Zero-shot generation of synthetic historical data with LLMs

Vera Danilova, Julia Reed, Andrew Burchell, Gijs Aangenendt, and Ylva Söderfeldt

1. Introduction

Synthetic data generation offers clear advantages for training classification models in low-resource settings and in domains where data privacy is critical, such as healthcare, finance, and the social sciences. Unlike traditional data augmentation, which modifies existing samples, synthetic data generation produces entirely new observations by specifying controlled input parameters. This enables researchers to test hypotheses and compare methods in a structured, reproducible manner.

Recent research has explored the use of Large Language Models (LLMs) to generate datasets that combine textual and numerical variables. These approaches rely on the model's ability to infer plausible values from variable descriptions provided in prompts. However, plausibility does not guarantee validity. A growing body of work on fairness and bias highlights that LLM outputs may reflect and amplify systematic biases. As shown by Yin et al. (2025), LLMs exhibit consistent bias patterns and failure modes that differ from those observed in human decision-making. Such biases can propagate into synthetic datasets. Moreover, while large-scale data generation is increasingly feasible, rigorous quality evaluation remains limited—even in sensitive domains. For example, Barr et al. (2025) use GPT-4o to generate synthetic vital sign data for surgical patients under both zero-shot conditions and with predefined statistical ranges. Although zero-shot outputs generally fall within plausible bounds, their confidence intervals often fail to align with those of real data, except for identifiers such as patient case IDs. Incorporating explicit parameter constraints improves alignment, as reflected in greater overlap of confidence intervals. While this is a promising result, more detailed analyses of distributional properties and generative patterns are needed to assess how faithfully synthetic data captures underlying structures.

In the social sciences, LLMs are increasingly employed using methodologies adapted from psychology, at times even as "surrogate" participants in experiments. This line of work often assumes that constructs and experimental designs developed for humans can be directly applied to language models. However, recent findings challenge this as-

sumption. Park, Schoenegger, and Zhu (2024), for instance, demonstrate that GPT-3.5 produces limited diversity when simulating survey responses, suggesting that LLM-generated data may fail to capture the heterogeneity of human attitudes and behaviors. A range of social, political, and cultural biases further undermines the reliability of such data as a proxy for human responses.

Overall, there is a lack of comprehensive frameworks for systematically evaluating synthetic data quality across domains, languages, and data types. This gap extends to Small Language Models (SLMs), which are increasingly adopted due to their lower computational cost (e.g., Binici et al., 2025; Wagner et al., 2024; Zhou et al., 2025; Kazdan et al., 2024), despite well-documented limitations in reasoning capacity and reliability (Wang, 2025).

Within natural language processing (NLP), it has been argued that once text generation reaches near-human performance, classification models can be derived directly, reducing the need for synthetic data (Nikolenko, 2019). However, despite substantial advances in fluency, both LLMs and SLMs continue to exhibit uneven performance across languages and domains, and remain prone to hallucinations, particularly in low-resource settings. In such scenarios, fine-tuning with a combination of human-annotated and synthetic data is a common strategy. Yet, for more subjective tasks—such as sarcasm or irony detection—models trained on synthetic data tend to underperform relative to those trained on human-annotated datasets (Li et al., 2023). This suggests that generated examples may lack sufficient diversity or may encode misleading contextual cues.

In this work, we focus on generating synthetic data for the classification of genre in historical medical periodicals, an especially underrepresented domain. A related and underexplored question concerns how the limitations and biases of LLMs affect the representation of temporal and historical information in synthetic data. This issue is broadly relevant, as most datasets contain implicit or explicit temporal dimensions that shape their structure and interpretation. Ensuring that such temporal aspects are accurately represented is essential to prevent the propagation of distortions into downstream applications and future training data.

This paper describes the initial experiments to generate and evaluate a synthetic dataset that is aimed to mimic genres within 20th-century historical medical periodicals in four European languages. We address the following research questions:

- **RQ1:** To what extent are the texts generated by SLMs faithful to the intended historical and stylistic parameters and how does this fidelity vary across different models and generation scenarios?
- **RQ2:** To what extent can SLMs produce diverse textual outputs and how does this diversity differ between models and under varying generation conditions?

To address these questions, we empirically assess the quality of outputs of SLMs (7b–14b) from LLaMA, Qwen, and Mistral families for domain-specific historical context in terms of faithfulness and diversity as established assessment criteria (Gan and Liu, 2025; Long et al, 2025). To evaluate faithfulness, we analyze model behavior across prompt configurations, focusing on generation failure and preference bias. The quality of outputs, par-

ticularly in terms of language and plausibility of historical details, is assessed through expert feedback from historians of medicine. Diversity is quantified using proxy metrics, including pairwise cosine distances, the proportion of unique topics, and self-BERTscore (Zhang, 2024).

We offer a cautious perspective on the use of SLMs for synthetic data generation, particularