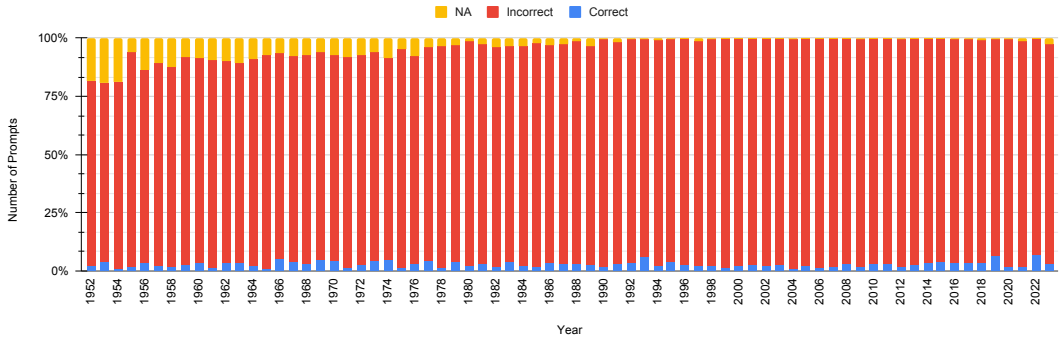

Figure 52: Plot for the Trend-based metric (TB) for year-wise count for `gemma-7b-it` in **Zeroshot evaluation**.

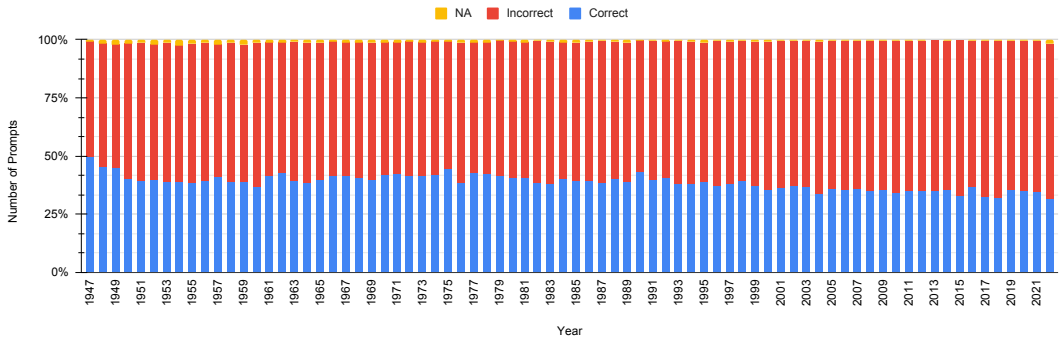

Figure 53: Plot for the Date-based metric (DB) for year-wise count for `llama-3-8b` in **Zeroshot evaluation**.

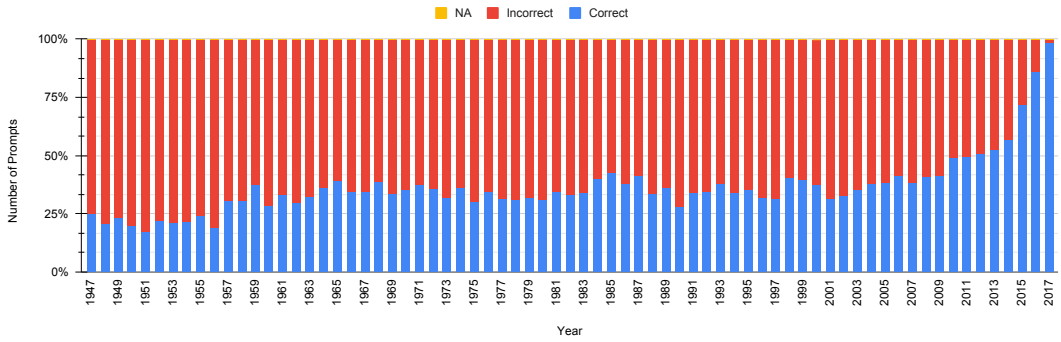

Figure 54: Plot for the Comparative-based metric (CP) for year-wise count for `llama-3-8b` in **Zeroshot evaluation**.

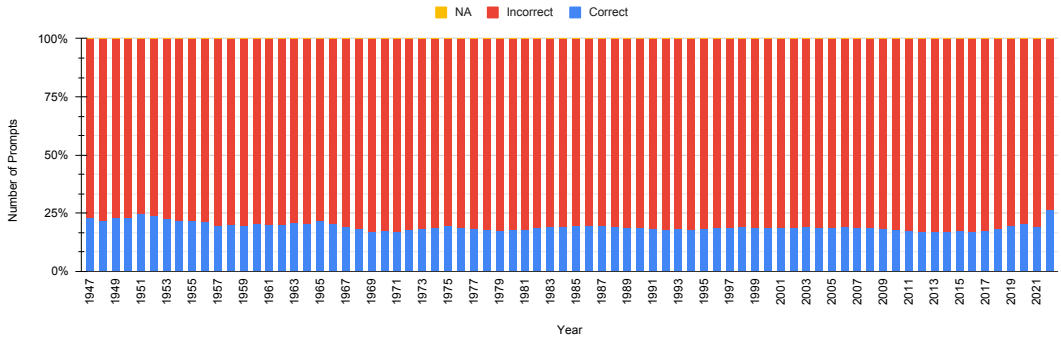

Figure 55: Plot for the Window-based metric (WB) for year-wise count for `llama-3-8b` in **Zeroshot evaluation**.

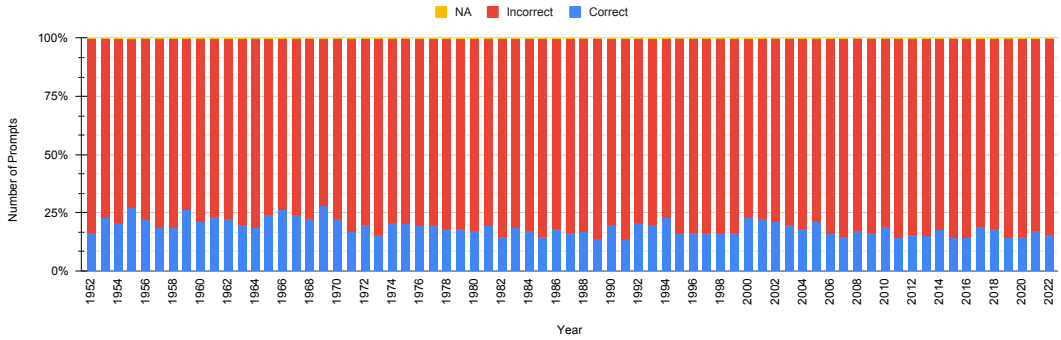

Figure 56: Plot for the Min/Max-based metric (MM) for year-wise count for `llama-3-8b` in **Zeroshot evaluation**.

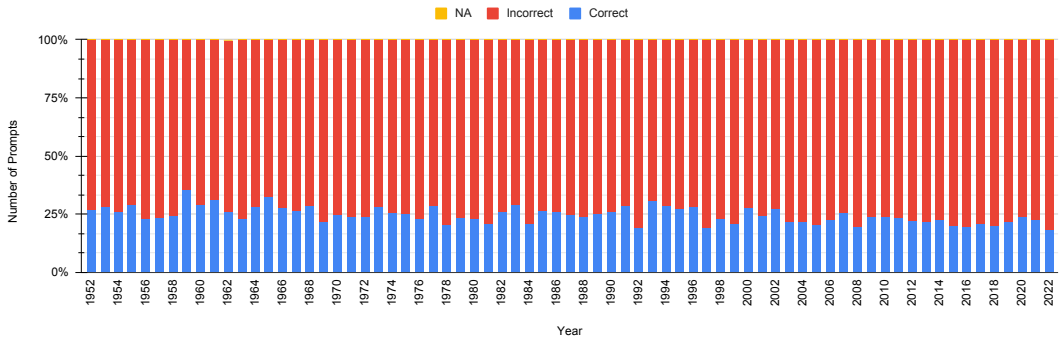

Figure 57: Plot for the Range-based metric (RB) for year-wise count for `llama-3-8b` in **Zeroshot evaluation**.

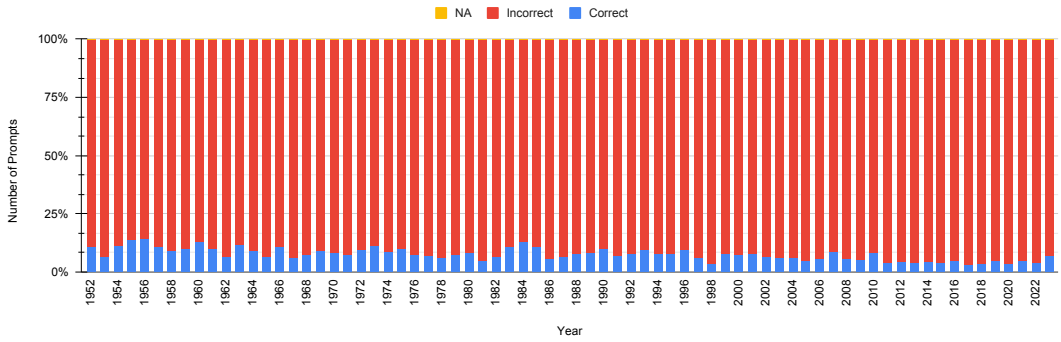

Figure 58: Plot for the Trend-based metric (TB) for year-wise count for `llama-3-8b` in **Zeroshot evaluation**.

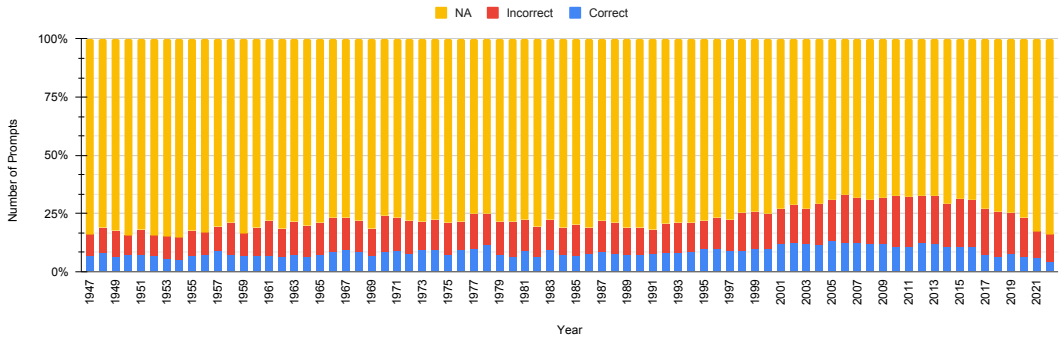

Figure 59: Plot for the Date-based metric (DB) for year-wise count for `phi-3-instruct` in **Zeroshot evaluation**.

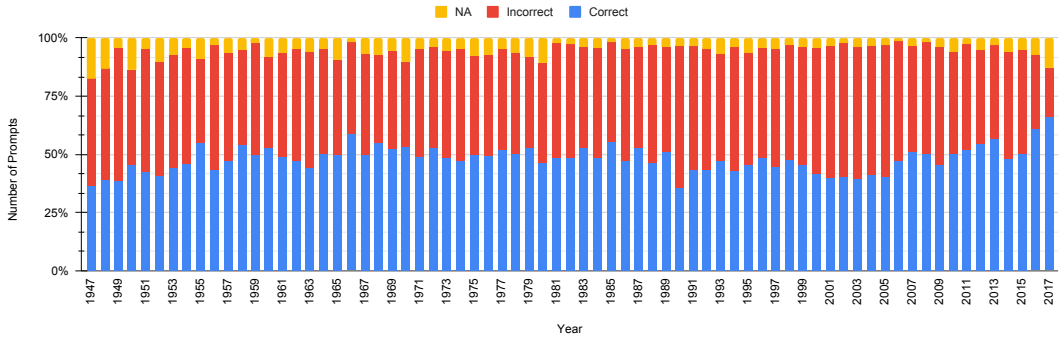

Figure 60: Plot for the Comparative-based metric (CP) for year-wise count for `phi-3-instruct2` in **Zeroshot evaluation**.

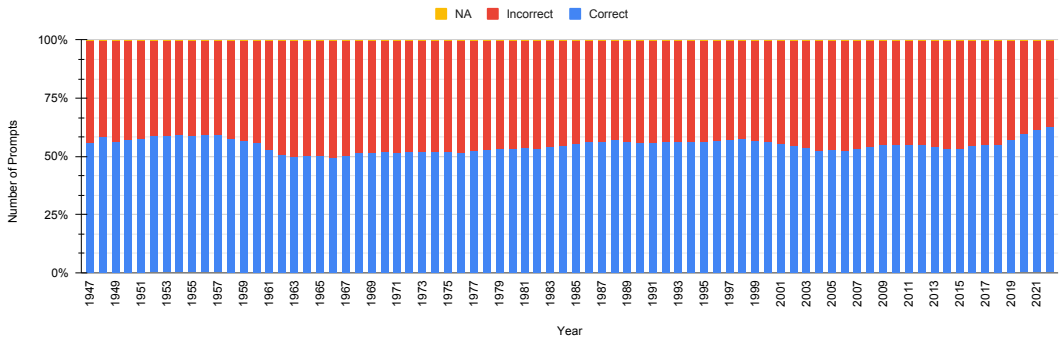

Figure 61: Plot for the Window-based metric (WB) for year-wise count for `phi-3-instruct2` in **Zeroshot evaluation**.

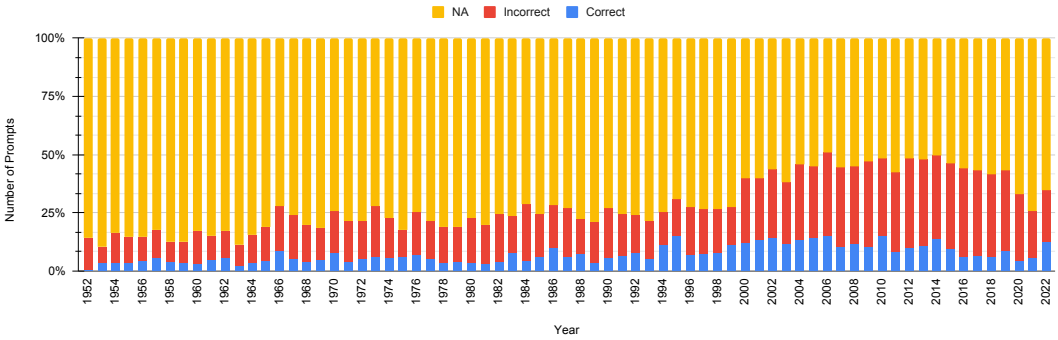

Figure 62: Plot for the Min/Max-based metric (MM) for year-wise count for `phi-3-instruct2` in **Zeroshot evaluation**.

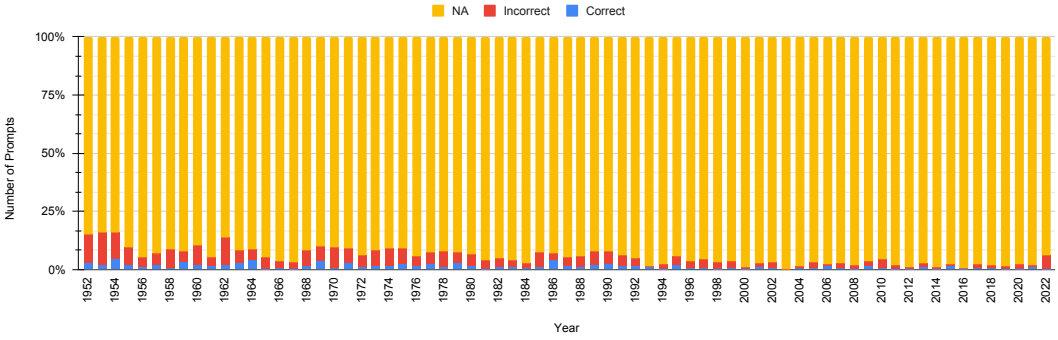

Figure 63: Plot for the Range-based metric (RB) for year-wise count for `phi-3-instruct2` in **Zeroshot evaluation**.

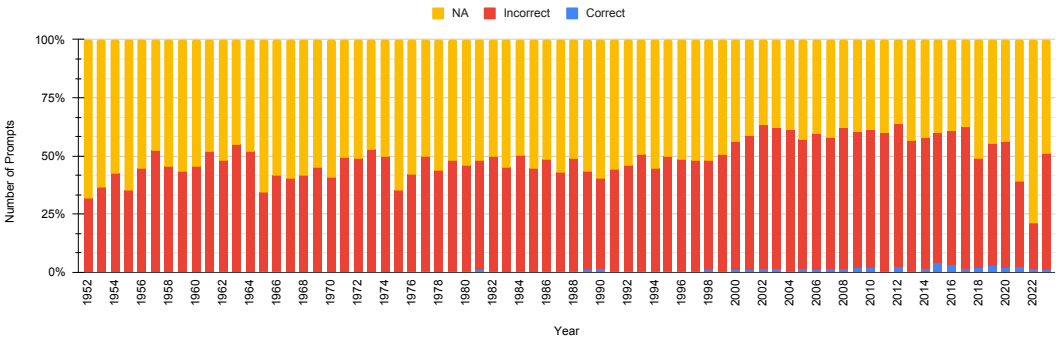

Figure 64: Plot for the Trend-based metric (TB) for year-wise count for `phi-3-instruct2` in **Zeroshot evaluation**.

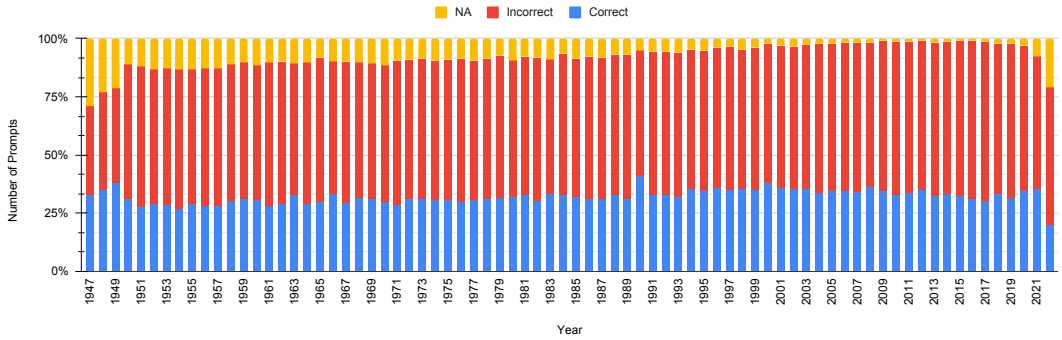

Figure 65: Plot for the Date-based metric (DB) for year-wise count for `mixtral-8x7b-32768` in **Zeroshot evaluation**.

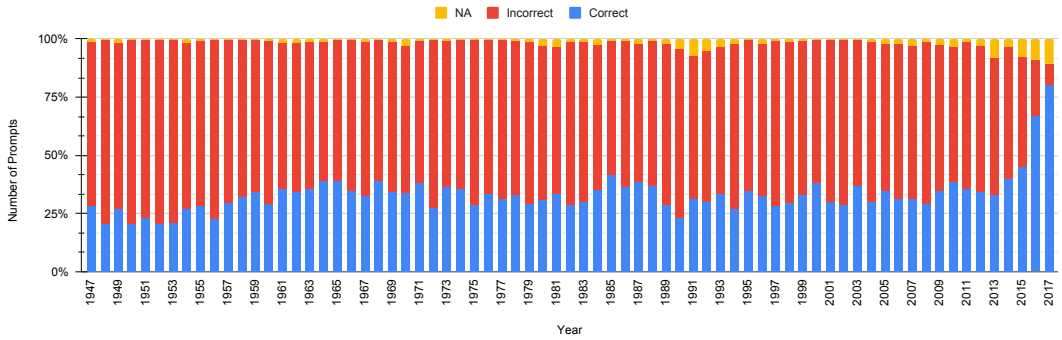

Figure 66: Plot for the Comparative-based metric (CP) for year-wise count for `mixtral-8x7b-32768` in **Zeroshot evaluation**.

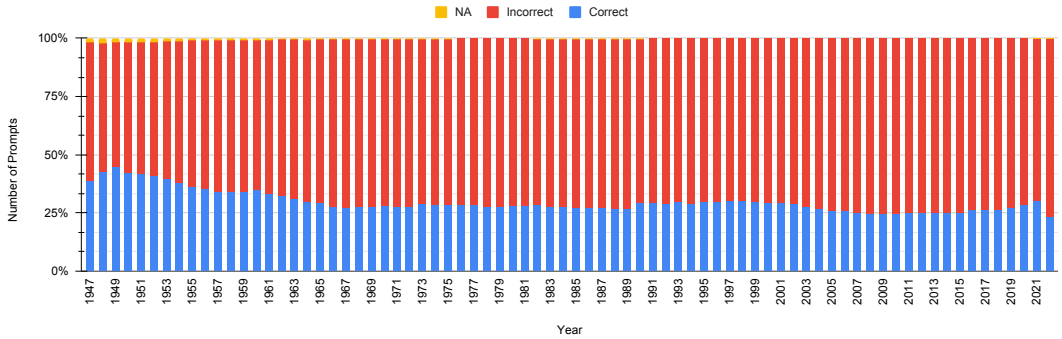

Figure 67: Plot for the Window-based metric (WB) for year-wise count for `mixtral-8x7b-32768` in **Zeroshot evaluation**.

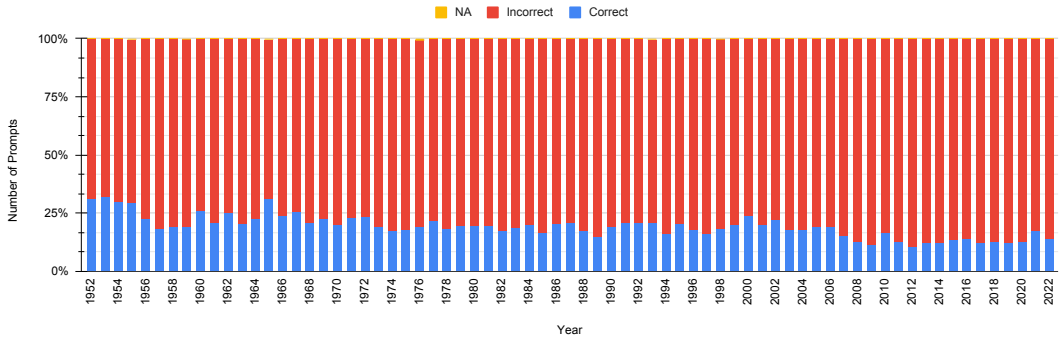

Figure 68: Plot for the Min/Max-based metric (MM) for year-wise count for `mixtral-8x7b-32768` in **Zeroshot evaluation**.

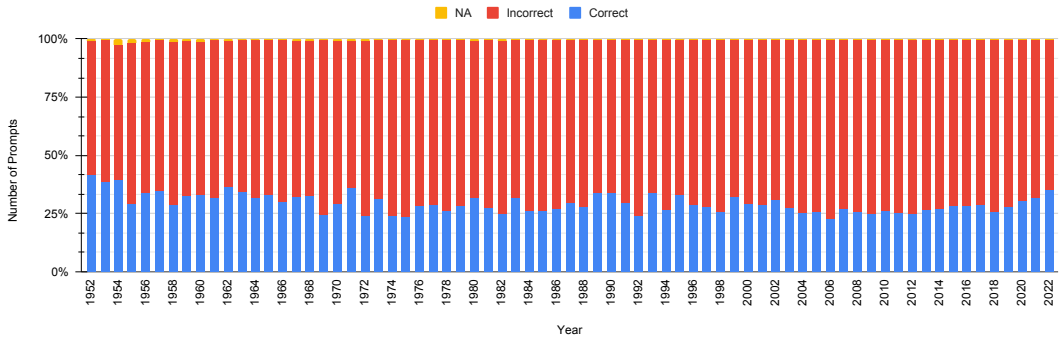

Figure 69: Plot for the Range-based metric (RB) for year-wise count for `mixtral-8x7b-32768` in **Zeroshot evaluation**.

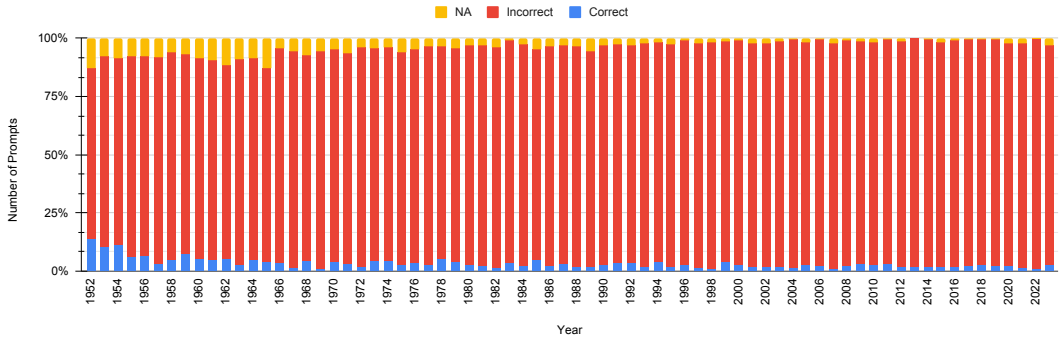

Figure 70: Plot for the Trend-based metric (TB) for year-wise count for `mixtral-8x7b-32768` in **Zeroshot evaluation**.

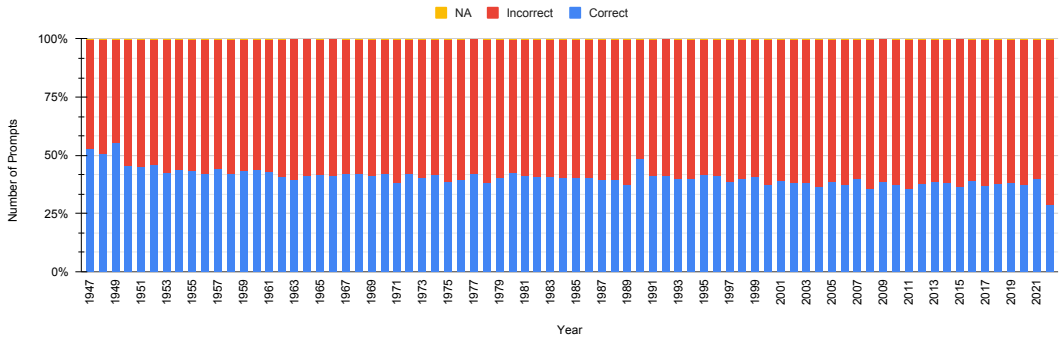

Figure 71: Plot for the Date-based metric (DB) for year-wise count for `llama-3-70b-8192` in **Zeroshot evaluation**.

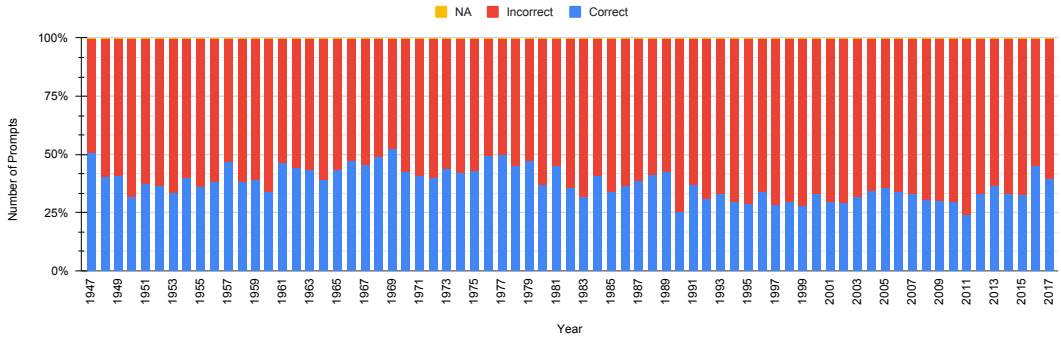

Figure 72: Plot for the Comparative-based metric (CP) for year-wise count for `llama-3-70b-8192` in **Zeroshot evaluation**.

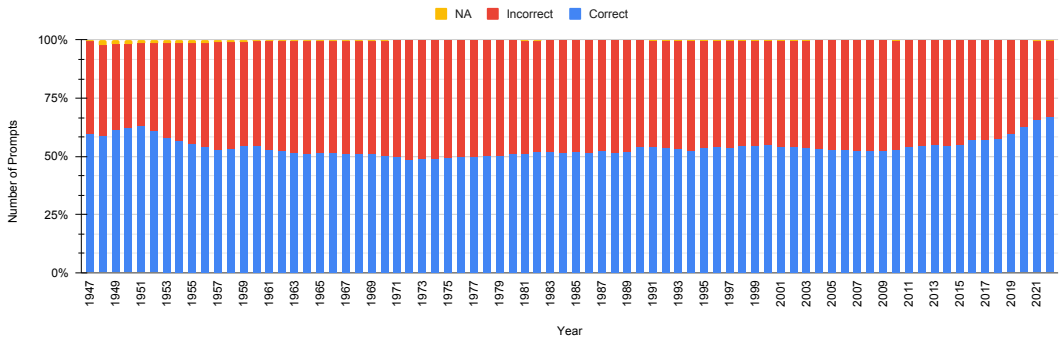

Figure 73: Plot for the Window-based metric (WB) for year-wise count for `llama-3-70b-8192` in **Zeroshot evaluation**.

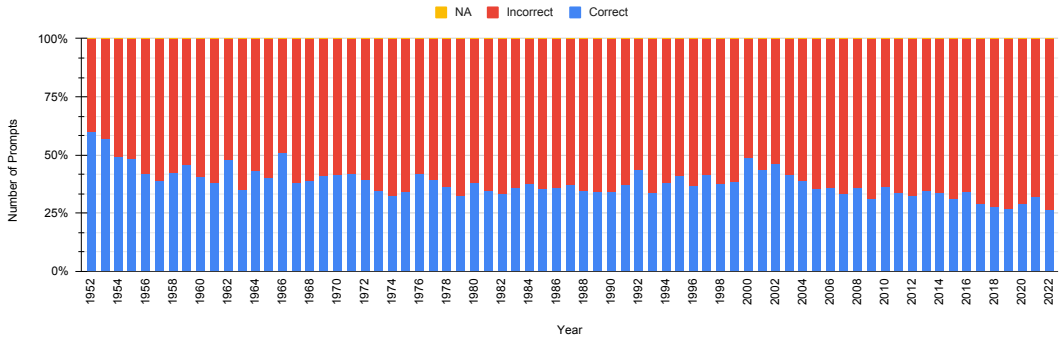

Figure 74: Plot for the Min/Max-based metric (MM) for year-wise count for `llama-3-70b-8192` in **Zeroshot evaluation**.

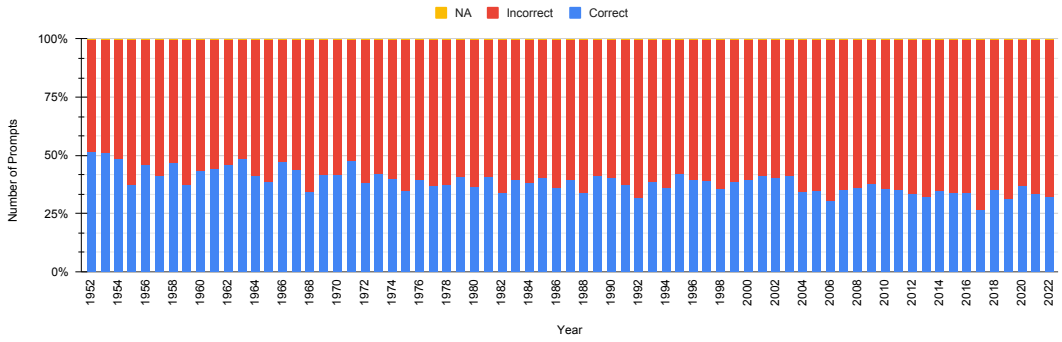

Figure 75: Plot for the Range-based metric (RB) for year-wise count for `llama-3-70b-8192` in **Zeroshot evaluation**.

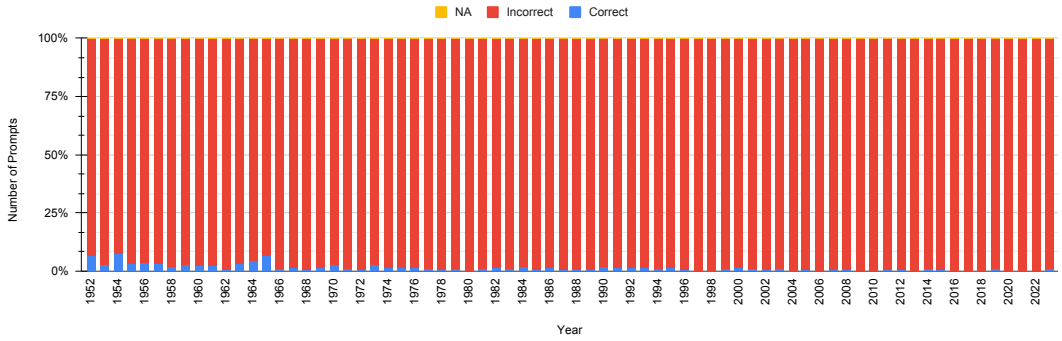

Figure 76: Plot for the Trend-based metric (TB) for year-wise count for `llama-3-70b-8192` in **Zeroshot evaluation**.

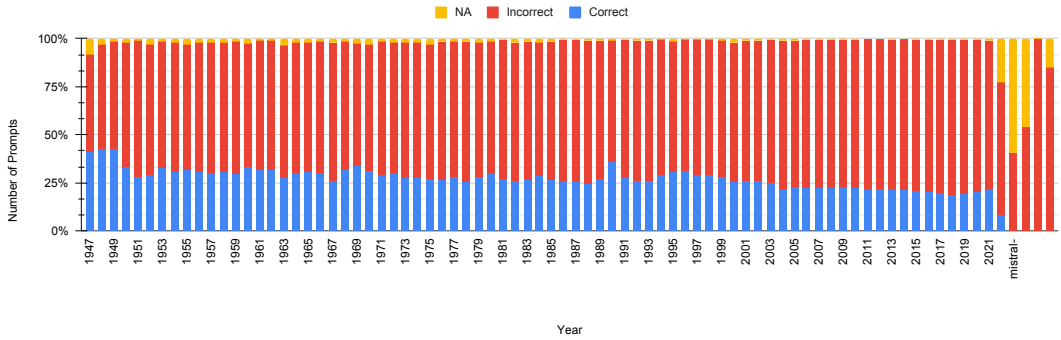

Figure 77: Plot for the Date-based metric (DB) for year-wise count for `gpt-3.5` in **Zeroshot evaluation**.

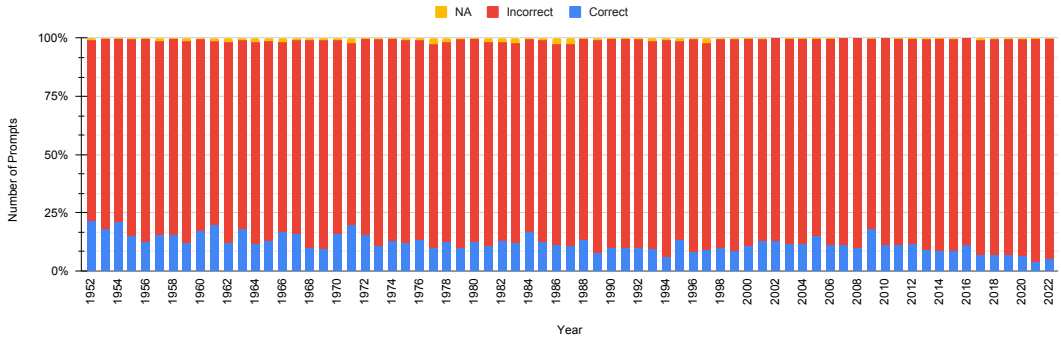

Figure 78: Plot for the Comparative-based metric (CP) for year-wise count for `gpt-3.5` in **Zeroshot evaluation**.

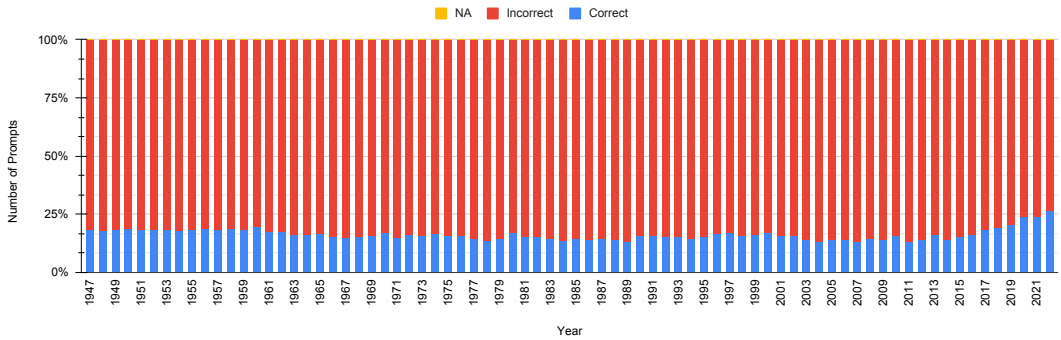

Figure 79: Plot for the Window-based metric (WB) for year-wise count for `gpt-3.5` in **Zeroshot evaluation**.

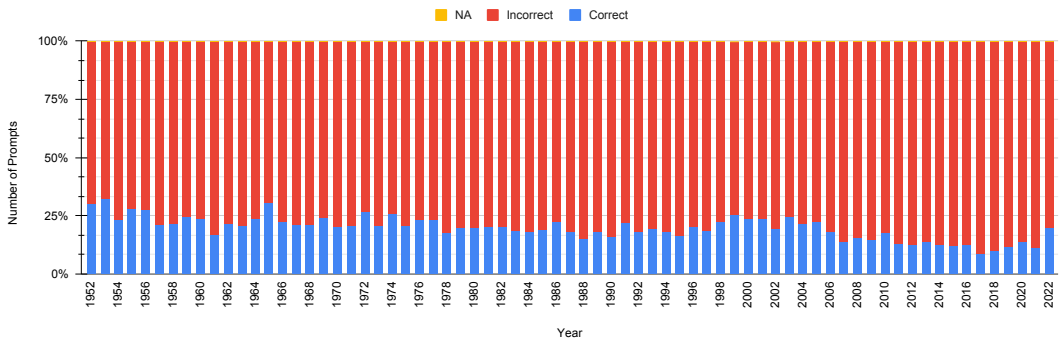

Figure 80: Plot for the Min/Max-based metric (MM) for year-wise count for `gpt-3.5` in **Zeroshot evaluation**.

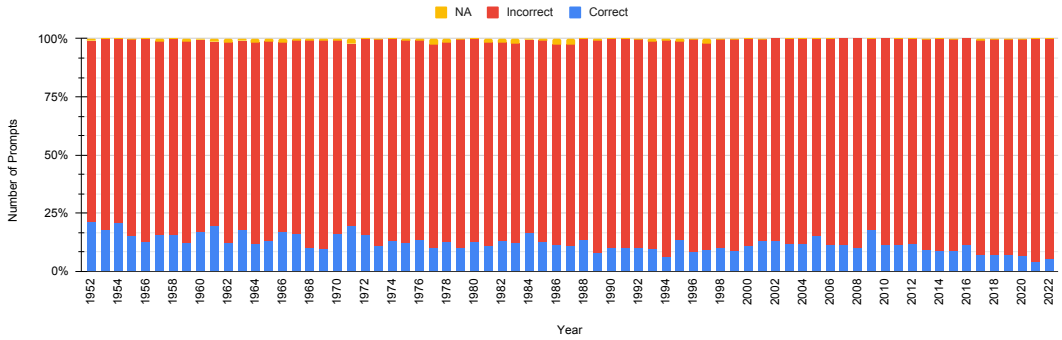

Figure 81: Plot for the Range-based metric (RB) for year-wise count for `gpt-3.5` in **Zeroshot evaluation**.

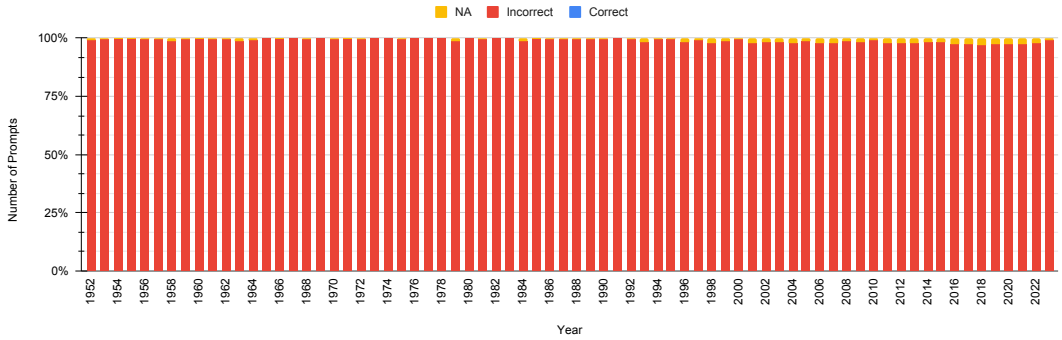

Figure 82: Plot for the Trend-based metric (TB) for year-wise count for `gpt-3.5` in **Zeroshot evaluation**.

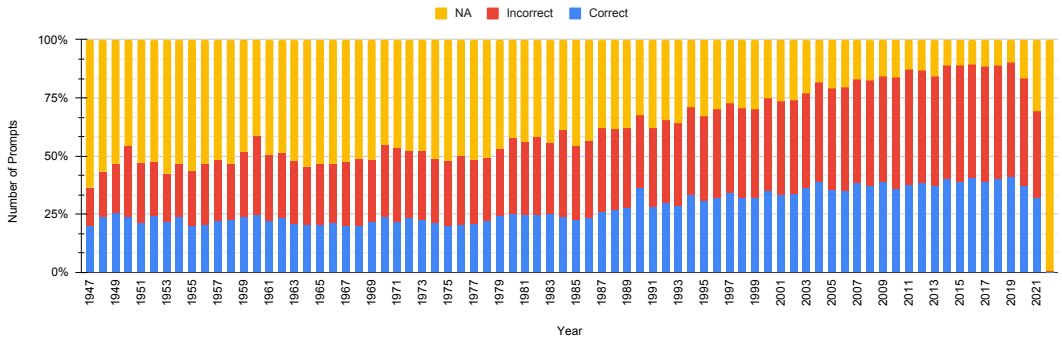

Figure 83: Plot for the Date-based metric (DB) for year-wise count for `gpt-4` in **Zeroshot evaluation**.

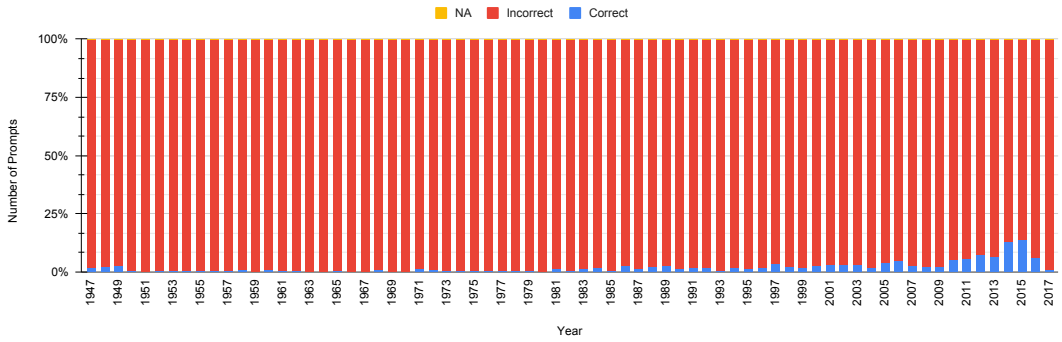

Figure 84: Plot for the Comparative-based metric (CP) for year-wise count for `gpt-4` in **Zeroshot evaluation**.

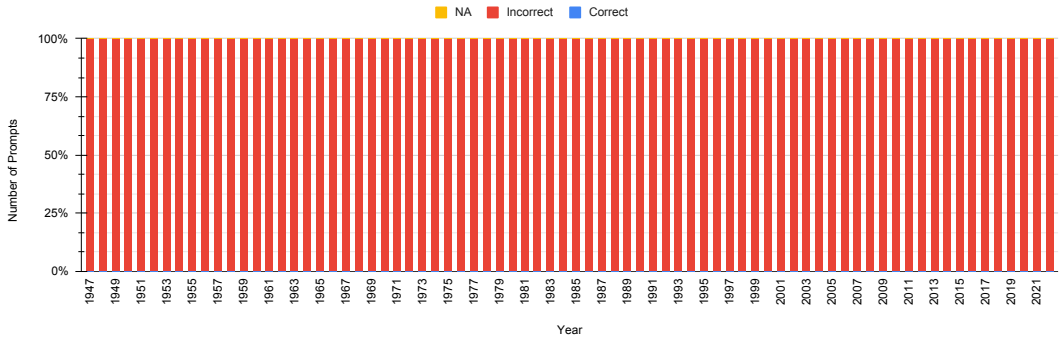

Figure 85: Plot for the Window-based metric (WB) for year-wise count for `gpt-4` in **Zeroshot evaluation**.

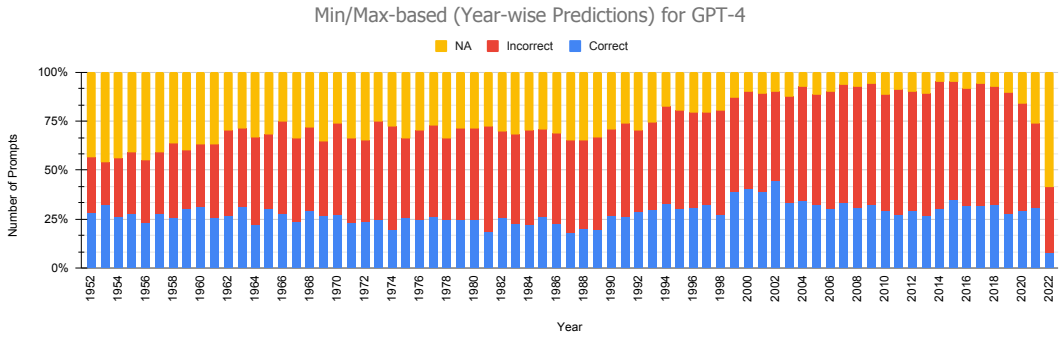

Figure 86: Plot for the Min/Max-based metric (MM) for year-wise count for `gpt-4` in **Zeroshot evaluation**.

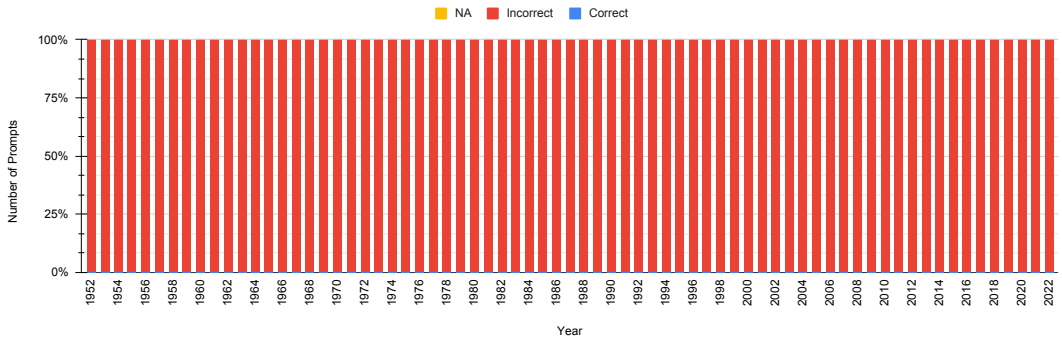

Figure 87: Plot for the Range-based metric (RB) for year-wise count for `gpt-4` in **Zeroshot evaluation**.

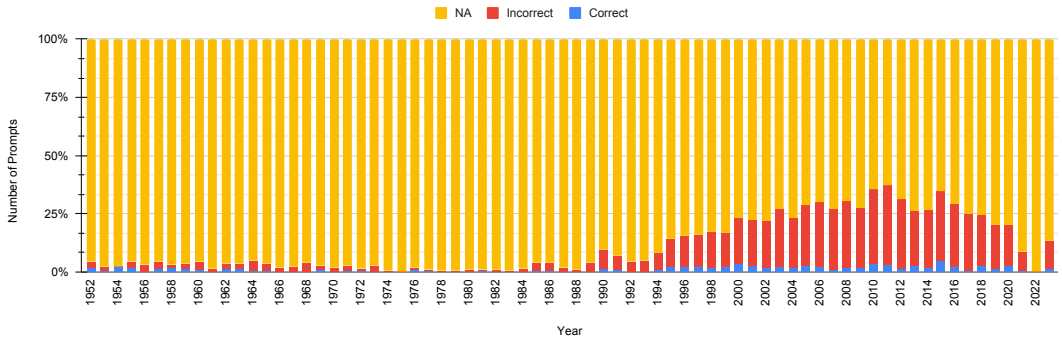

Figure 88: Plot for the Trend-based metric (TB) for year-wise count for `gpt-4` in **Zeroshot evaluation**.

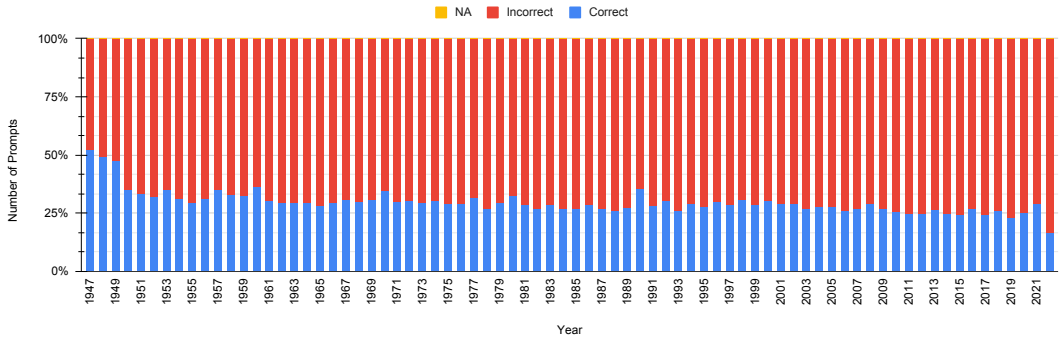

Figure 89: Plot for the Date-based metric (DB) for year-wise count for `gemini-pro` in **Zeroshot evaluation**.

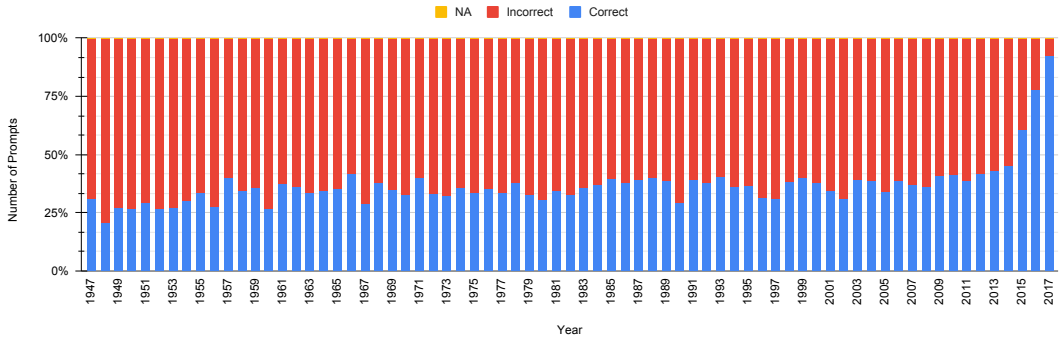

Figure 90: Plot for the Comparative-based metric (CP) for year-wise count for `gemini-pro` in **Zeroshot evaluation**.

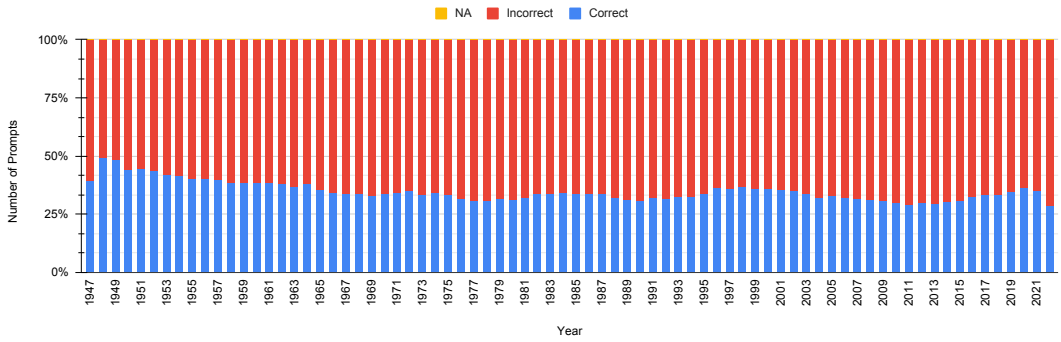

Figure 91: Plot for the Window-based metric (WB) for year-wise count for `gemini-pro` in **Zeroshot evaluation**.

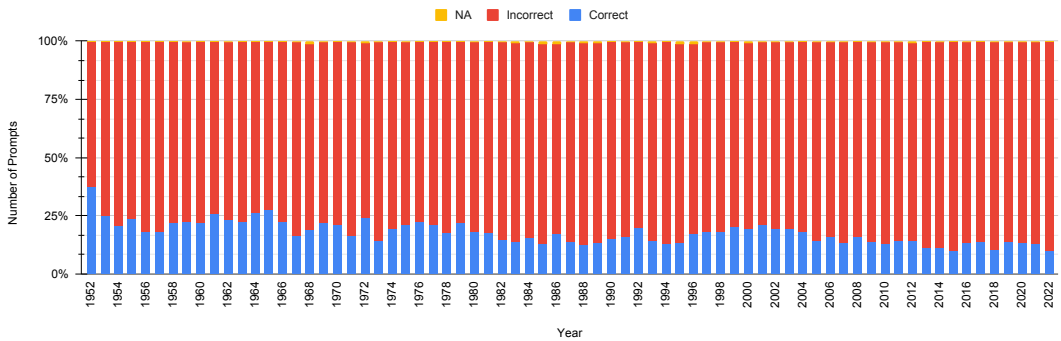

Figure 92: Plot for the Min/Max-based metric (MM) for year-wise count for `gemini-pro` in **Zeroshot evaluation**.

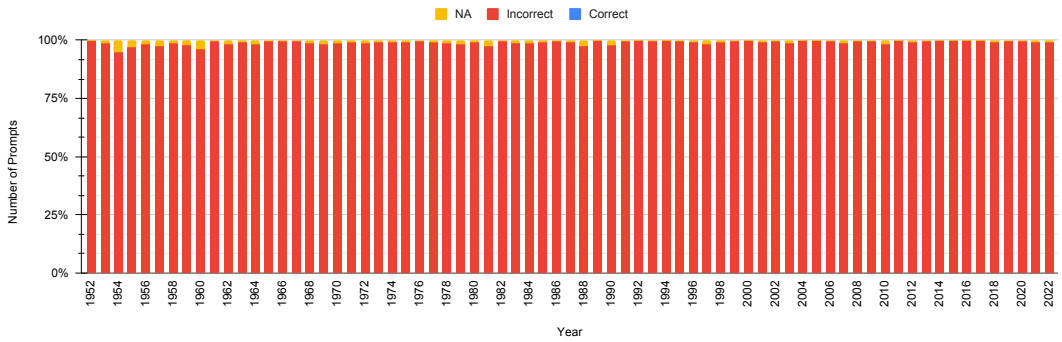

Figure 93: Plot for the Range-based metric (RB) for year-wise count for `gemini-pro` in **Zeroshot evaluation**.

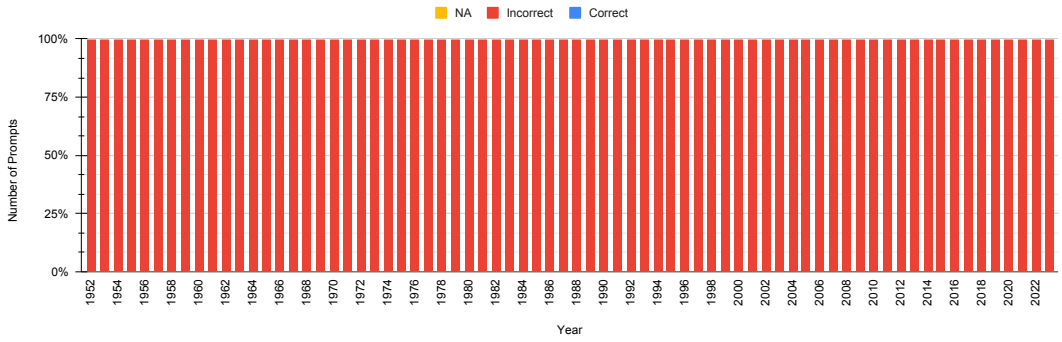

Figure 94: Plot for the Trend-based metric (TB) for year-wise count for `gemini-pro` in **Zeroshot evaluation**.

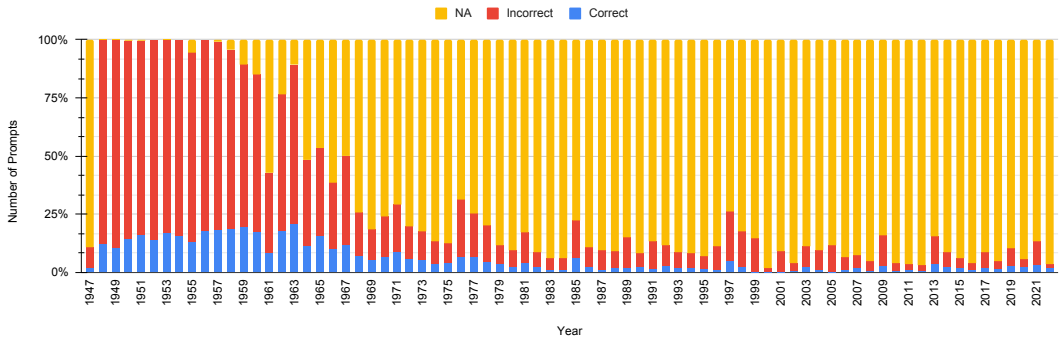

Figure 95: Plot for the Date-based metric (DB) as year-wise count (In percentage) for **continual fine-tuning** for `phi-2`.

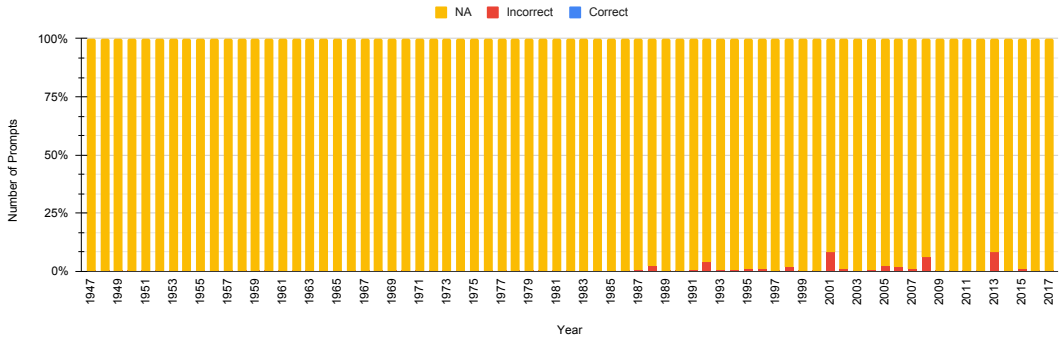

Figure 96: Plot for the Comparative-based metric (CP) as year-wise count (In percentage) for **continual fine-tuning** for `phi-2`.

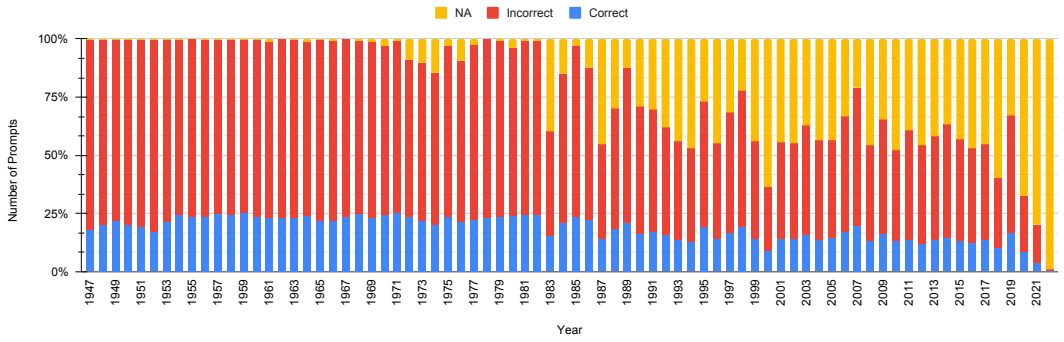

Figure 97: Plot for the Window-based metric (WB) as year-wise count (In percentage) for **continual fine-tuning** for `phi-2`.

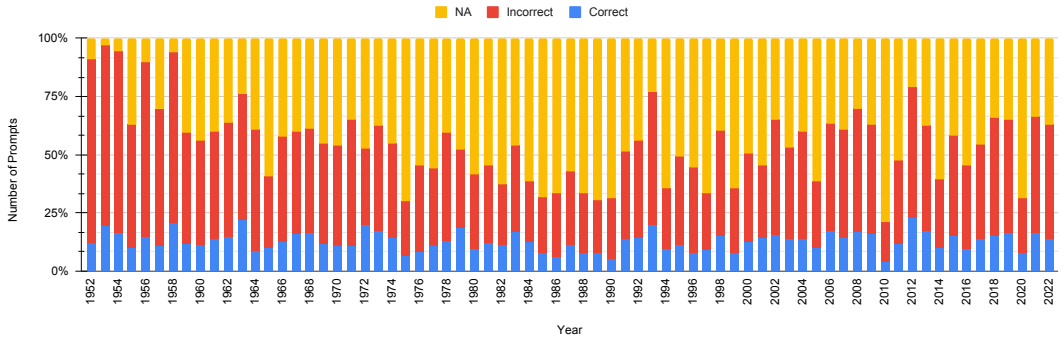

Figure 98: Plot for the Min/Max-based metric (MM) as year-wise count (In percentage) for **continual fine-tuning** for `phi-2`.

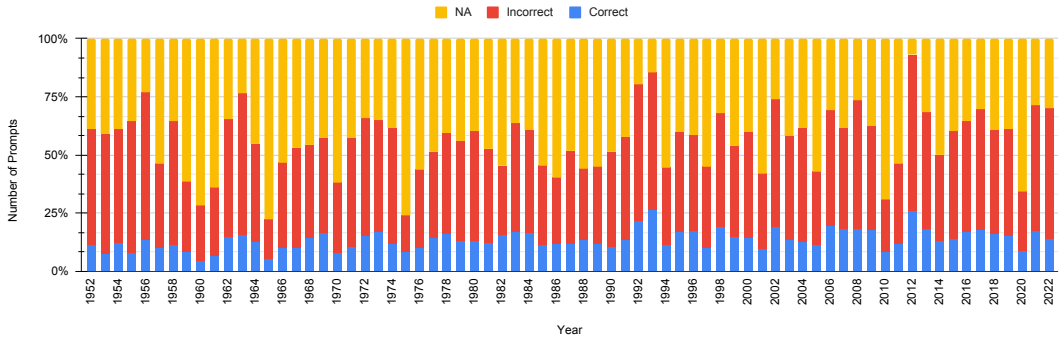

Figure 99: Plot for the Range-based metric (RB) as year-wise count (In percentage) for **continual fine-tuning** for `phi-2`.

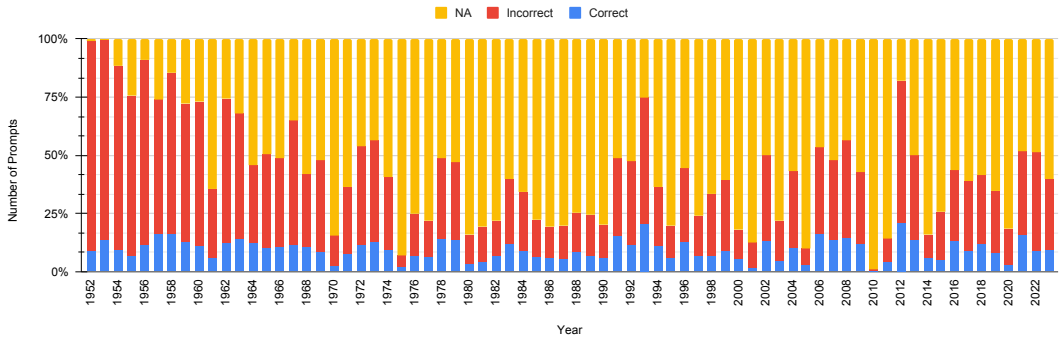

Figure 100: Plot for the Trend-based metric (TB) as year-wise count (In percentage) for **continual fine-tuning** for `phi-2`.

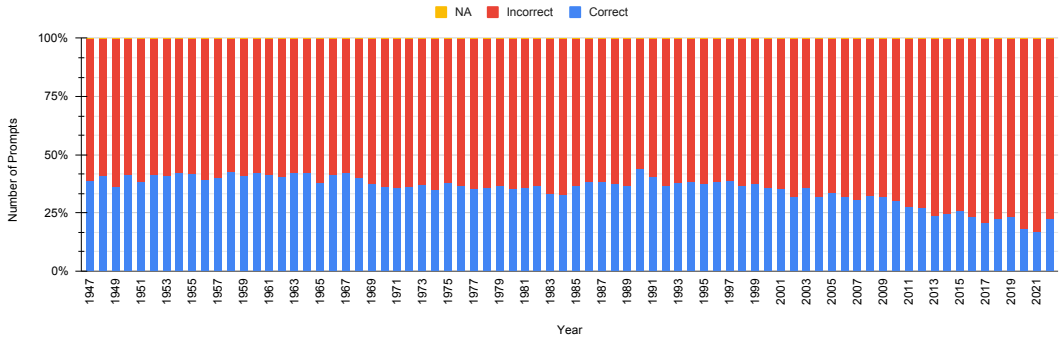

Figure 101: Plot for the Date-based metric (DB) as year-wise count (In percentage) for **continual fine-tuning** for `flan-t5-xl`.

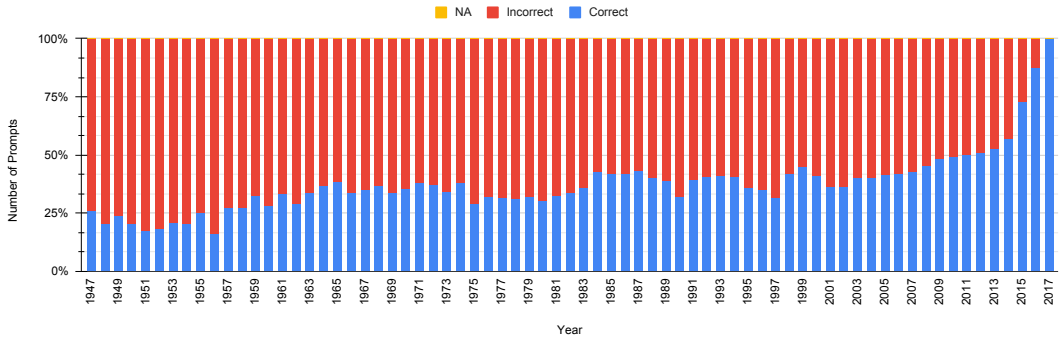

Figure 102: Plot for the Comparative-based metric (CP) as year-wise count (In percentage) for **continual fine-tuning** for `flan-t5-xl`.

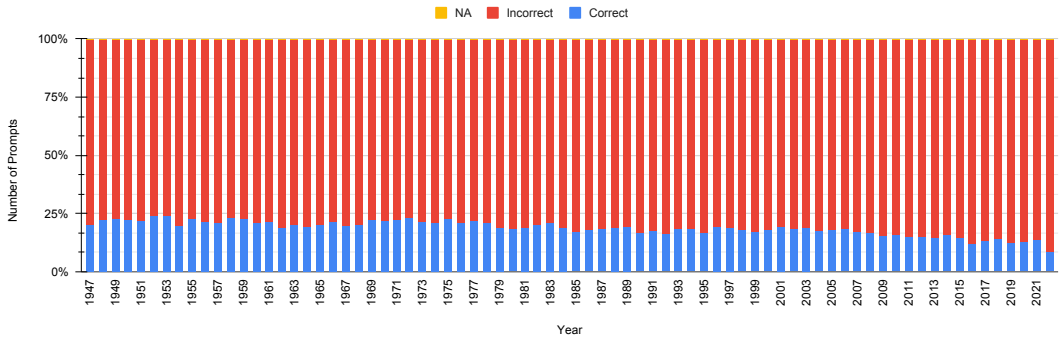

Figure 103: Plot for the Window-based metric (WB) as year-wise count (In percentage) for **continual fine-tuning** for `flan-t5-xl`.

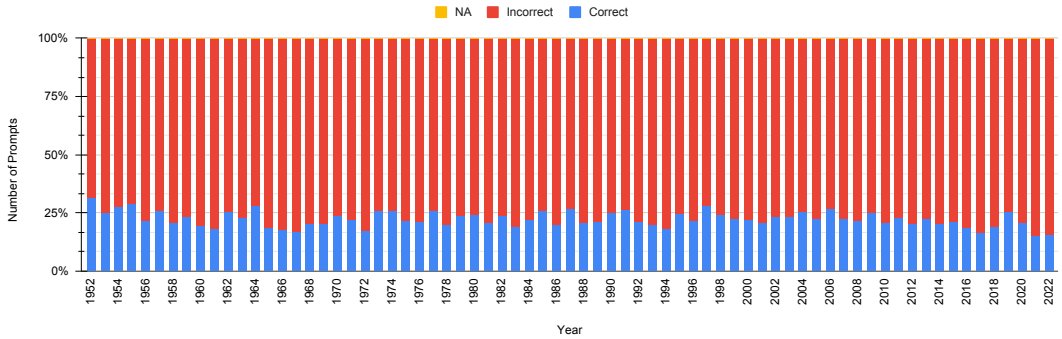

Figure 104: Plot for the Min/Max-based metric (MM) as year-wise count (In percentage) for **continual fine-tuning** for `flan-t5-xl`.

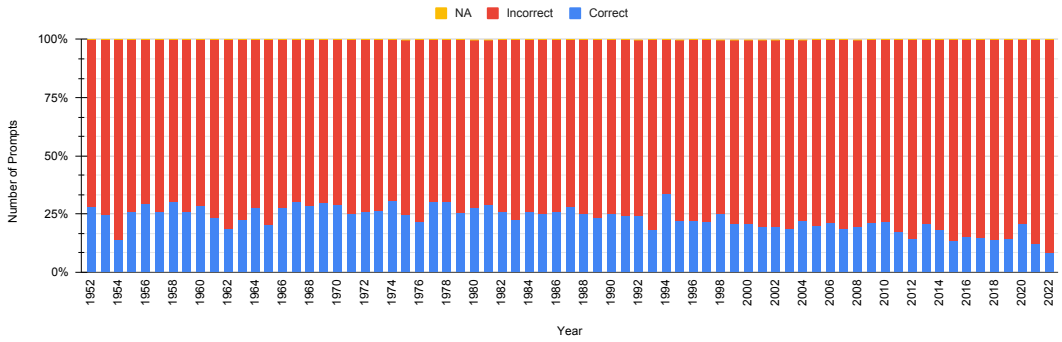

Figure 105: Plot for the Range-based metric (RB) as year-wise count (In percentage) for **continual fine-tuning** for `flan-t5-xl`.

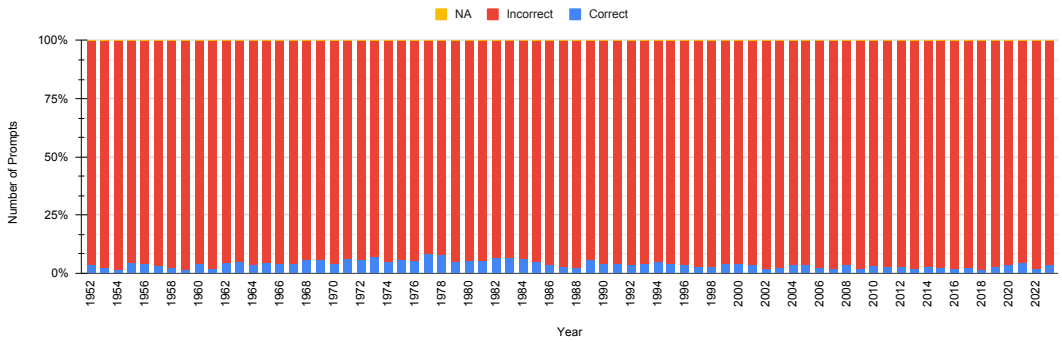

Figure 106: Plot for the Trend-based metric (TB) as year-wise count (In percentage) for **continual fine-tuning** for `flan-t5-xl`.

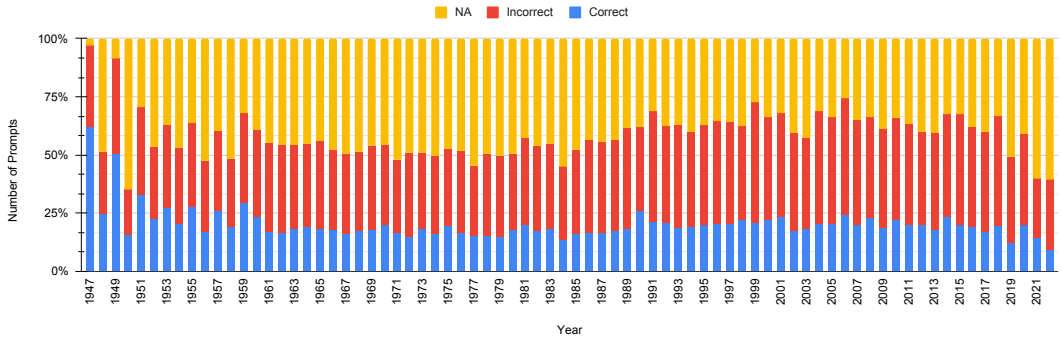

Figure 107: Plot for the Date-based metric (DB) as year-wise count (In percentage) for **continual fine-tuning** for `mistral-instruct`.

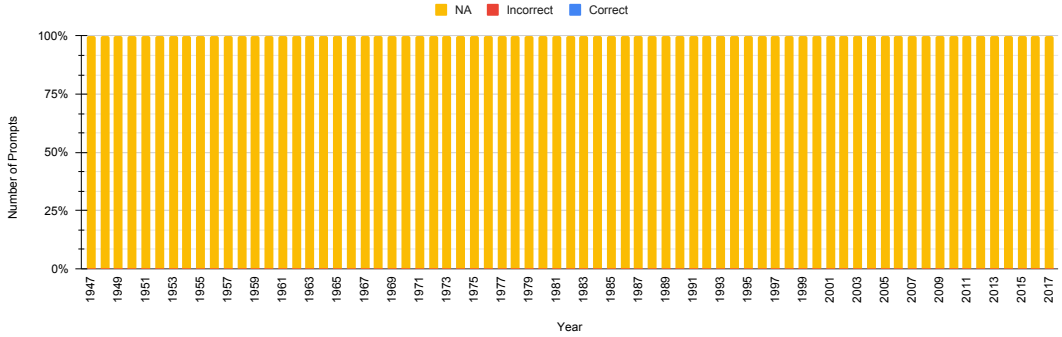

Figure 108: Plot for the Comparative-based metric (CP) as year-wise count (In percentage) for **continual fine-tuning** for `mistral-instruct`.

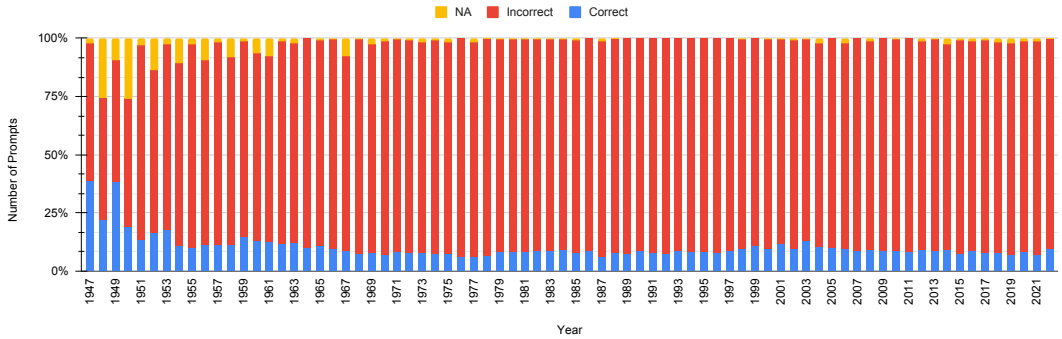

Figure 109: Plot for the Window-based metric (WB) as year-wise count (In percentage) for **continual fine-tuning** for `mistral-instruct`.

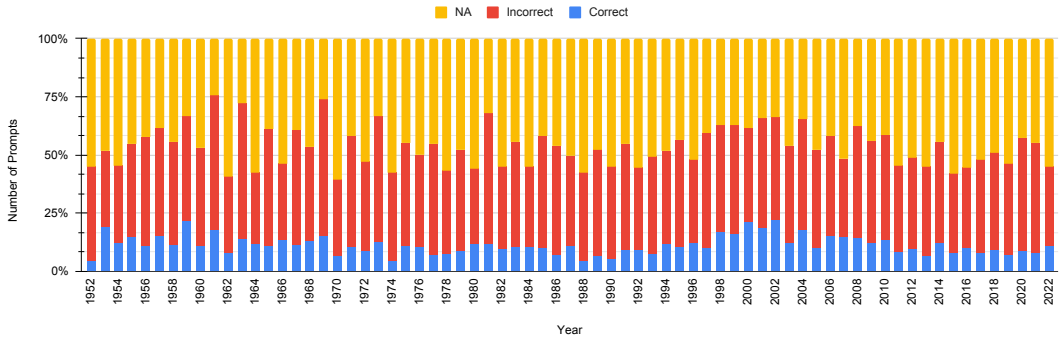

Figure 110: Plot for the Min/Max-based metric (MM) as year-wise count (In percentage) for **continual fine-tuning** for `mistral-instruct`.

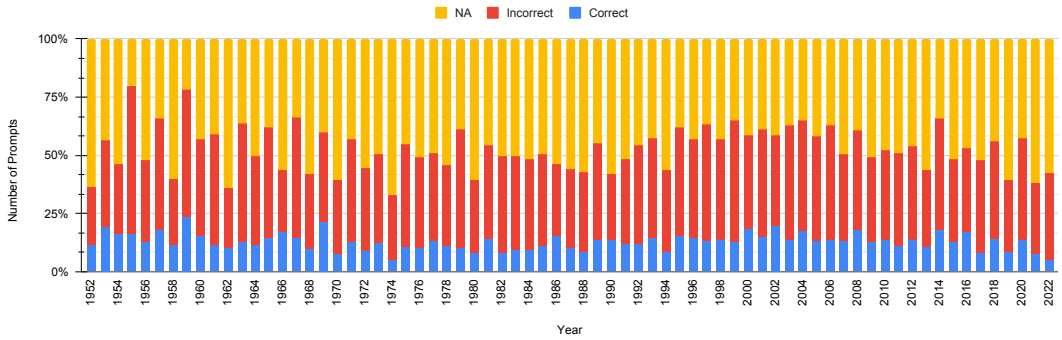

Figure 111: Plot for the Range-based metric (RB) as year-wise count (In percentage) for **continual fine-tuning** for `mistral-instruct`.

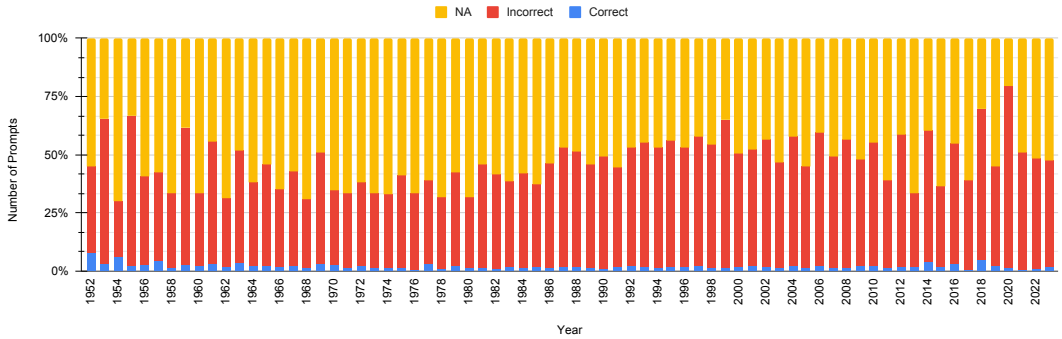

Figure 112: Plot for the Trend-based metric (TB) as year-wise count (In percentage) for **continual fine-tuning** for `mistral-instruct`.

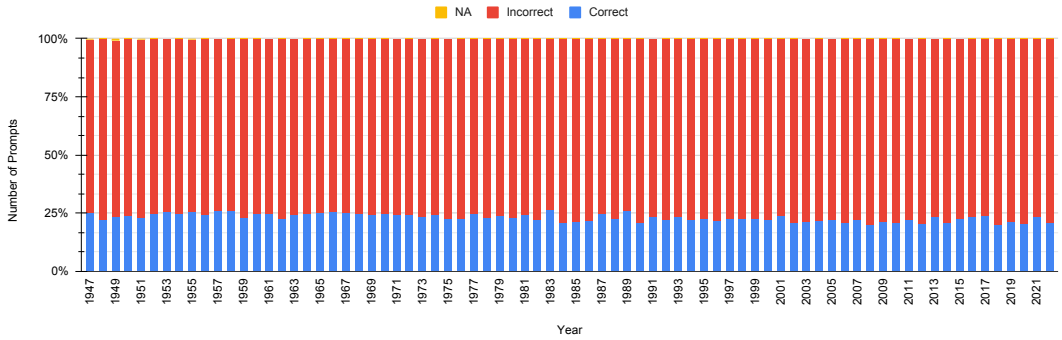

Figure 113: Plot for the Date-based metric (DB) as year-wise count (In percentage) for **continual fine-tuning** for `llama-2`.

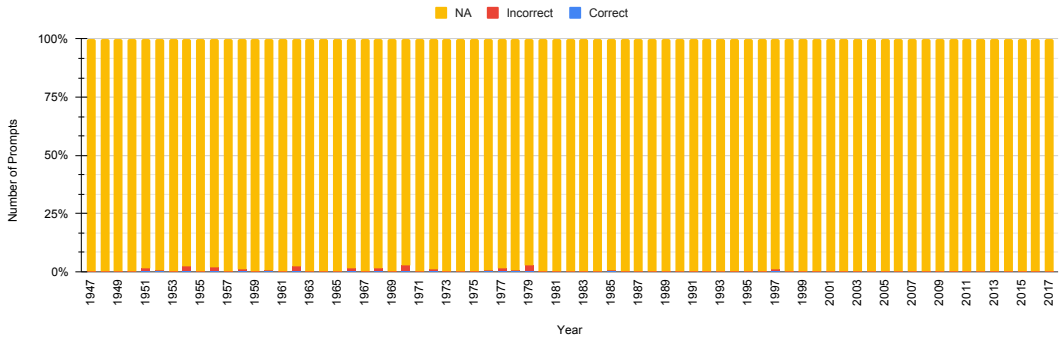

Figure 114: Plot for the Comparative-based metric (CP) as year-wise count (In percentage) for **continual fine-tuning** for `llama-2`.

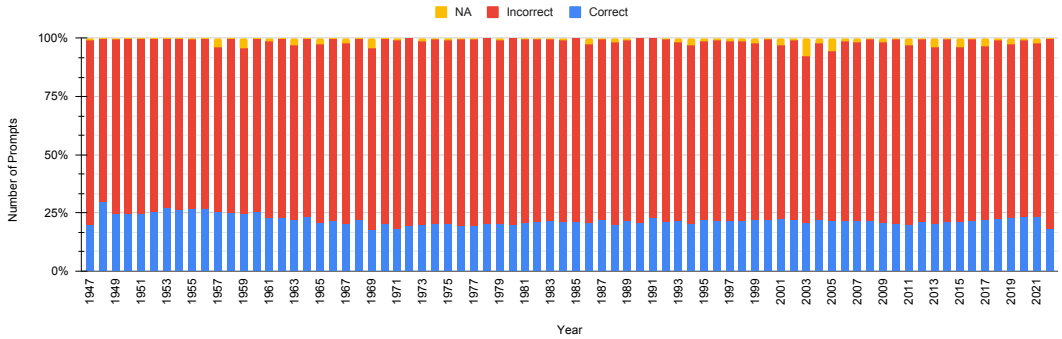

Figure 115: Plot for the Window-based metric (WB) as year-wise count (In percentage) for **continual fine-tuning** for `llama-2`.

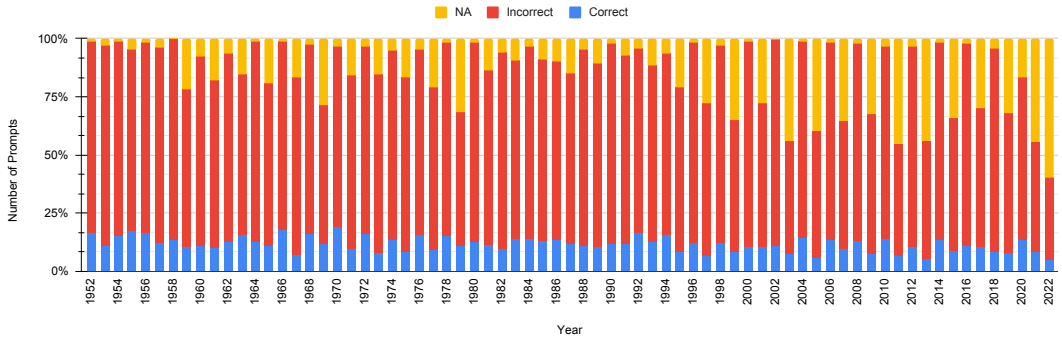

Figure 116: Plot for the Min/Max-based metric (MM) as year-wise count (In percentage) for **continual fine-tuning** for `llama-2`.

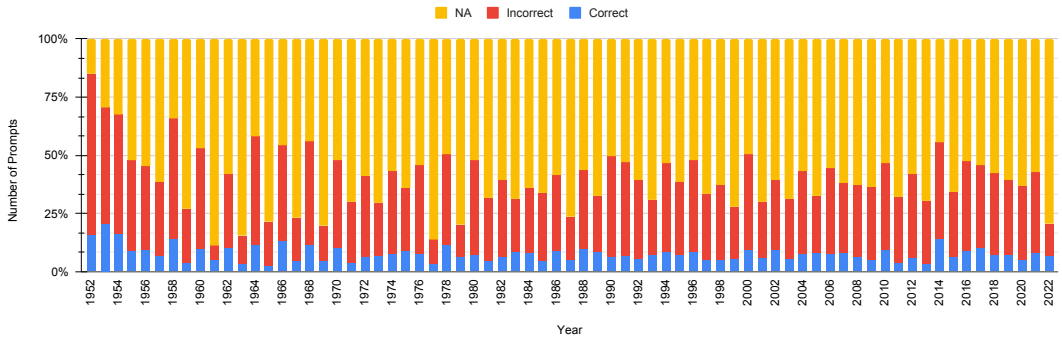

Figure 117: Plot for the Range-based metric (RB) as year-wise count (In percentage) for **continual fine-tuning** for `llama-2`.

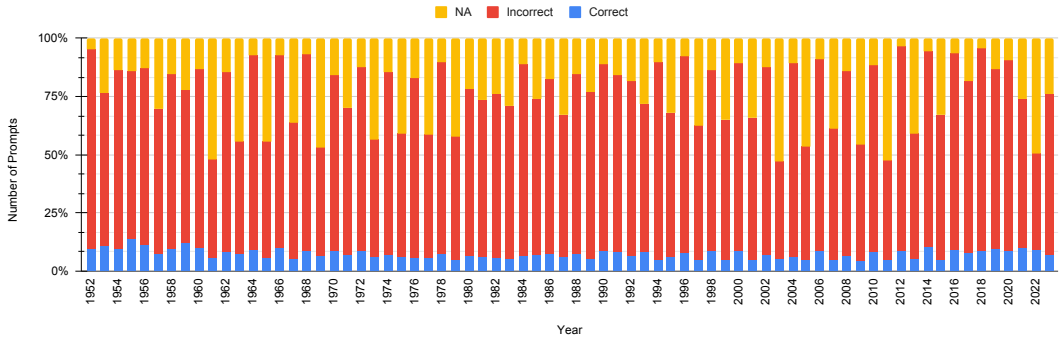

Figure 118: Plot for the Trend-based metric (TB) as year-wise count (In percentage) for **continual fine-tuning** for `llama-2`.

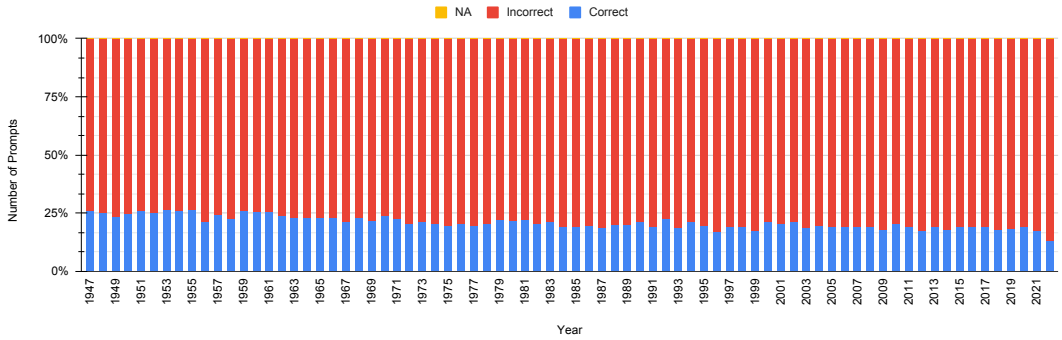

Figure 119: Plot for the Date-based metric (DB) as year-wise count (In percentage) for **continual fine-tuning** for `gemma-7b-it`.

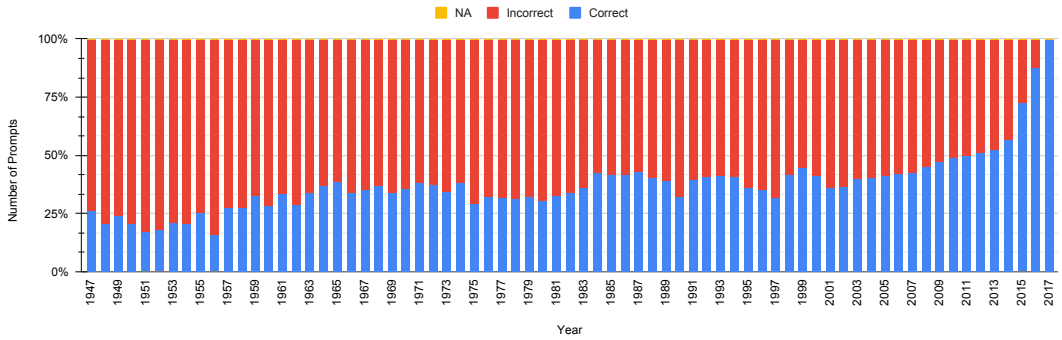

Figure 120: Plot for the Comparative-based metric (CP) as year-wise count (In percentage) for **continual fine-tuning** for `gemma-7b-it`.

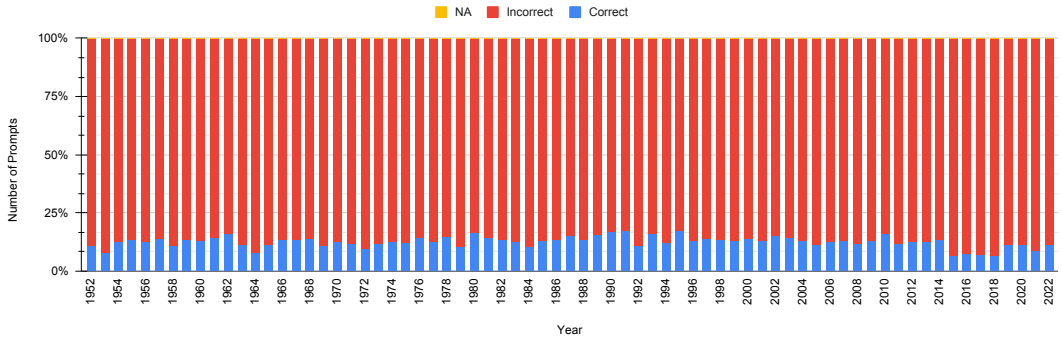

Figure 121: Plot for the Window-based metric (WB) as year-wise count (In percentage) for **continual fine-tuning** for `gemma-7b-it`.

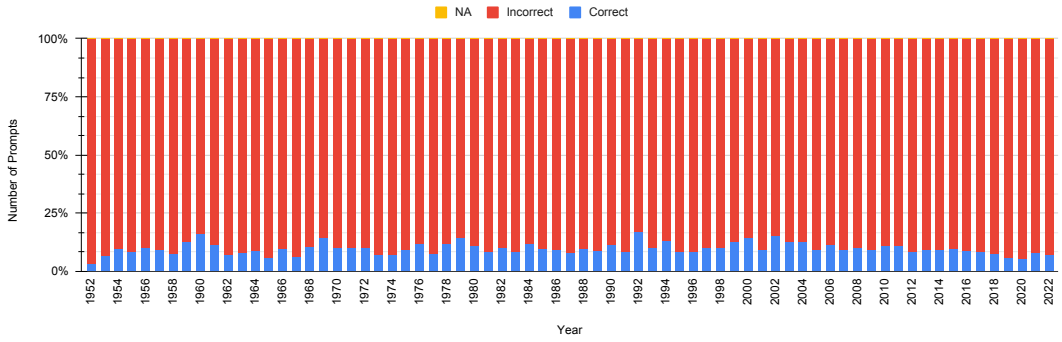

Figure 122: Plot for the Min/Max-based metric (MM) as year-wise count (In percentage) for **continual fine-tuning** for `gemma-7b-it`.

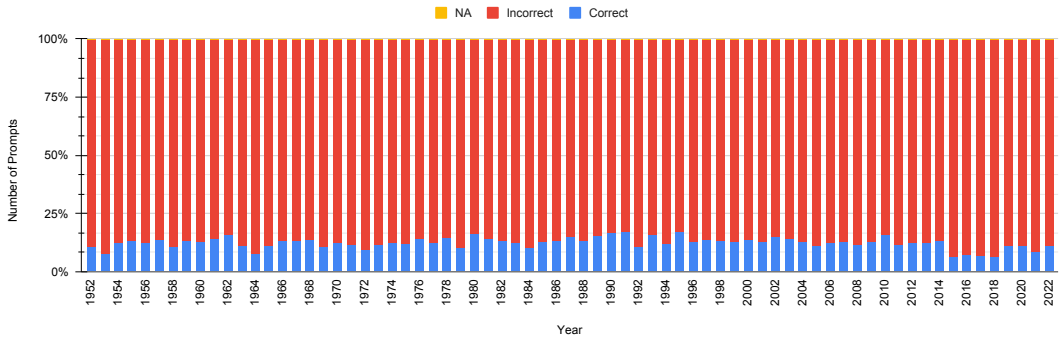

Figure 123: Plot for the Range-based metric (RB) as year-wise count (In percentage) for **continual fine-tuning** for `gemma-7b-it`.

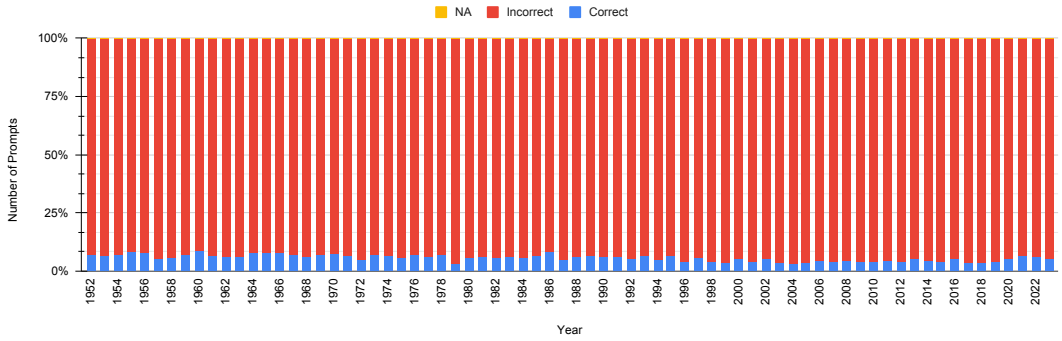

Figure 124: Plot for the Trend-based metric (TB) as year-wise count (In percentage) for **continual fine-tuning** for `gemma-7b-it`.

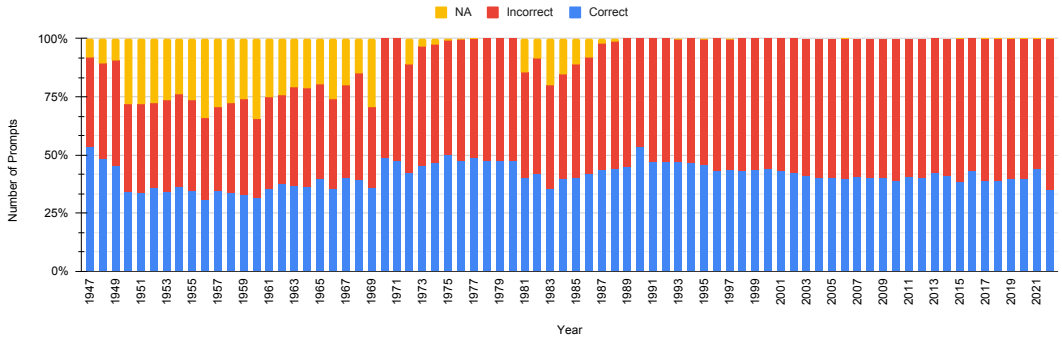

Figure 125: Plot for the Date-based metric (DB) as year-wise count (In percentage) for **continual fine-tuning** for `llama-3-8b`.

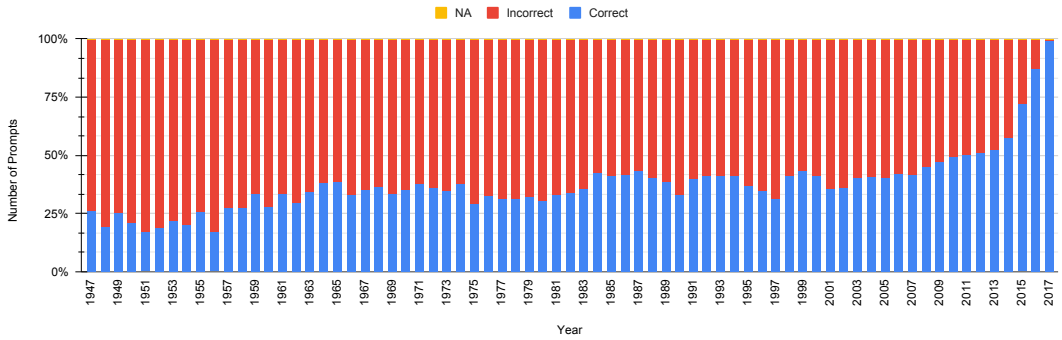

Figure 126: Plot for the Comparative-based metric (CP) as year-wise count (In percentage) for **continual fine-tuning** for `llama-3-8b`.

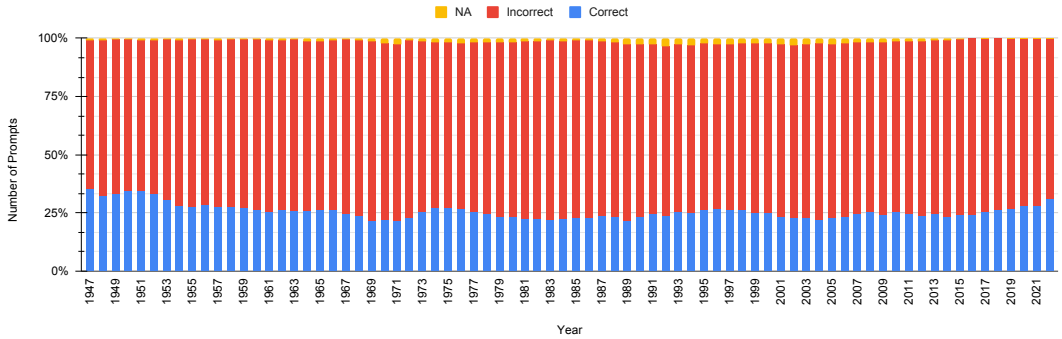

Figure 127: Plot for the Window-based metric (WB) as year-wise count (In percentage) for **continual fine-tuning** for `llama-3-8b`.

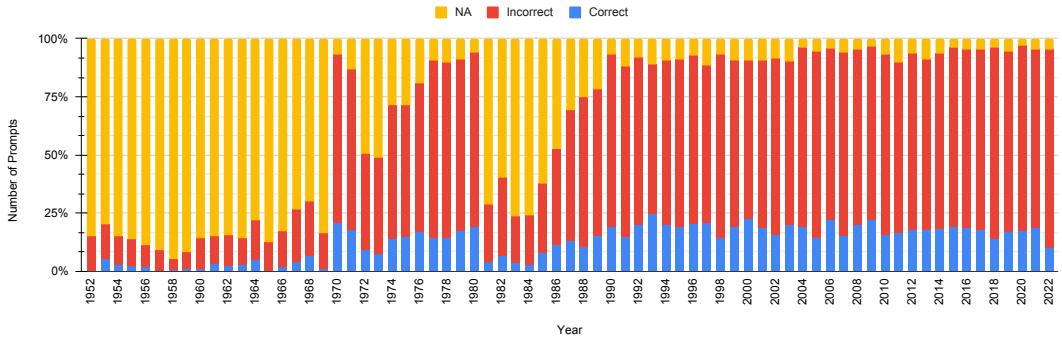

Figure 128: Plot for the Min/Max-based metric (MM) as year-wise count (In percentage) for **continual fine-tuning** for `llama-3-8b`.

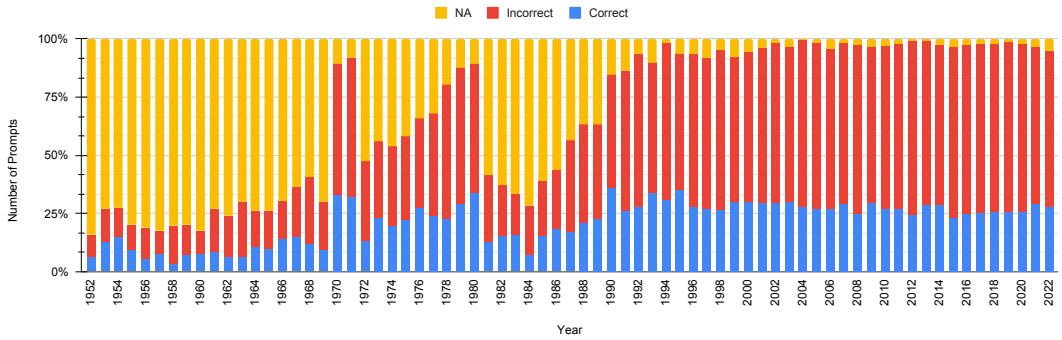

Figure 129: Plot for the Range-based metric (RB) as year-wise count (In percentage) for **continual fine-tuning** for `llama-3-8b`.

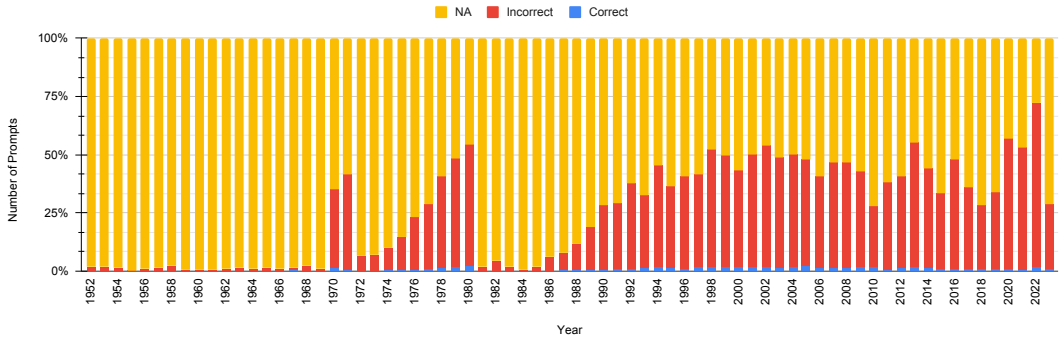

Figure 130: Plot for the Trend-based metric (TB) as year-wise count (In percentage) for **continual fine-tuning** for `llama-3-8b`.

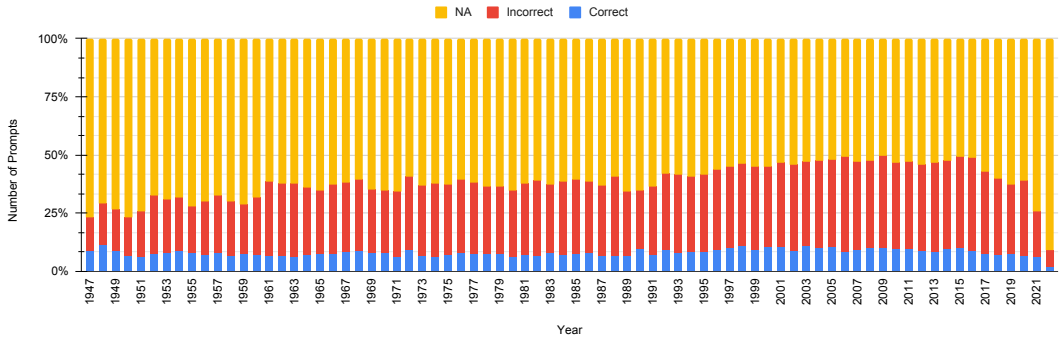

Figure 131: Plot for the Date-based metric (DB) as year-wise count (In percentage) for **continual fine-tuning** for `phi-3-medium`.

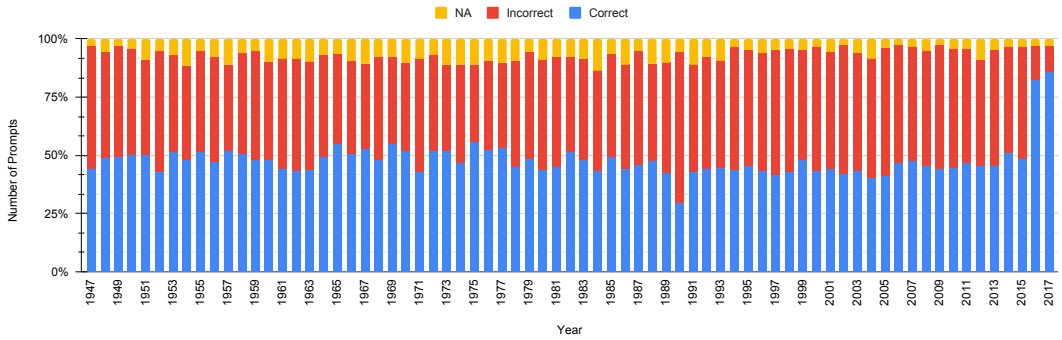

Figure 132: Plot for the Comparative-based metric (CP) as year-wise count (In percentage) for **continual fine-tuning** for `phi-3-medium`.

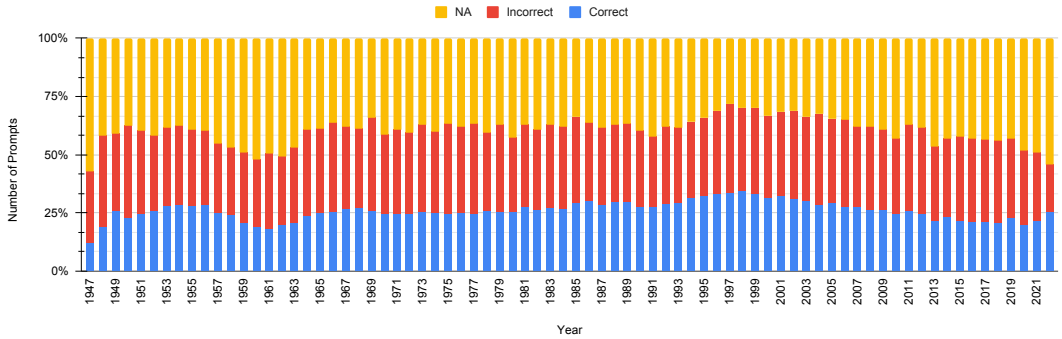

Figure 133: Plot for the Window-based metric (WB) as year-wise count (In percentage) for **continual fine-tuning** for `phi-3-medium`.

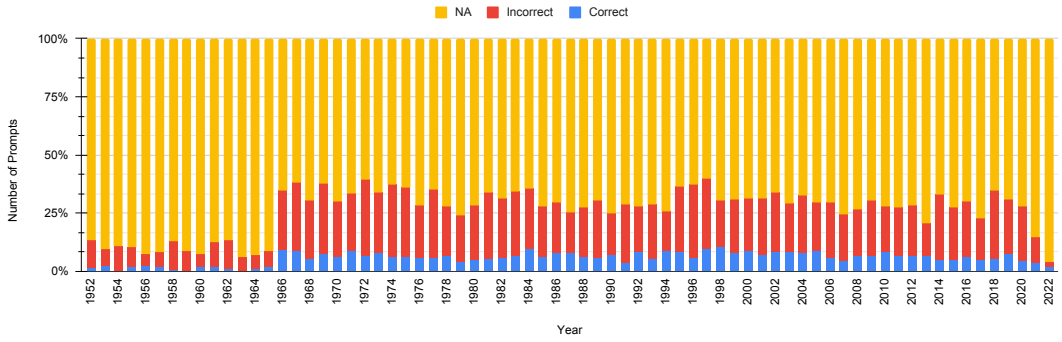

Figure 134: Plot for the Min/Max-based metric (MM) as year-wise count (In percentage) for **continual fine-tuning** for `phi-3-medium`.

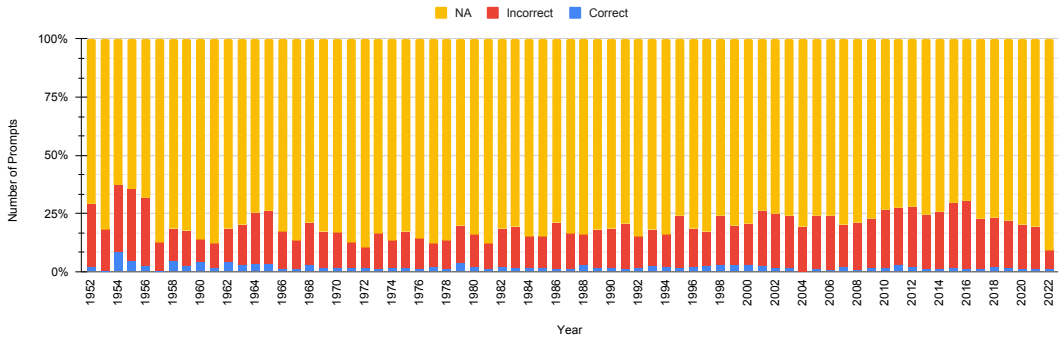

Figure 135: Plot for the Range-based metric (RB) as year-wise count (In percentage) for **continual fine-tuning** for `phi-3-medium`.

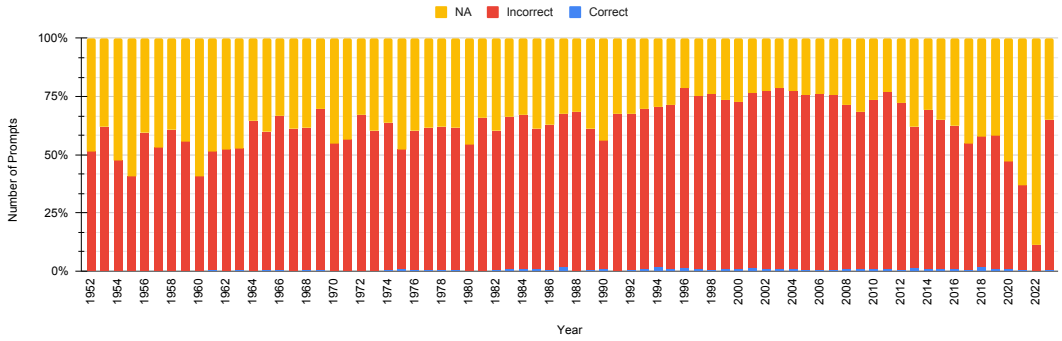

Figure 136: Plot for the Trend-based metric (TB) as year-wise count (In percentage) for **continual fine-tuning** for `phi-3-medium`.

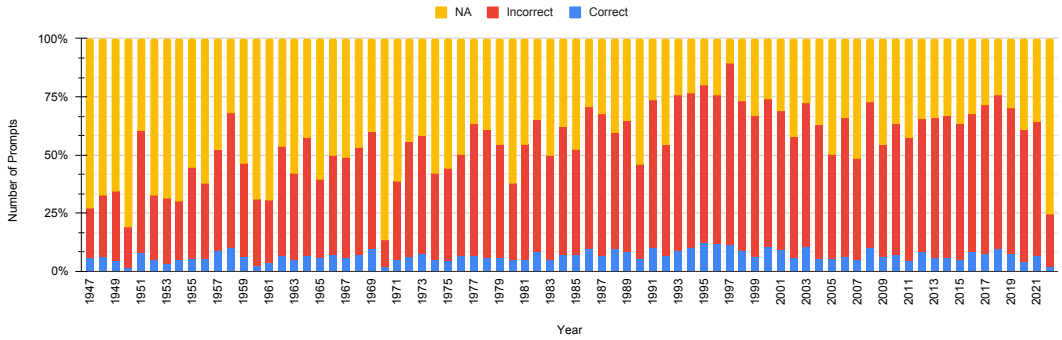

Figure 137: Plot for the Date-based metric (DB) as year-wise count (In percentage) for **yearwise fine-tuning** for `phi-2`.

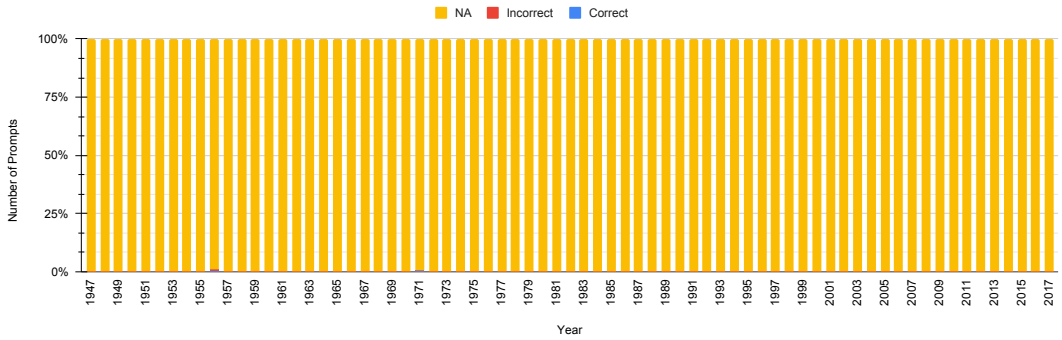

Figure 138: Plot for the Comparative-based metric (CP) as year-wise count (In percentage) for **yearwise fine-tuning** for `phi-2`.

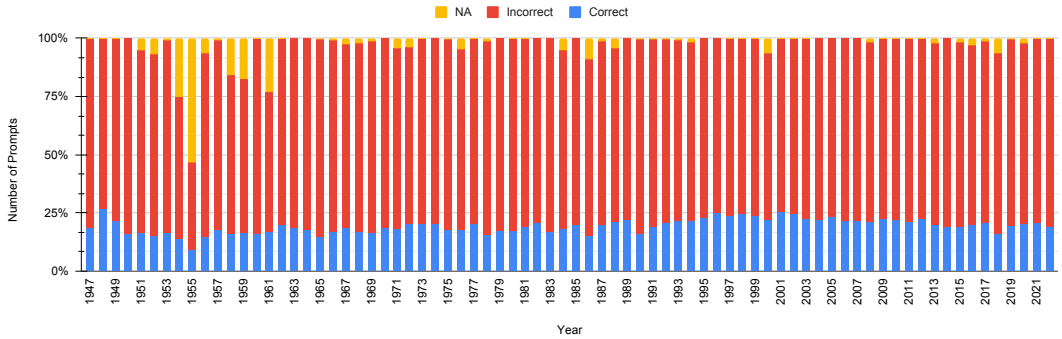

Figure 139: Plot for the Window-based metric (WB) as year-wise count (In percentage) for **yearwise fine-tuning** for `phi-2`.

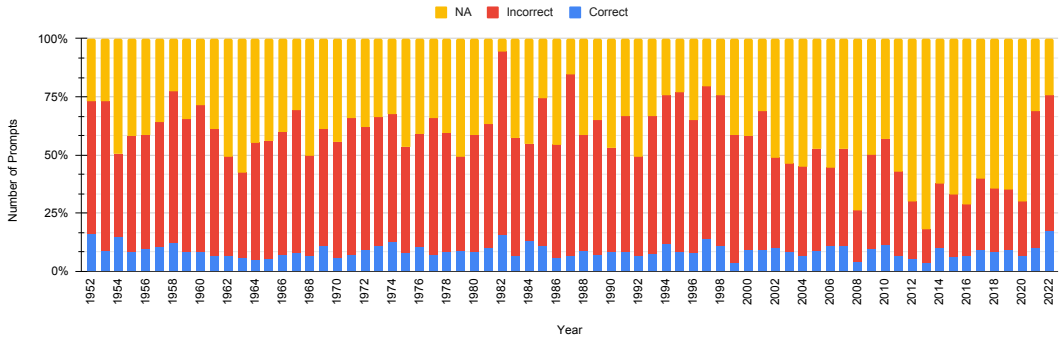

Figure 140: Plot for the Min/Max-based metric (MM) as year-wise count (In percentage) for **yearwise fine-tuning** for `phi-2`.

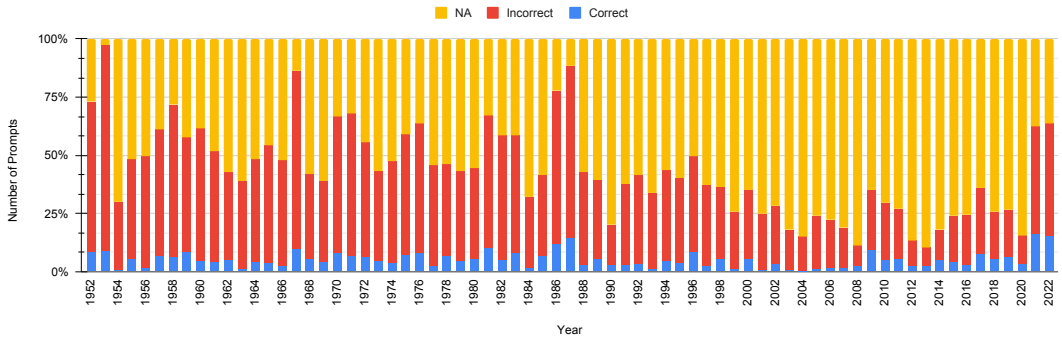

Figure 141: Plot for the Range-based metric (RB) as year-wise count (In percentage) for **yearwise fine-tuning** for `phi-2`.

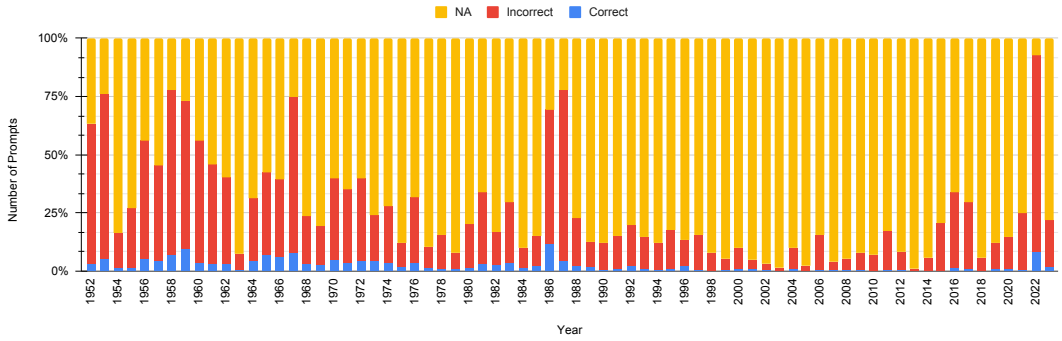

Figure 142: Plot for the Trend-based metric (TB) as year-wise count (In percentage) for **yearwise fine-tuning** for `phi-2`.

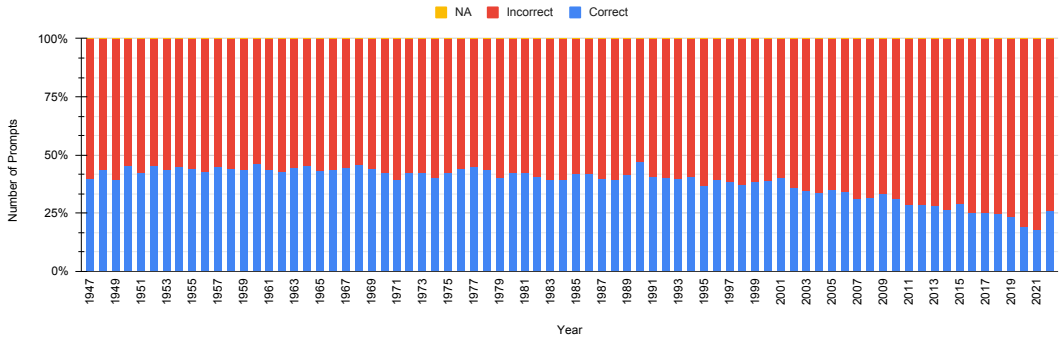

Figure 143: Plot for the Date-based metric (DB) as year-wise count (In percentage) for **yearwise fine-tuning** for `flan-t5-xl`.

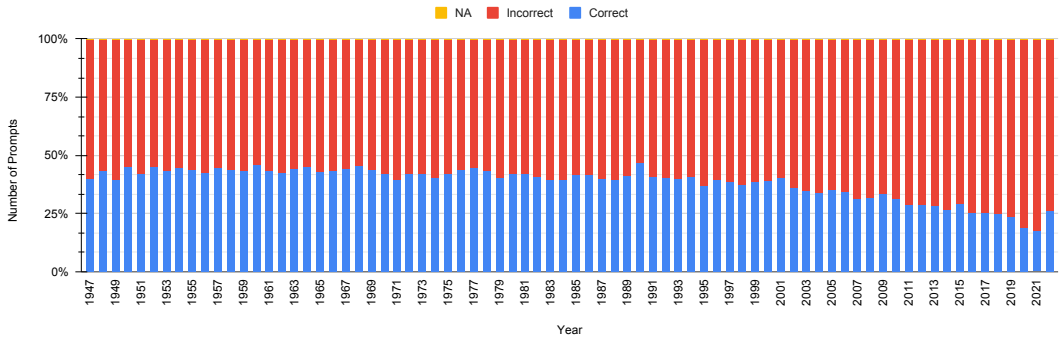

Figure 144: Plot for the Comparative-based metric (CP) as year-wise count (In percentage) for **yearwise fine-tuning** for `flan-t5-xl`.

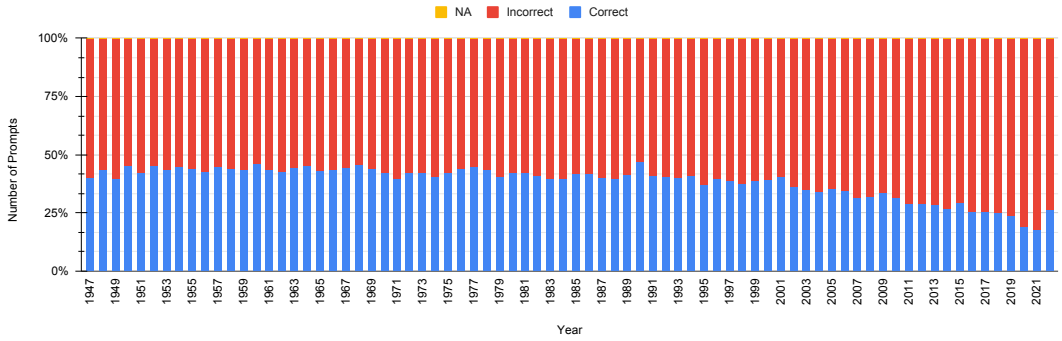

Figure 145: Plot for the Window-based metric (WB) as year-wise count (In percentage) for **yearwise fine-tuning** for `flan-t5-xl`.

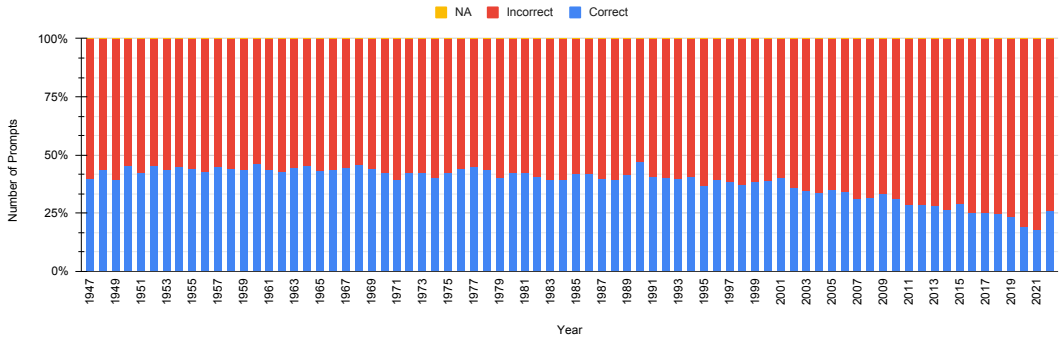

Figure 146: Plot for the Min/Max-based metric (MM) as year-wise count (In percentage) for **yearwise fine-tuning** for `flan-t5-xl`.

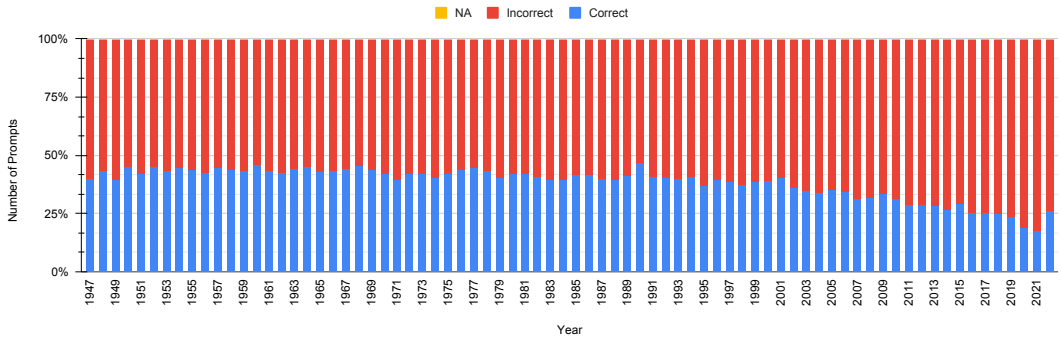

Figure 147: Plot for the Range-based metric (RB) as year-wise count (In percentage) for **yearwise fine-tuning** for `flan-t5-xl`.

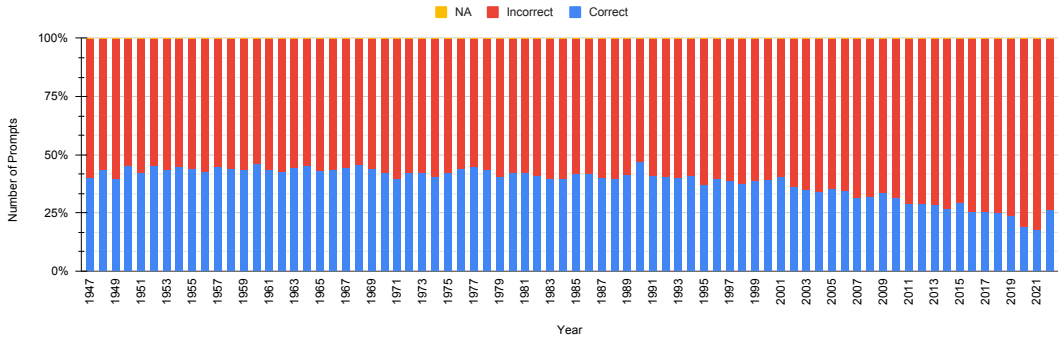

Figure 148: Plot for the Trend-based metric (TB) as year-wise count (In percentage) for **yearwise fine-tuning** for `flan-t5-xl`.

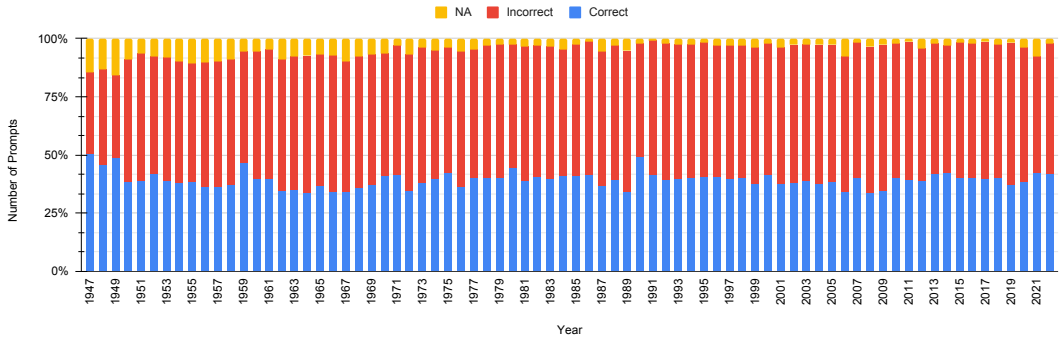

Figure 149: Plot for the Date-based metric (DB) as year-wise count (In percentage) for **yearwise fine-tuning** for for `mistral-instruct`.

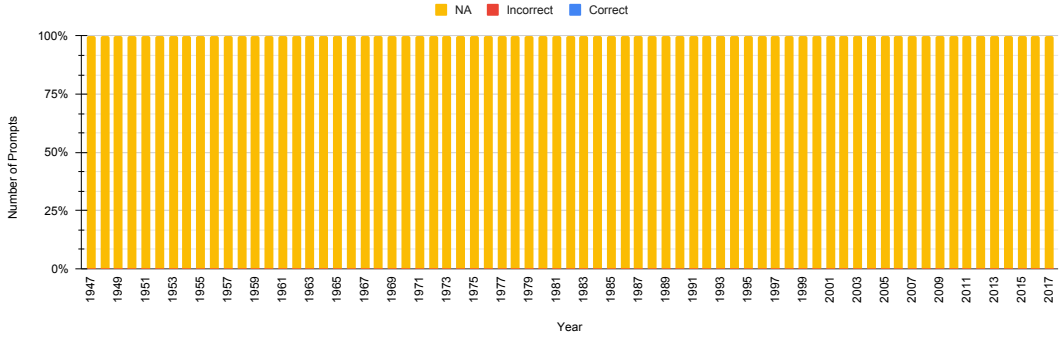

Figure 150: Plot for the Comparative-based metric (CP) as year-wise count (In percentage) for **yearwise fine-tuning** for `mistral-instruct`.

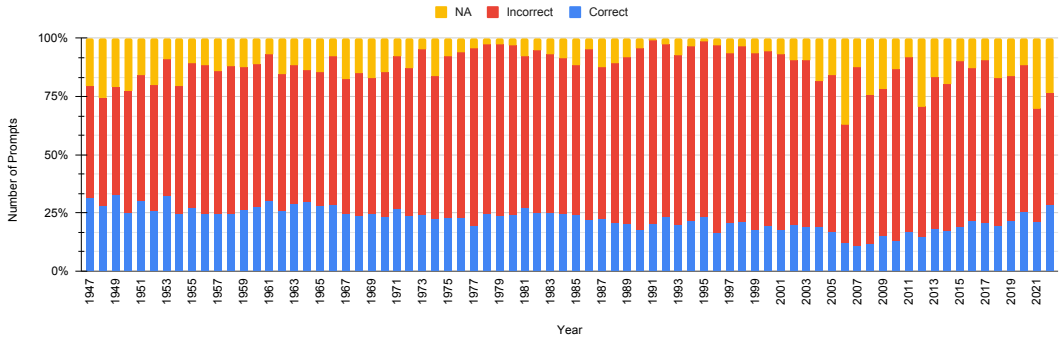

Figure 151: Plot for the Window-based metric (WB) as year-wise count (In percentage) for **yearwise fine-tuning** for `mistral-instruct`.

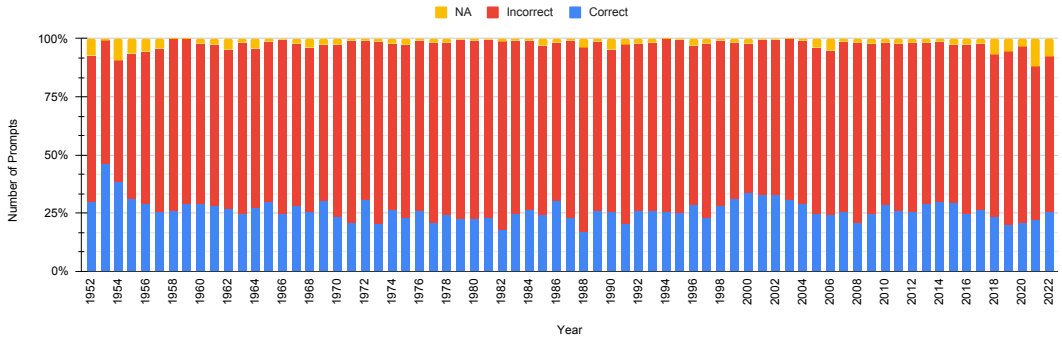

Figure 152: Plot for the Min/Max-based metric (MM) as year-wise count (In percentage) for **yearwise fine-tuning** for `mistral-instruct`.

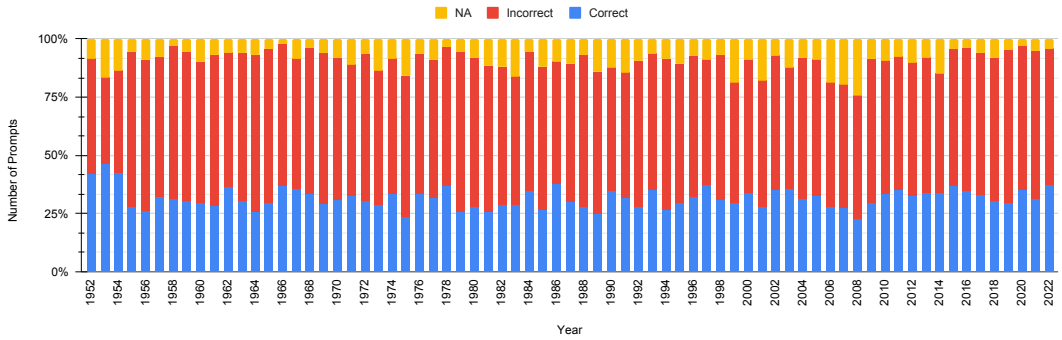

Figure 153: Plot for the Range-based metric (RB) as year-wise count (In percentage) for **yearwise fine-tuning** for `mistral-instruct`.

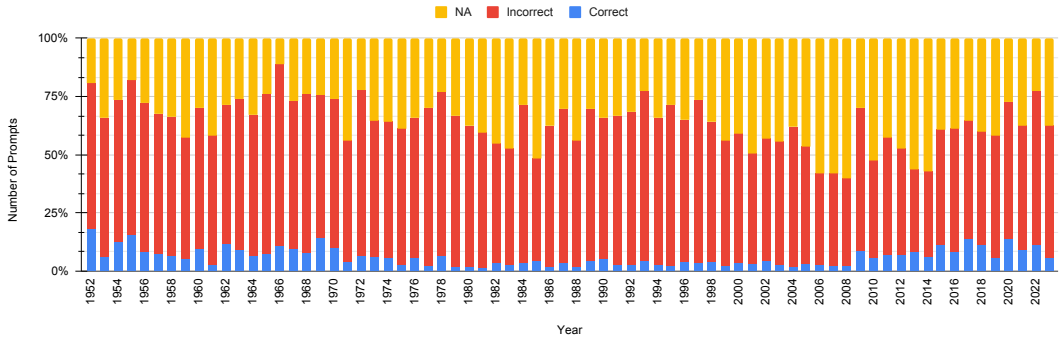

Figure 154: Plot for the Trend-based metric (TB) as year-wise count (In percentage) for **yearwise fine-tuning** for `mistral-instruct`.

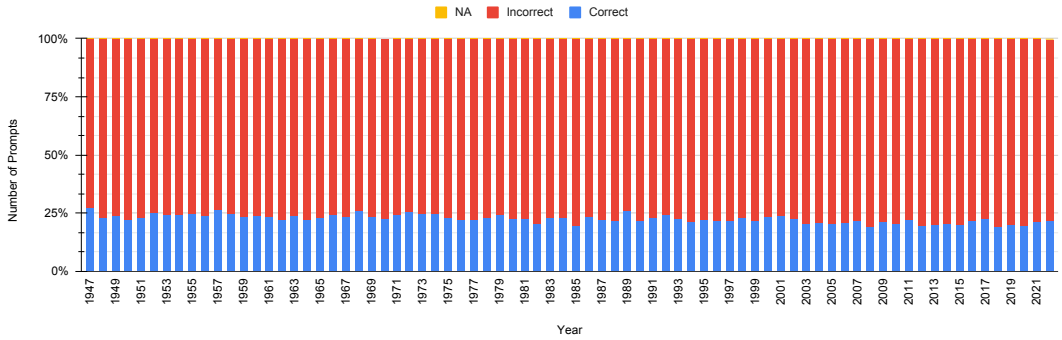

Figure 155: Plot for the Date-based metric (DB) as year-wise count (In percentage) for **yearwise fine-tuning** for `llama-2`.

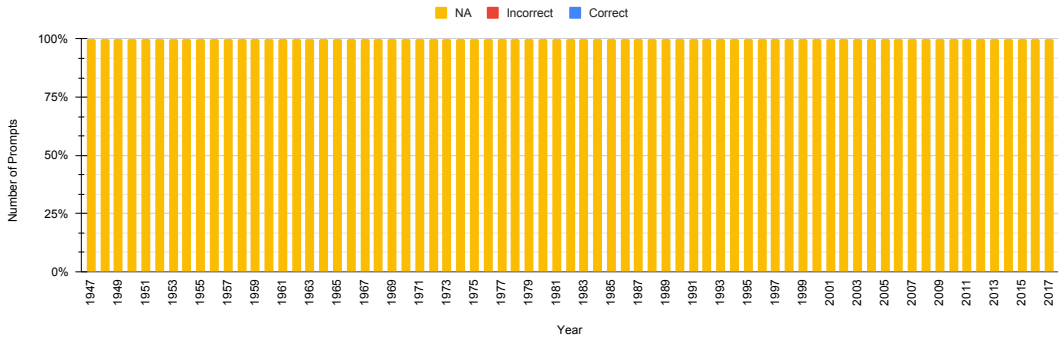

Figure 156: Plot for the Comparative-based metric (CP) as year-wise count (In percentage) for **yearwise fine-tuning** for `llama-2`.

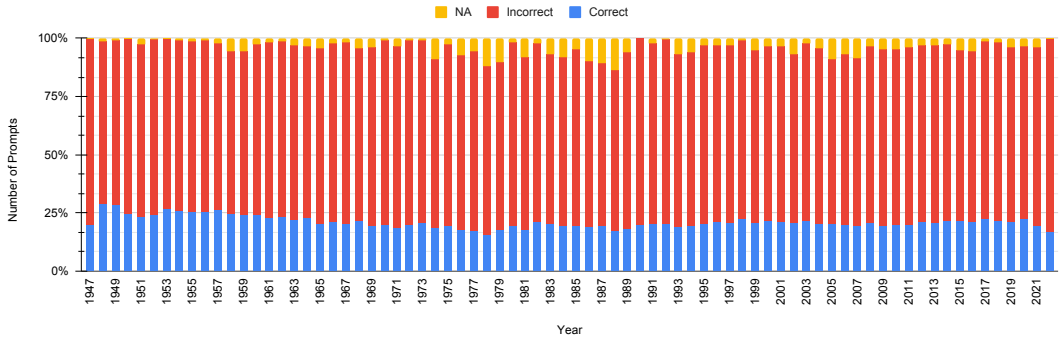

Figure 157: Plot for the Window-based metric (WB) as year-wise count (In percentage) for **yearwise fine-tuning** for `llama-2`.

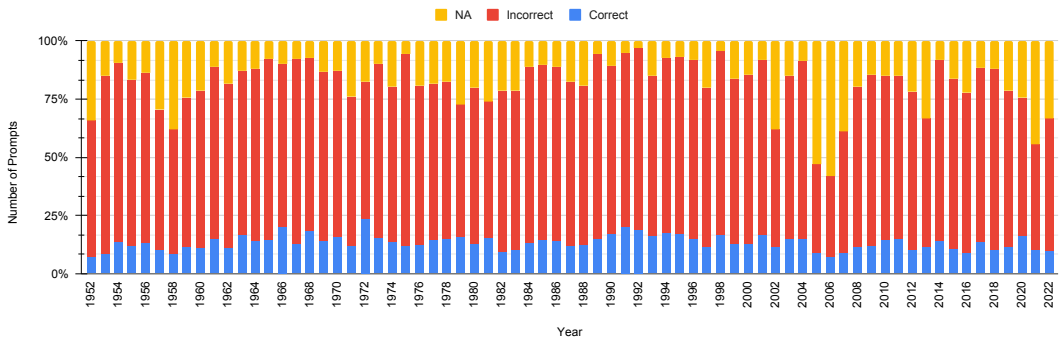

Figure 158: Plot for the Min/Max-based metric (MM) as year-wise count (In percentage) for **yearwise fine-tuning** for `llama-2`.

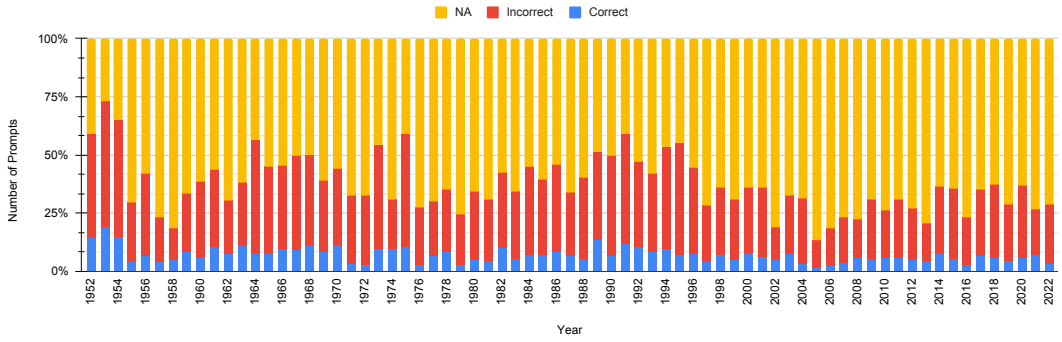

Figure 159: Plot for the Range-based metric (RB) as year-wise count (In percentage) for **yearwise fine-tuning** for `llama-2`.

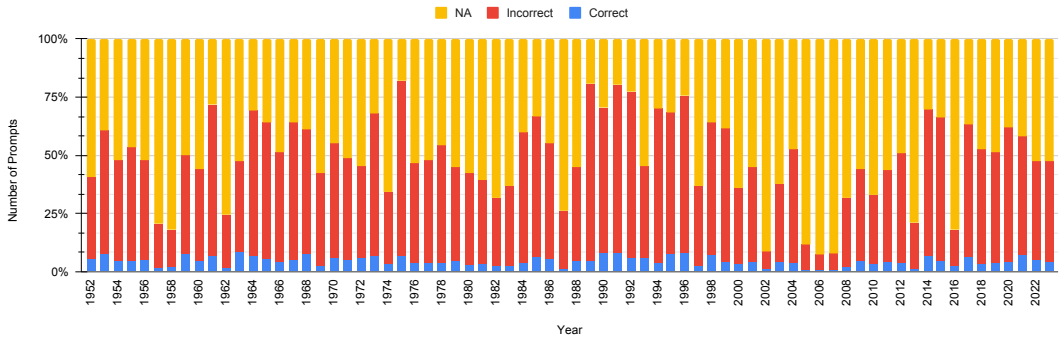

Figure 160: Plot for the Trend-based metric (TB) as year-wise count (In percentage) for **yearwise fine-tuning** for `llama-2`.

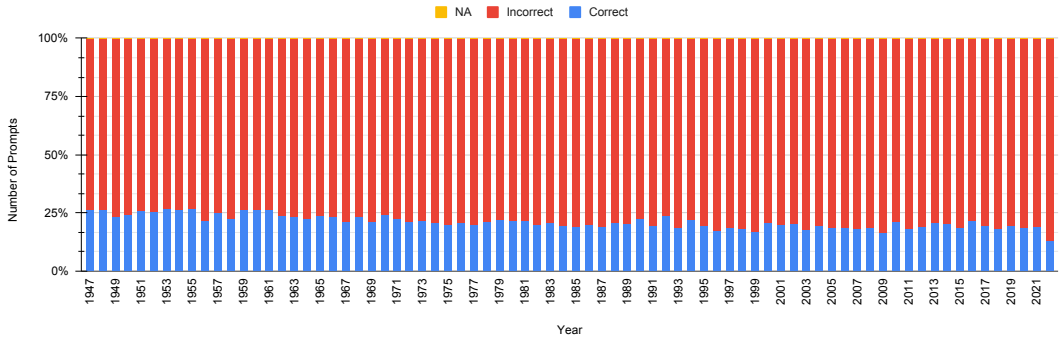

Figure 161: Plot for the Date-based metric (DB) as year-wise count (In percentage) for **yearwise fine-tuning** for `gemma-7b-it`.

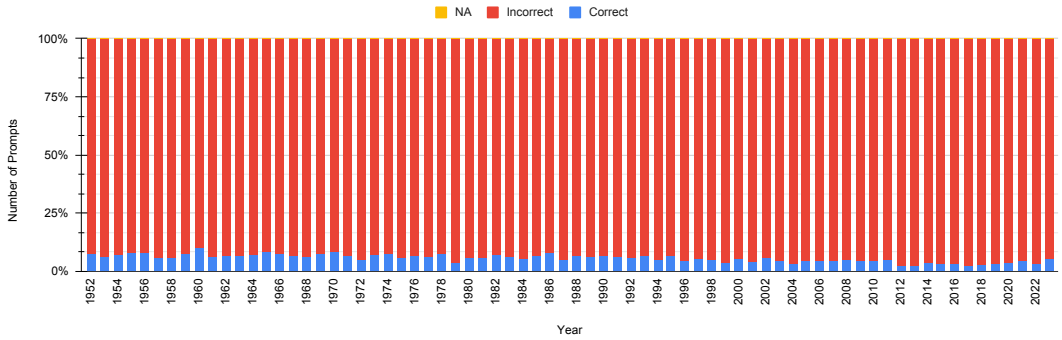

Figure 162: Plot for the Comparative-based metric (CP) as year-wise count (In percentage) for **yearwise fine-tuning** for `gemma-7b-it`.

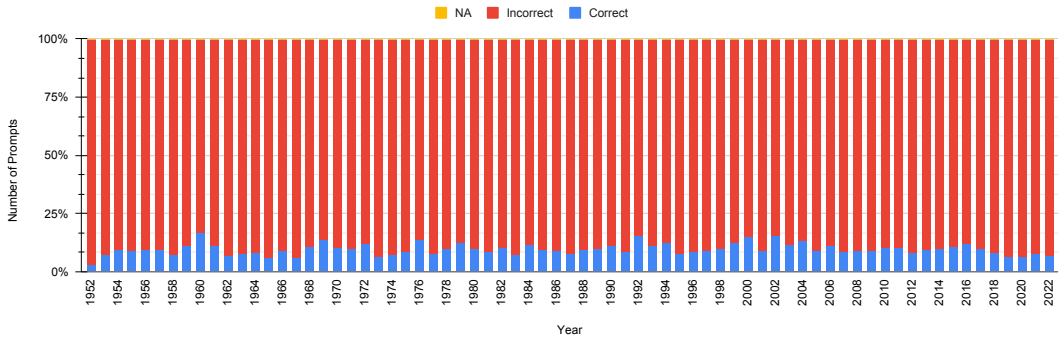

Figure 163: Plot for the Window-based metric (WB) as year-wise count (In percentage) for **yearwise fine-tuning** for `gemma-7b-it`.

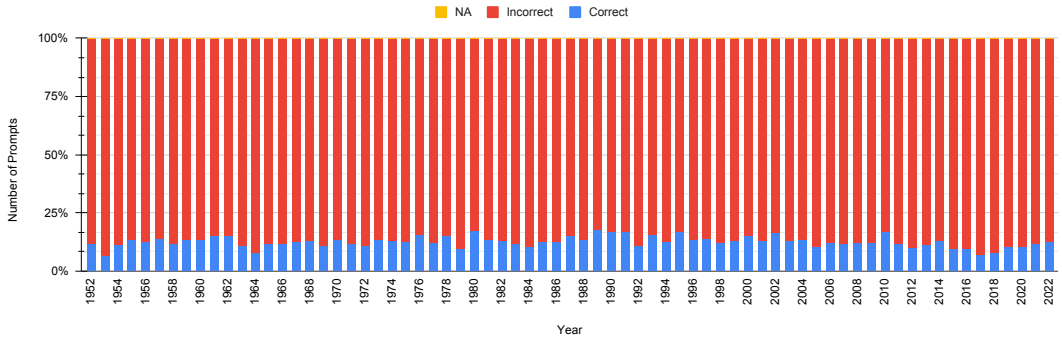

Figure 164: Plot for the Min/Max-based metric (MM) as year-wise count (In percentage) for **yearwise fine-tuning** for `gemma-7b-it`.

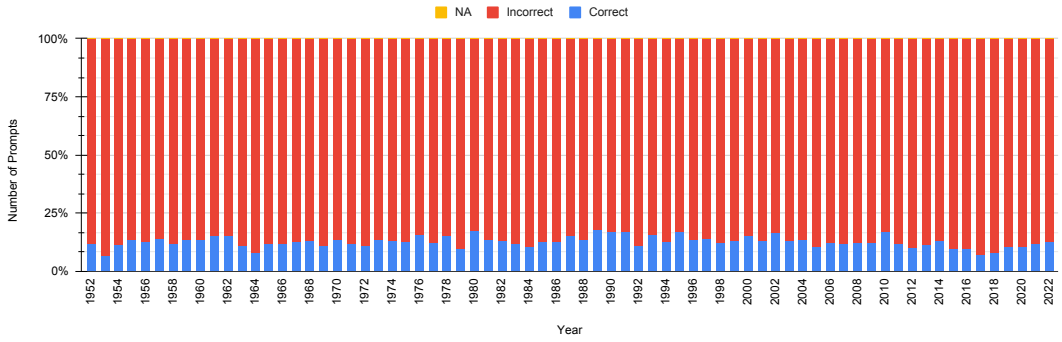

Figure 165: Plot for the Range-based metric (RB) as year-wise count (In percentage) for **yearwise fine-tuning** for `gemma-7b-it`.

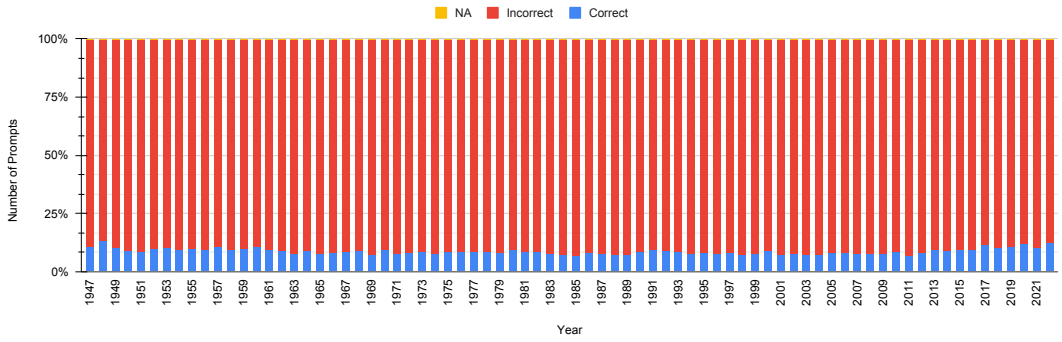

Figure 166: Plot for the Trend-based metric (TB) as year-wise count (In percentage) for **yearwise fine-tuning** for `gemma-7b-it`.

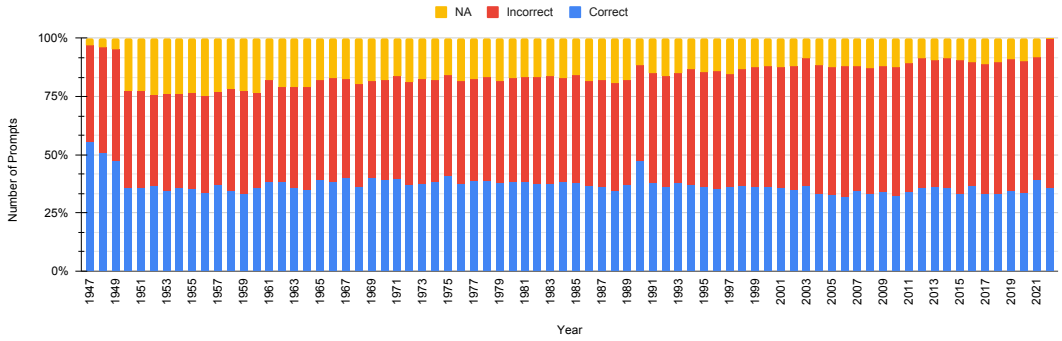

Figure 167: Plot for the Date-based metric (DB) as year-wise count (In percentage) for **yearwise fine-tuning** for `llama-3-8b`.

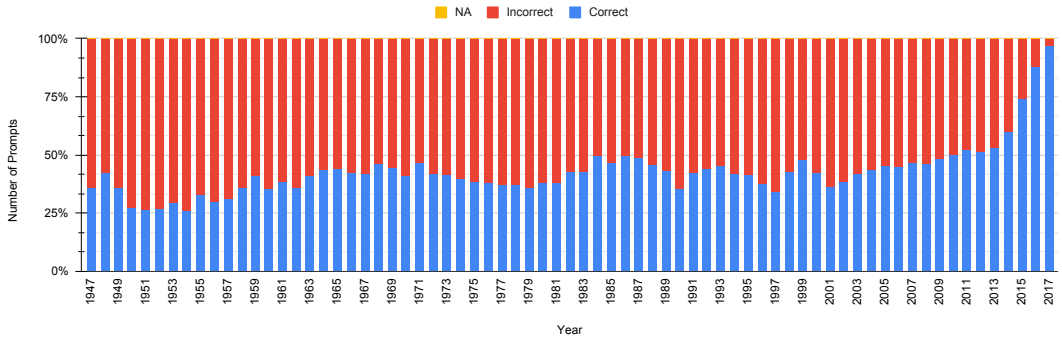

Figure 168: Plot for the Comparative-based metric (CP) as year-wise count (In percentage) for **yearwise fine-tuning** for `llama-3-8b`.

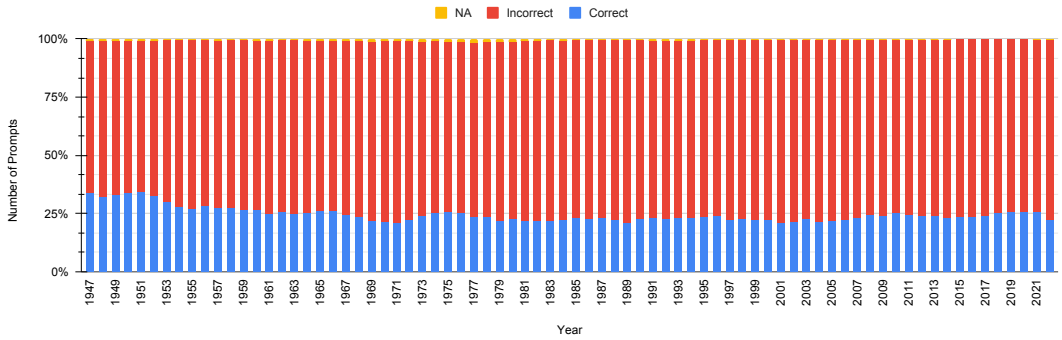

Figure 169: Plot for the Window-based metric (WB) as year-wise count (In percentage) for **yearwise fine-tuning** for `llama-3-8b`.

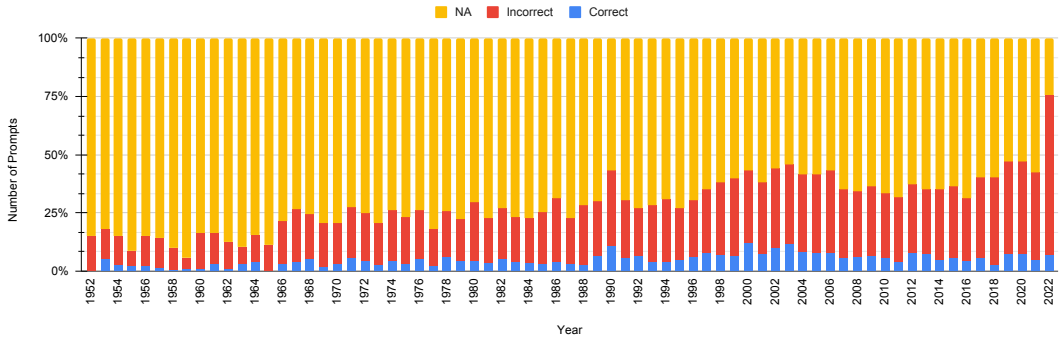

Figure 170: Plot for the Min/Max-based metric (MM) as year-wise count (In percentage) for **yearwise fine-tuning** for `llama-3-8b`.

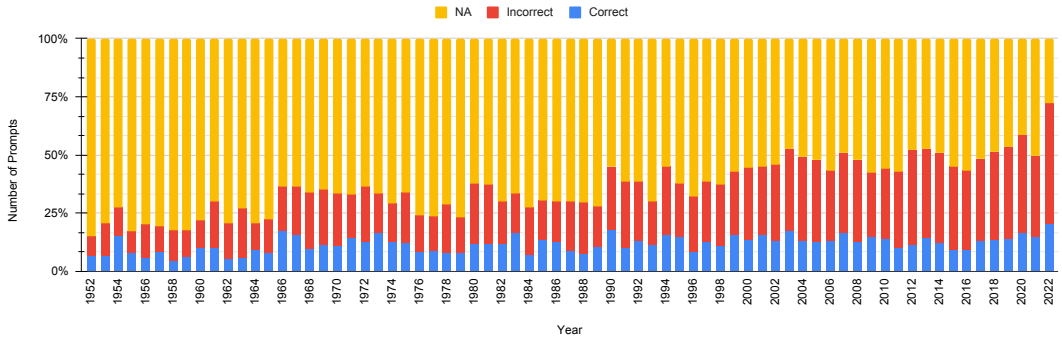

Figure 171: Plot for the Range-based metric (RB) as year-wise count (In percentage) for **yearwise fine-tuning** for `llama-3-8b`.

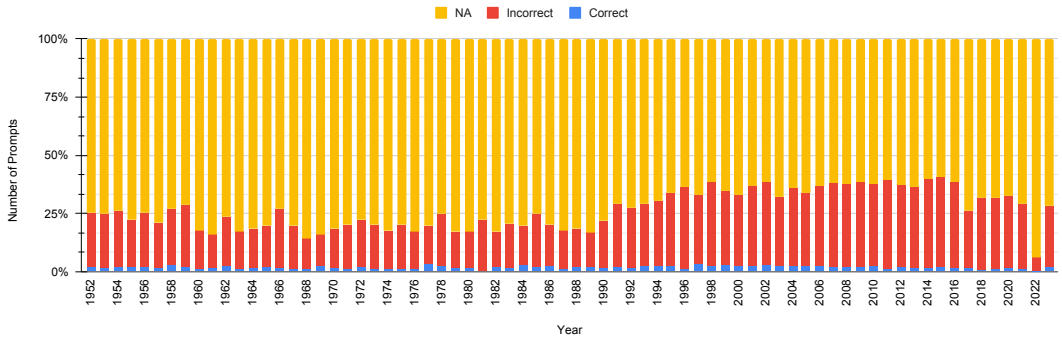

Figure 172: Plot for the Trend-based metric (TB) as year-wise count (In percentage) for **yearwise fine-tuning** for `llama-3-8b`.

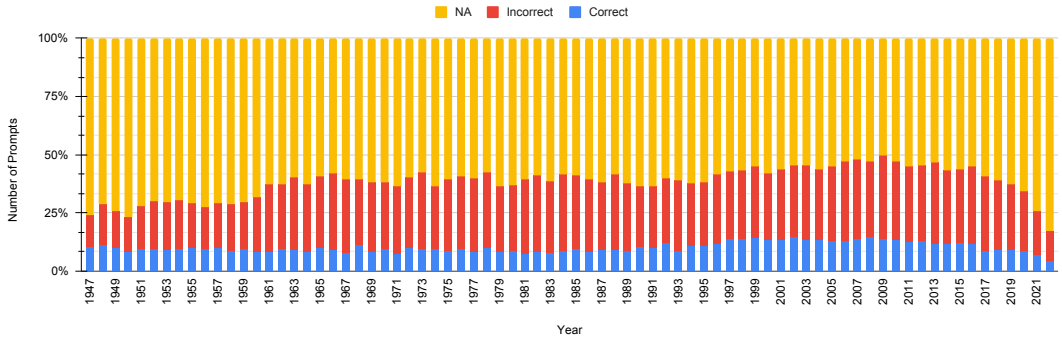

Figure 173: Plot for the Date-based metric (DB) as year-wise count (In percentage) for **yearwise fine-tuning** for `phi-3-medium`.

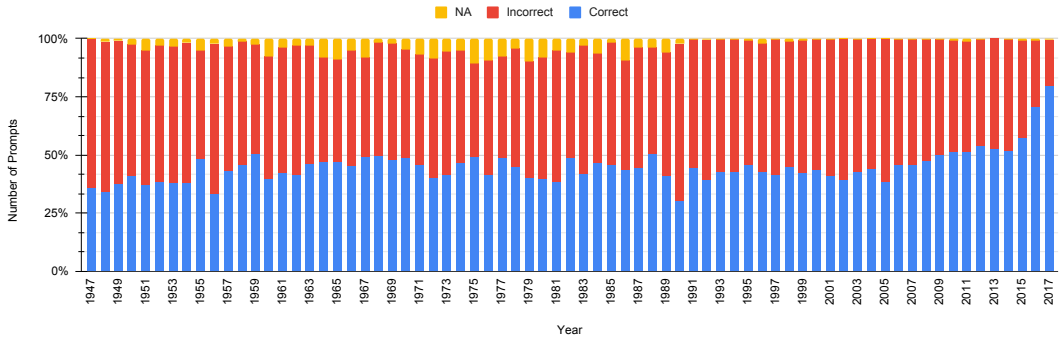

Figure 174: Plot for the Comparative-based metric (CP) as year-wise count (In percentage) for **yearwise fine-tuning** for `phi-3-medium`.

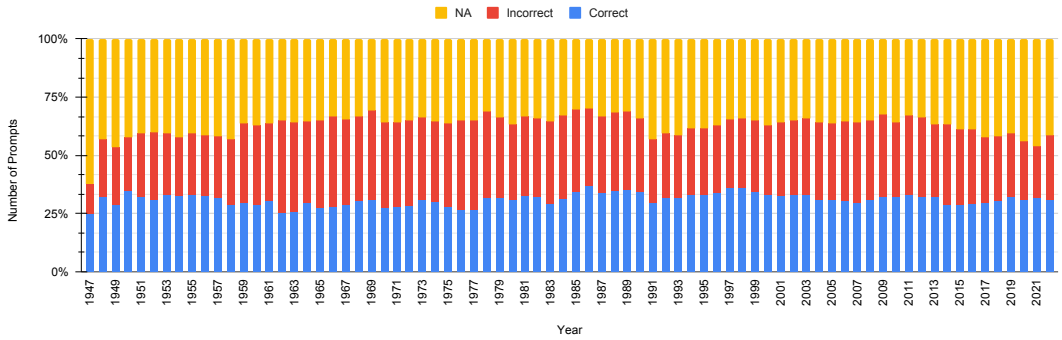

Figure 175: Plot for the Window-based metric (WB) as year-wise count (In percentage) for **yearwise fine-tuning** for `phi-3-medium`.

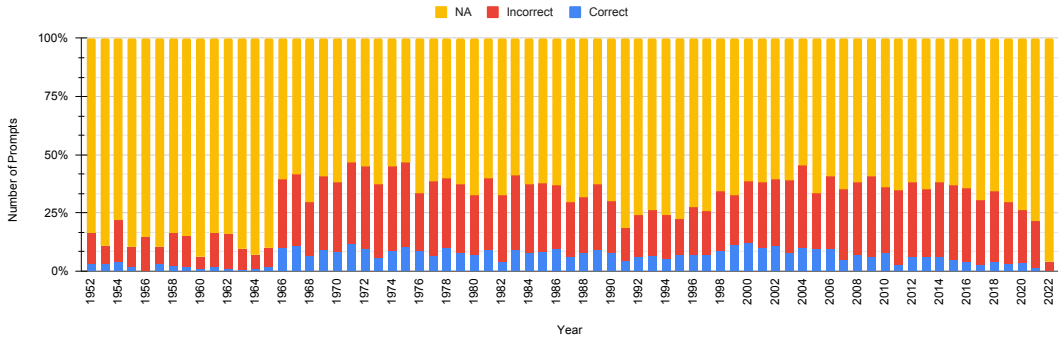

Figure 176: Plot for the Min/Max-based metric (MM) as year-wise count (In percentage) for **yearwise fine-tuning** for `phi-3-medium`.

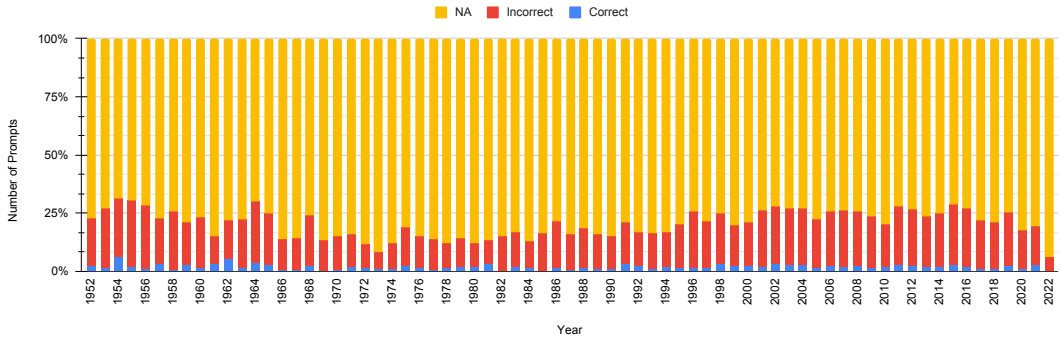

Figure 177: Plot for the Range-based metric (RB) as year-wise count (In percentage) for **yearwise fine-tuning** for `phi-3-medium`.

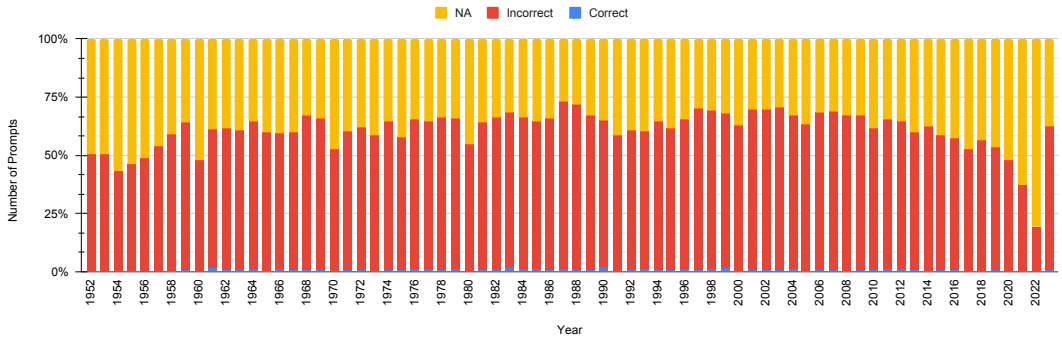

Figure 178: Plot for the Trend-based metric (TB) as year-wise count (In percentage) for **yearwise fine-tuning** for `phi-3-medium`.

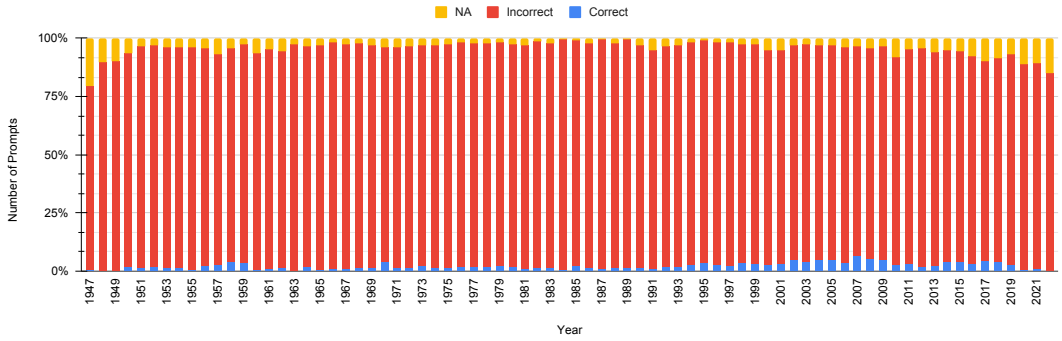

Figure 179: Plot for the Date-based metric (DB) as year-wise count (In percentage) for **random fine-tuning** for `phi-2`.

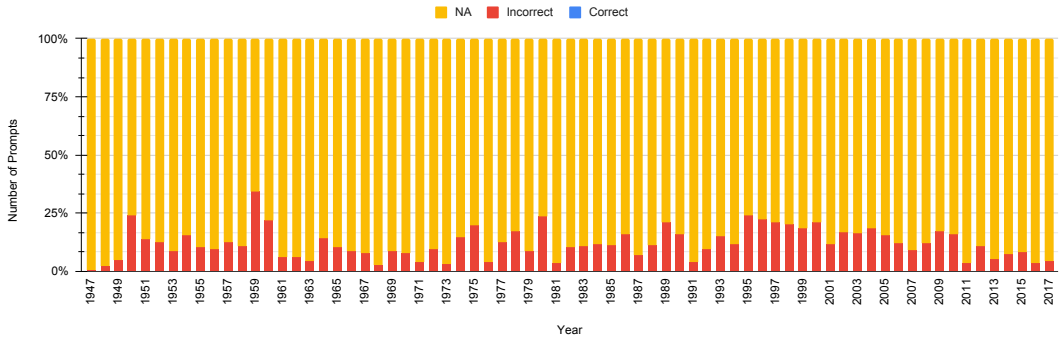

Figure 180: Plot for the Comparative-based metric (CP) as year-wise count (In percentage) for **random fine-tuning** for `phi-2`.

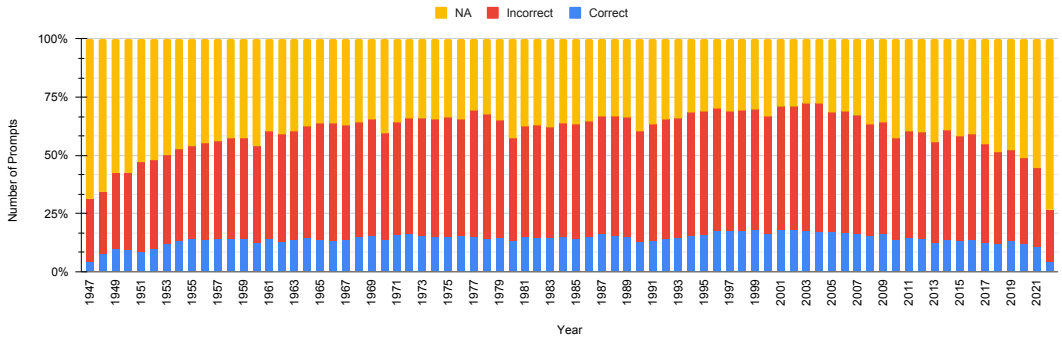

Figure 181: Plot for the Window-based metric (WB) as year-wise count (In percentage) for **random fine-tuning** for `phi-2`.

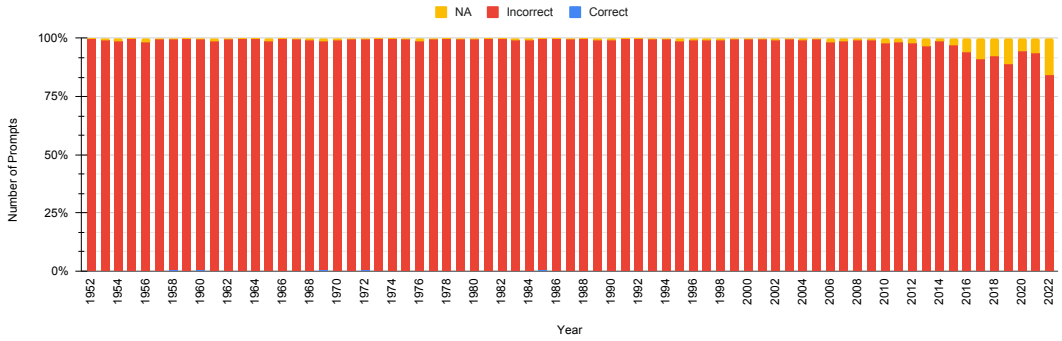

Figure 182: Plot for the Min/Max-based metric (MM) as year-wise count (In percentage) for **random fine-tuning** for `phi-2`.

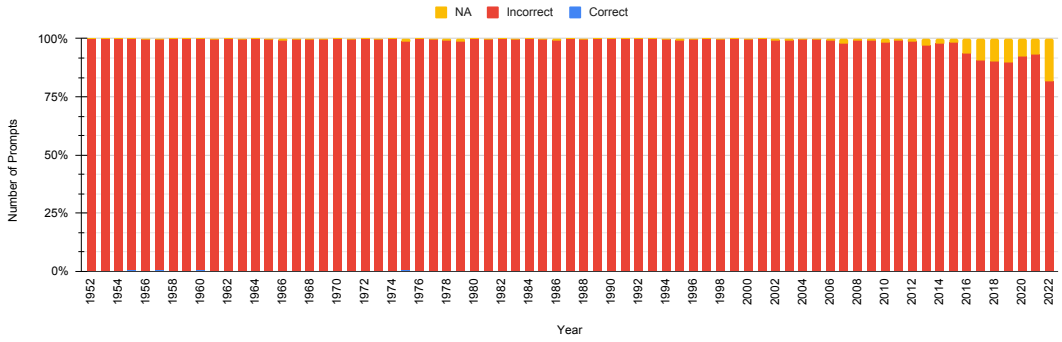

Figure 183: Plot for the Range-based metric (RB) as year-wise count (In percentage) for **random fine-tuning** for `phi-2`.

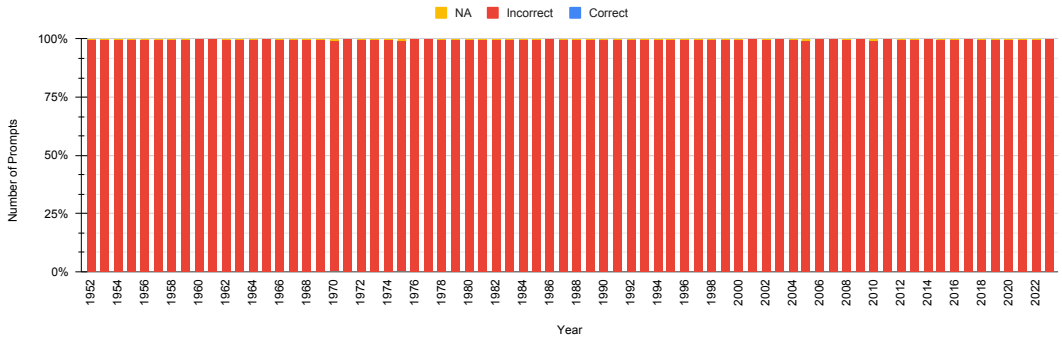

Figure 184: Plot for the Trend-based metric (TB) as year-wise count (In percentage) for **random fine-tuning** for `phi-2`.

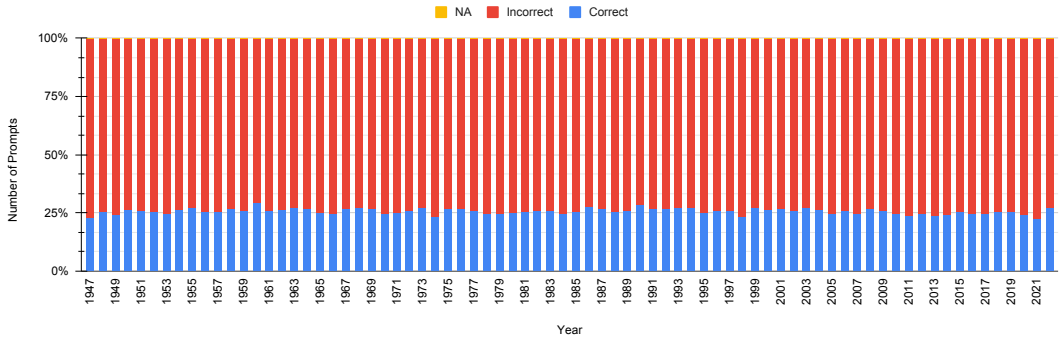

Figure 185: Plot for the Date-based metric (DB) as year-wise count (In percentage) for **random fine-tuning** for `flan-t5-xl`.

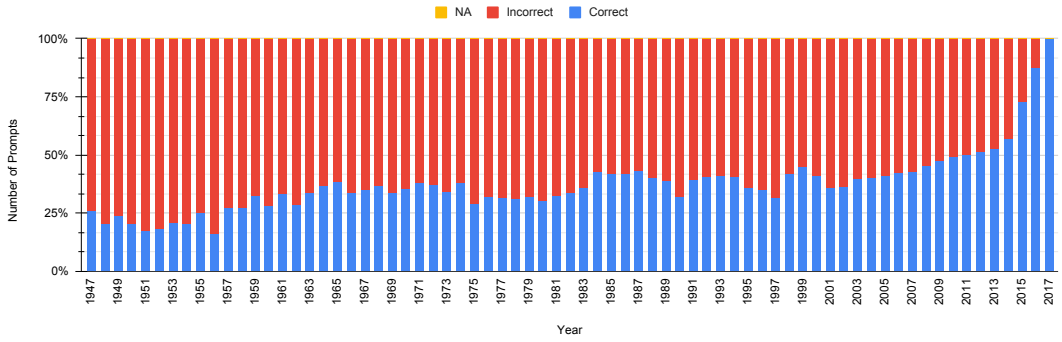

Figure 186: Plot for the Comparative-based metric (CP) as year-wise count (In percentage) for **random fine-tuning** for `flan-t5-xl`.

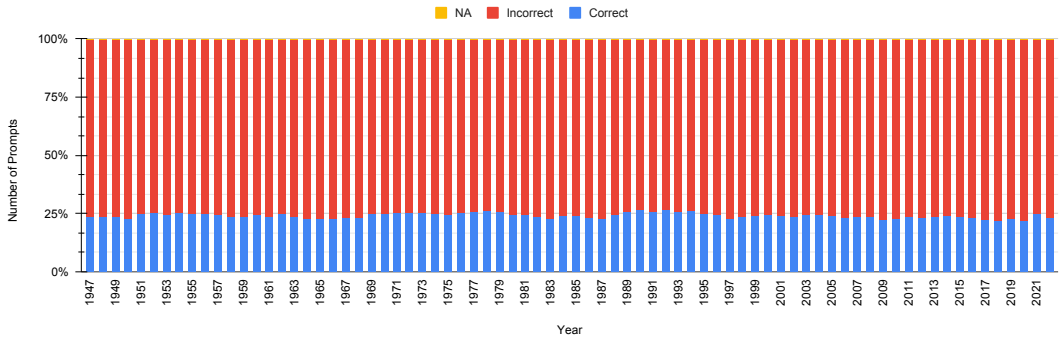

Figure 187: Plot for the Window-based metric (WB) as year-wise count (In percentage) for **random fine-tuning** for `flan-t5-xl`.

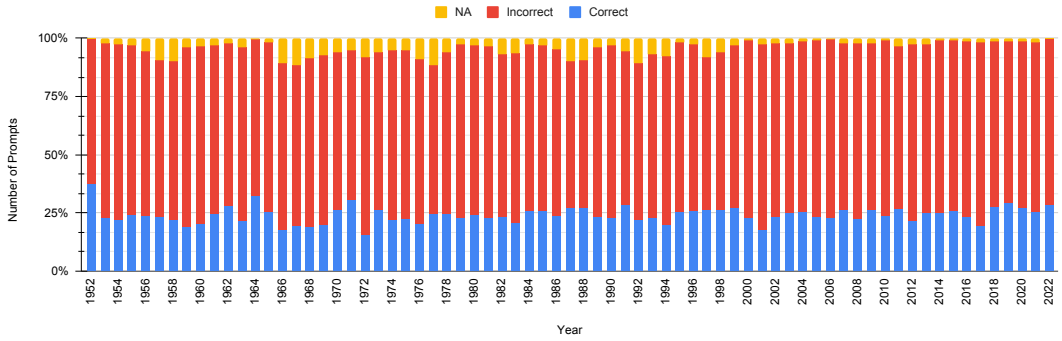

Figure 188: Plot for the Min/Max-based metric (MM) as year-wise count (In percentage) for **random fine-tuning** for `flan-t5-xl`.

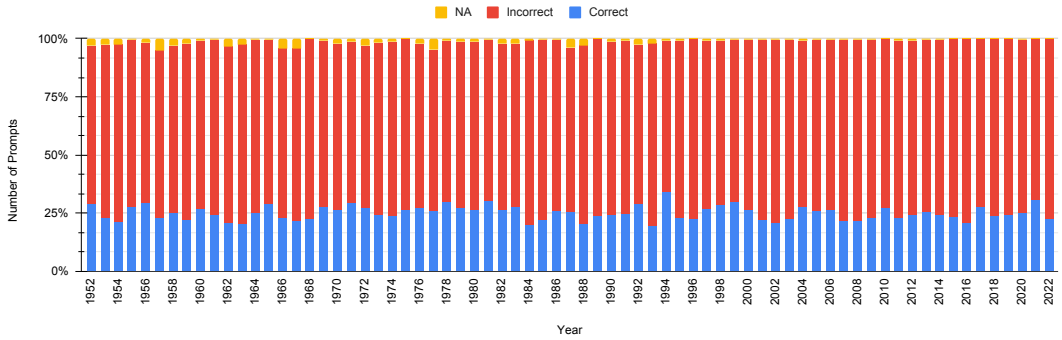

Figure 189: Plot for the Range-based metric (RB) as year-wise count (In percentage) for **random fine-tuning** for `flan-t5-xl`.

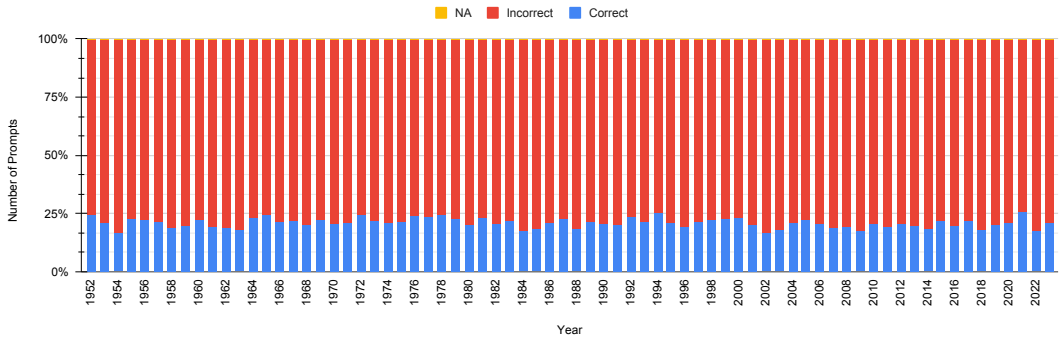

Figure 190: Plot for the Trend-based metric (TB) as year-wise count (In percentage) for **random fine-tuning** for `flan-t5-xl`.

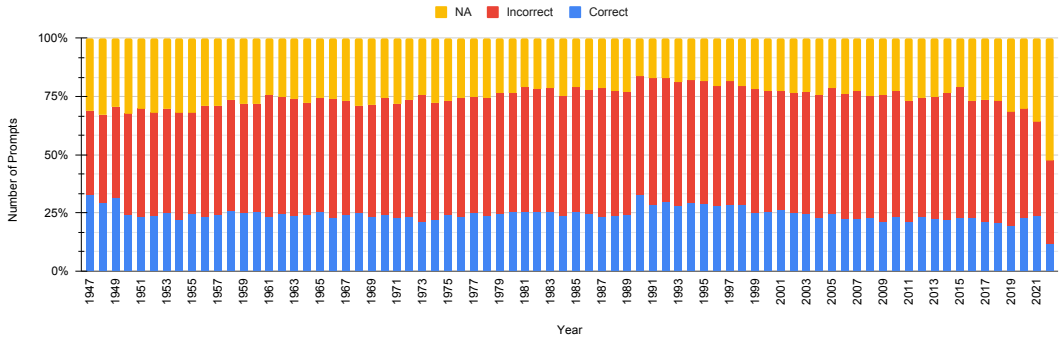

Figure 191: Plot for the Date-based metric (DB) as year-wise count (In percentage) for **random fine-tuning** for `mistral-instruct`.

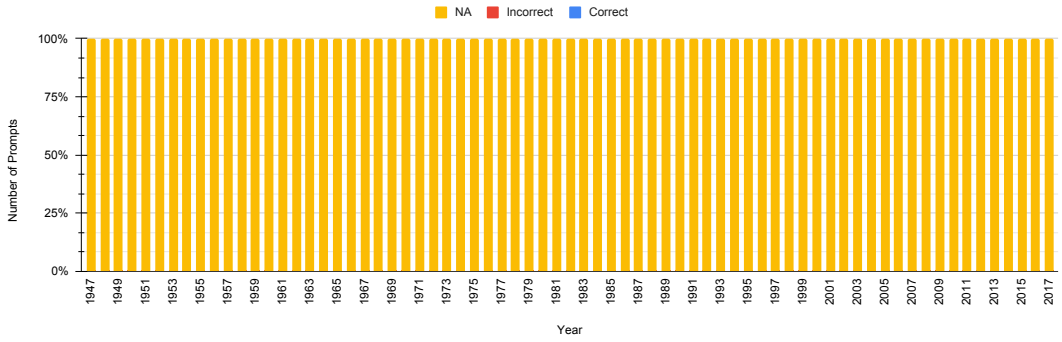

Figure 192: Plot for the Comparative-based metric (CP) as year-wise count (In percentage) for **random fine-tuning** for `mistral-instruct`.

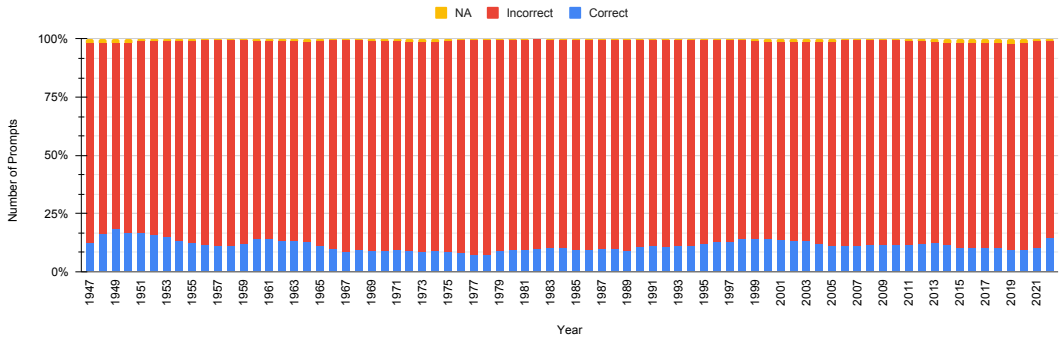

Figure 193: Plot for the Window-based metric (WB) as year-wise count (In percentage) for **random fine-tuning** for `mistral-instruct`.

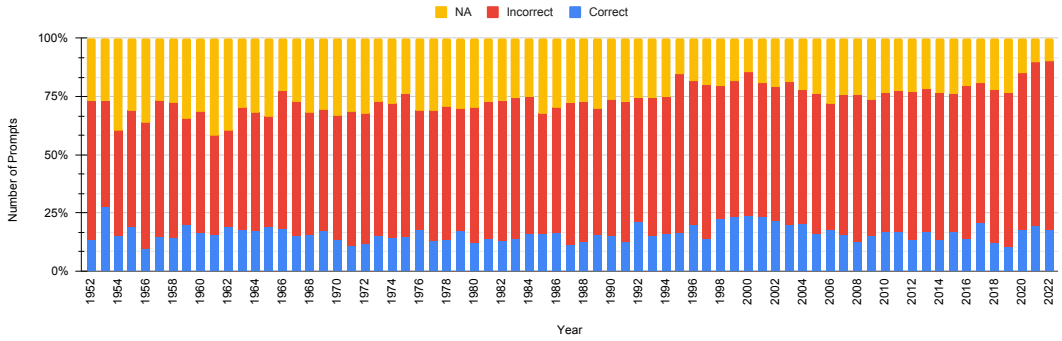

Figure 194: Plot for the Min/Max-based metric (MM) as year-wise count (In percentage) for **random fine-tuning** for `mistral-instruct`.

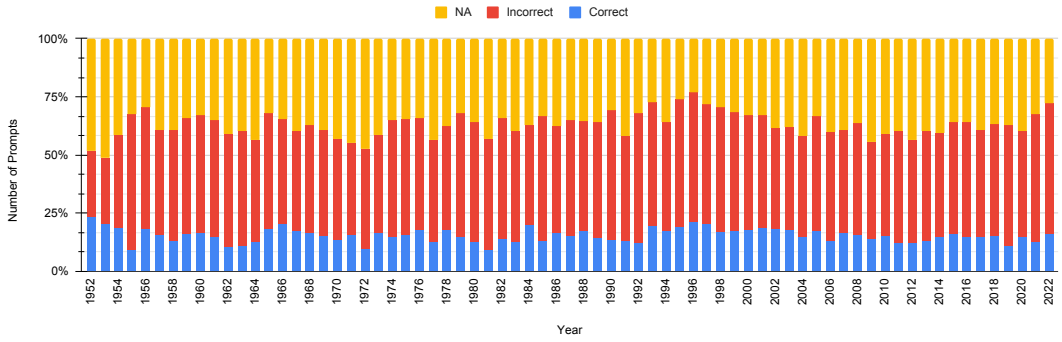

Figure 195: Plot for the Range-based metric (RB) as year-wise count (In percentage) for **random fine-tuning** for `mistral-instruct`.

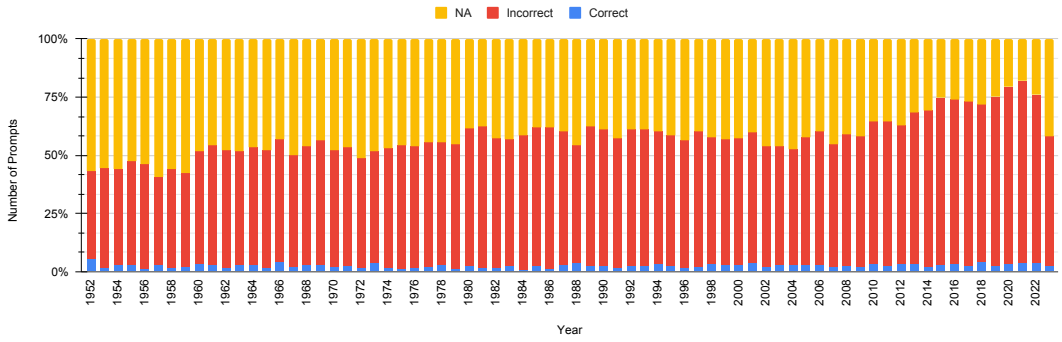

Figure 196: Plot for the Trend-based metric (TB) as year-wise count (In percentage) for **random fine-tuning** for `mistral-instruct`.

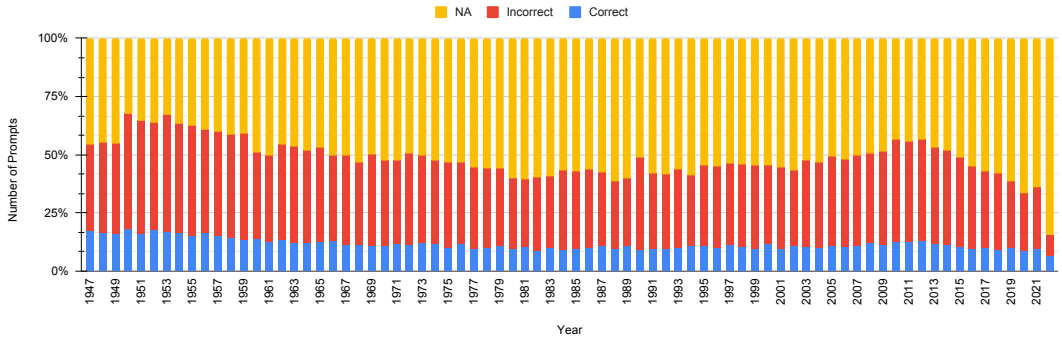

Figure 197: Plot for the Date-based metric (DB) as year-wise count (In percentage) for **random fine-tuning** for `llama-2`.

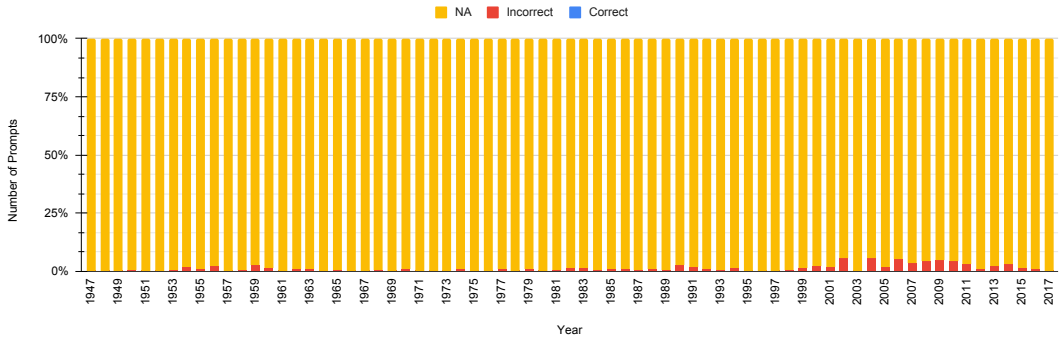

Figure 198: Plot for the Comparative-based metric (CP) as year-wise count (In percentage) for **random fine-tuning** for `llama-2`.

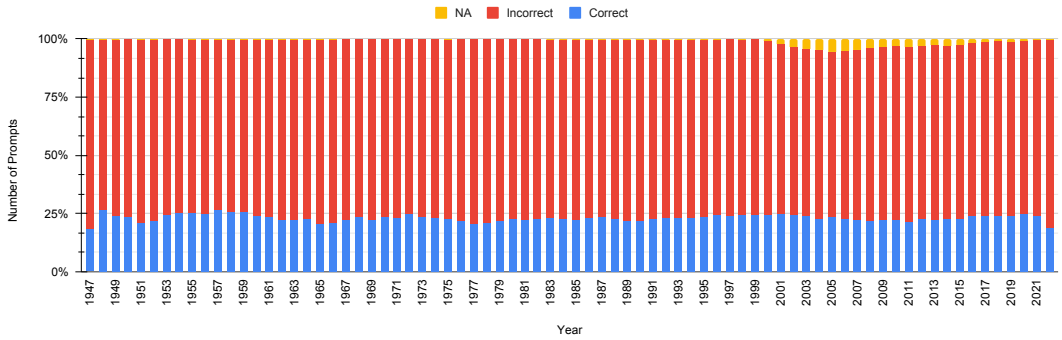

Figure 199: Plot for the Window-based metric (WB) as year-wise count (In percentage) for **random fine-tuning** for `llama-2`.

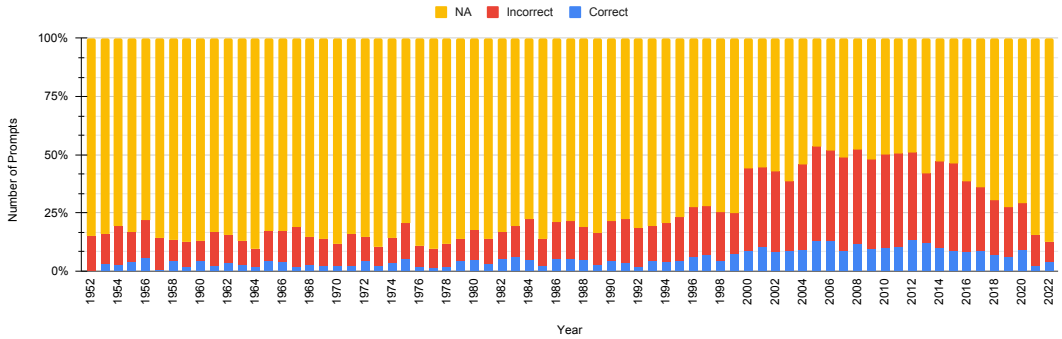

Figure 200: Plot for the Min/Max-based metric (MM) as year-wise count (In percentage) for **random fine-tuning** for `llama-2`.

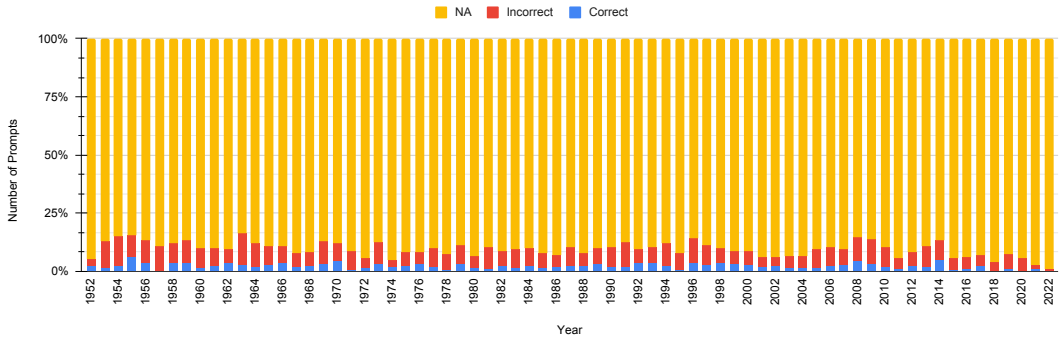

Figure 201: Plot for the Range-based metric (RB) as year-wise count (In percentage) for **random fine-tuning** for `llama-2`.

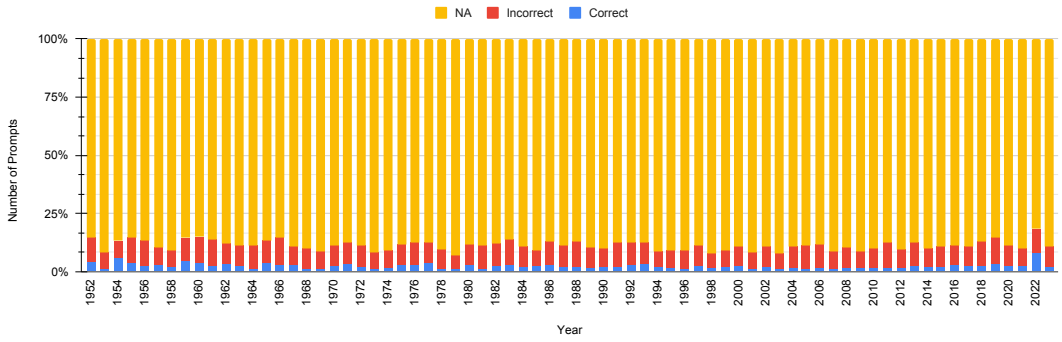

Figure 202: Plot for the Trend-based metric (TB) as year-wise count (In percentage) for **random fine-tuning** for `llama-2`.

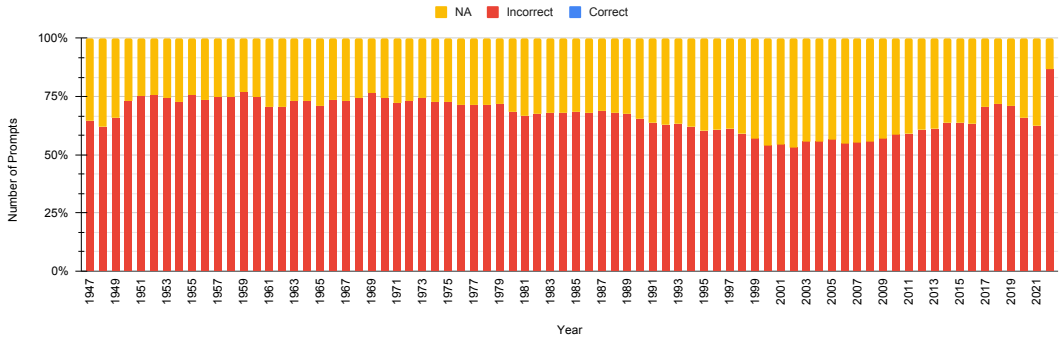

Figure 203: Plot for the Date-based metric (DB) as year-wise count (In percentage) for **random fine-tuning** for `gemma-7b-it`.

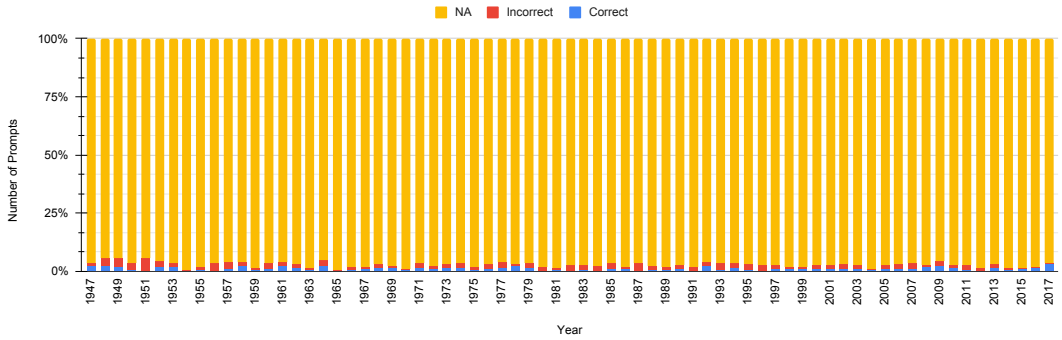

Figure 204: Plot for the Comparative-based metric (CP) as year-wise count (In percentage) for **random fine-tuning** for `gemma-7b-it`.

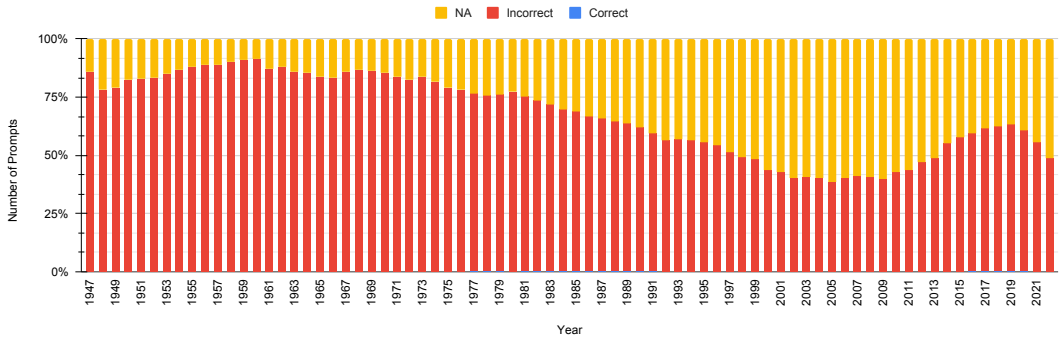

Figure 205: Plot for the Window-based metric (WB) as year-wise count (In percentage) for **random fine-tuning** for `gemma-7b-it`.

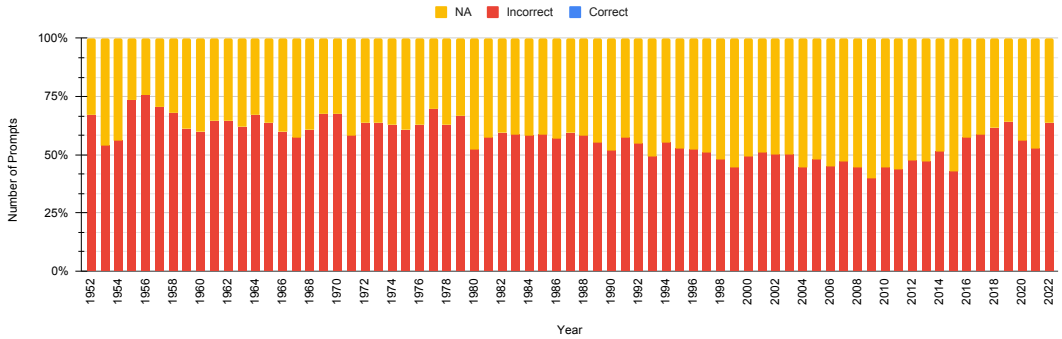

Figure 206: Plot for the Min/Max-based metric (MM) as year-wise count (In percentage) for **random fine-tuning** for `gemma-7b-it`.

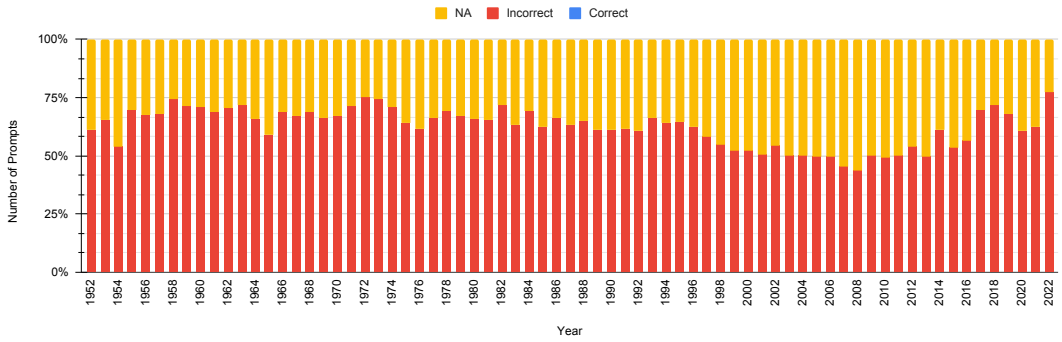

Figure 207: Plot for the Range-based metric (RB) as year-wise count (In percentage) for **random fine-tuning** for `gemma-7b-it`.

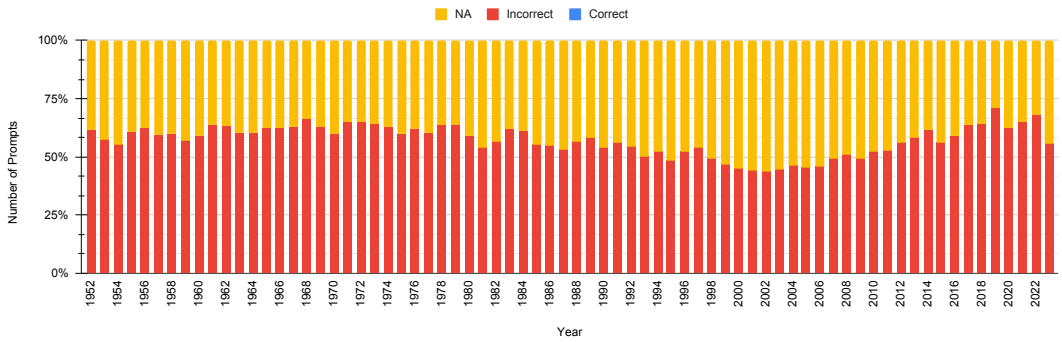

Figure 208: Plot for the Trend-based metric (TB) as year-wise count (In percentage) for **random fine-tuning** for `gemma-7b-it`.

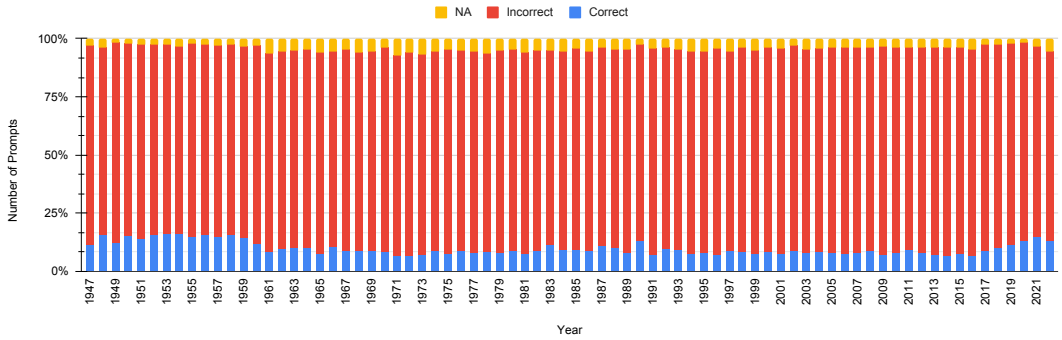

Figure 209: Plot for the Date-based metric (DB) as year-wise count (In percentage) for **random fine-tuning** for `llama-3-8b`.

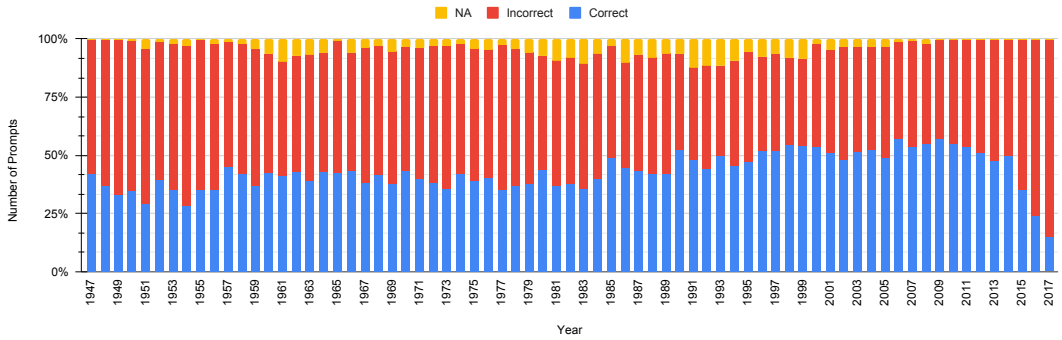

Figure 210: Plot for the Comparative-based metric (CP) as year-wise count (In percentage) for **random fine-tuning** for `llama-3-8b`.

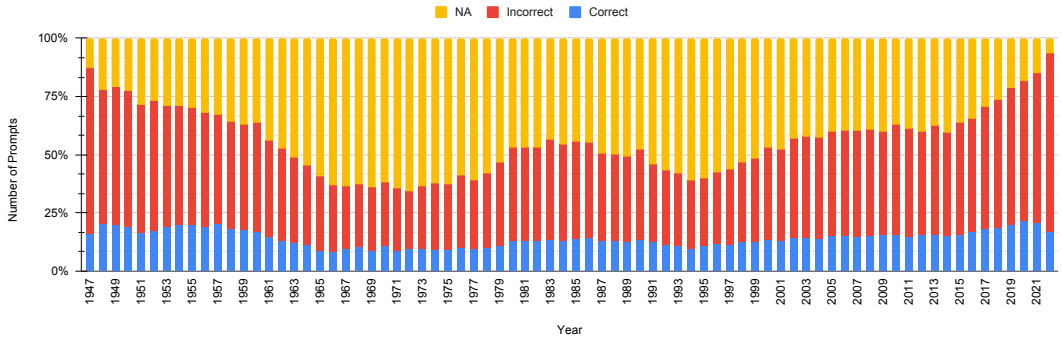

Figure 211: Plot for the Window-based metric (WB) as year-wise count (In percentage) for **random fine-tuning** for `llama-3-8b`.

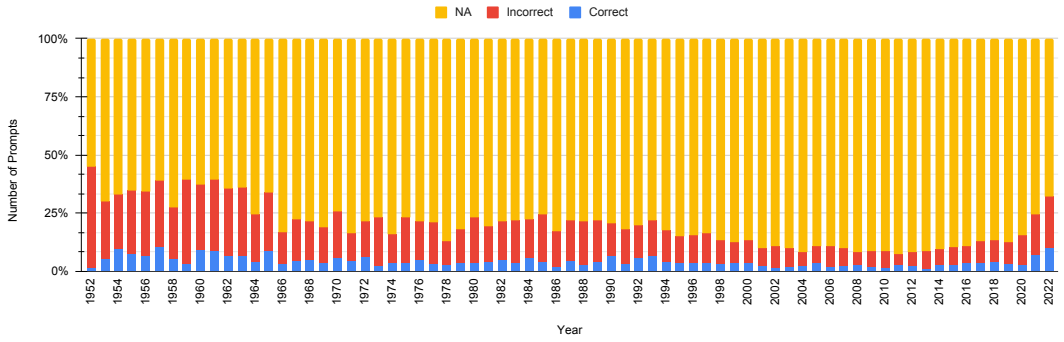

Figure 212: Plot for the Min/Max-based metric (MM) as year-wise count (In percentage) for **random fine-tuning** for `llama-3-8b`.

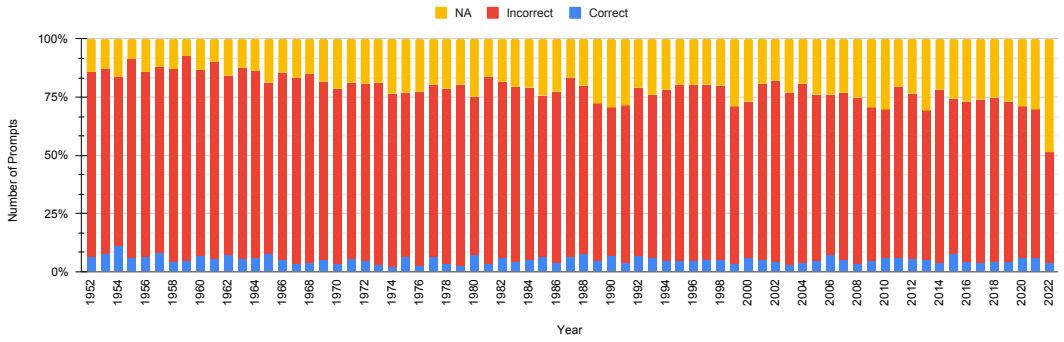

Figure 213: Plot for the Range-based metric (RB) as year-wise count (In percentage) for **random fine-tuning** for `llama-3-8b`.

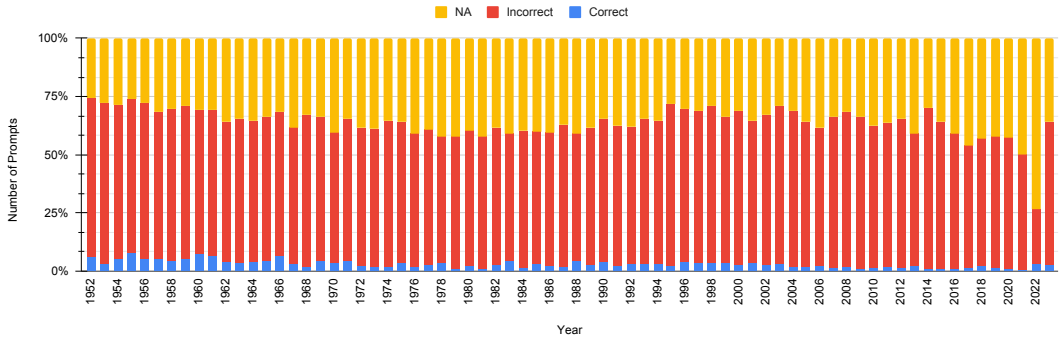

Figure 214: Plot for the Trend-based metric (TB) as year-wise count (In percentage) for **random fine-tuning** for `llama-3-8b`.

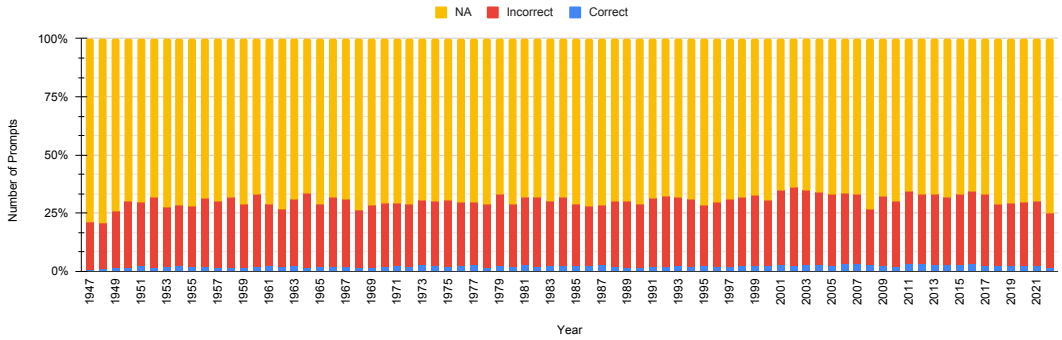

Figure 215: Plot for the Date-based metric (DB) as year-wise count (In percentage) for **random fine-tuning** for `phi-3-medium`.

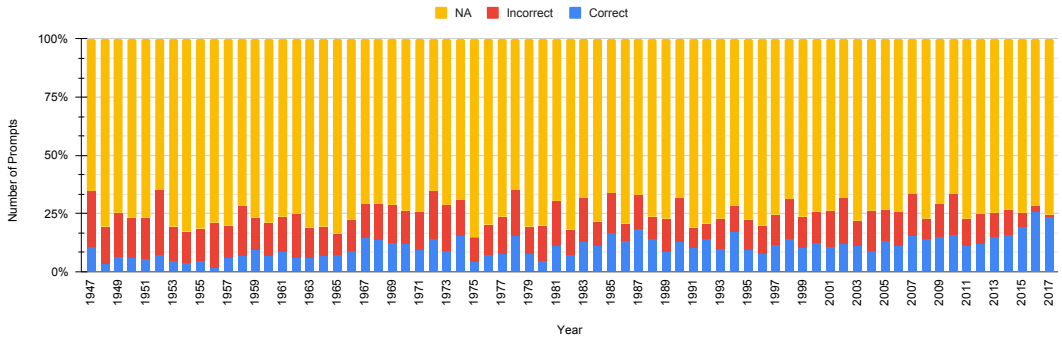

Figure 216: Plot for the Comparative-based metric (CP) as year-wise count (In percentage) for **random fine-tuning** for `phi-3-medium`.

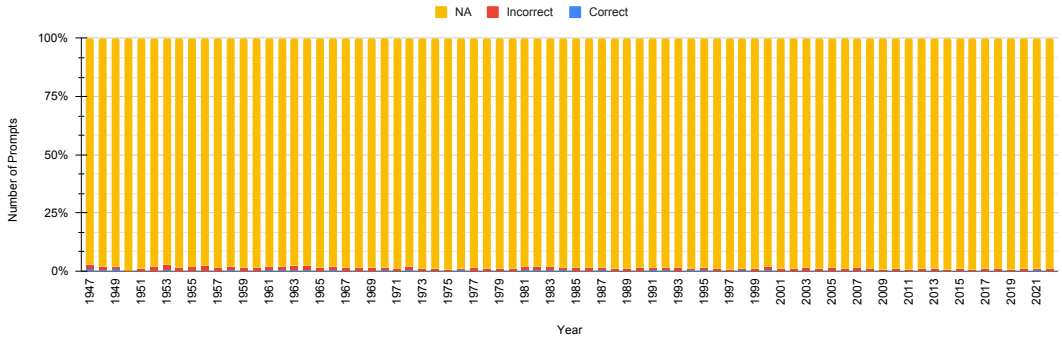

Figure 217: Plot for the Window-based metric (WB) as year-wise count (In percentage) for **random fine-tuning** for `phi-3-medium`.

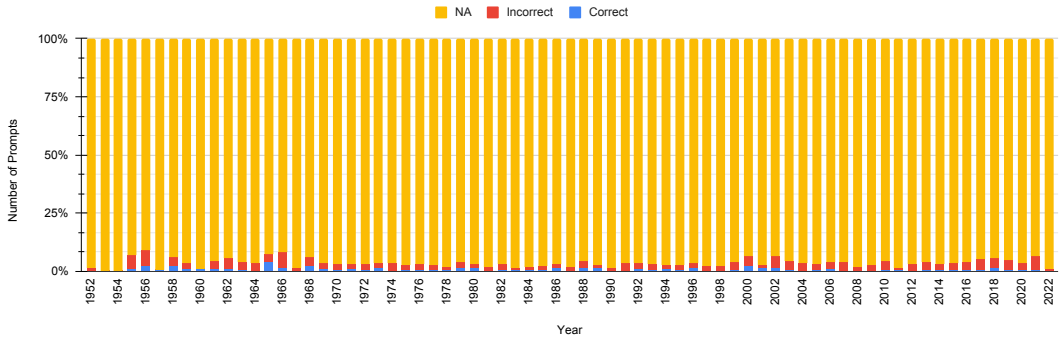

Figure 218: Plot for the Min/Max-based metric (MM) as year-wise count (In percentage) for **random fine-tuning** for `phi-3-medium`.

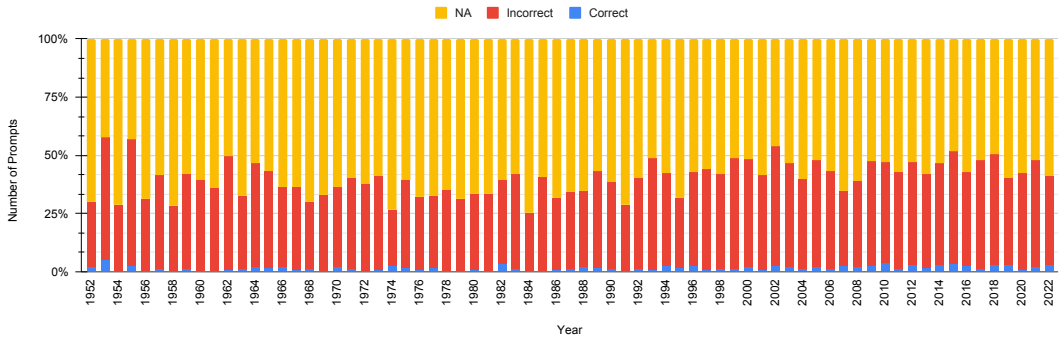

Figure 219: Plot for the Range-based metric (RB) as year-wise count (In percentage) for **random fine-tuning** for `phi-3-medium`.

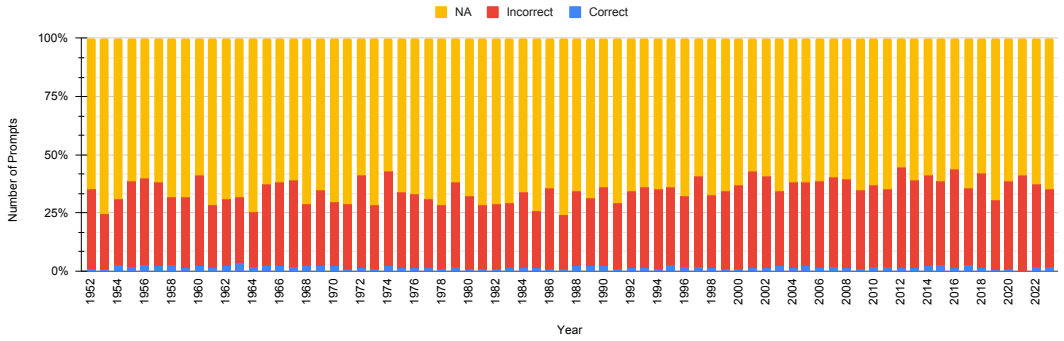

Figure 220: Plot for the Trend-based metric (TB) as year-wise count (In percentage) for **random fine-tuning** for `phi-3-medium`.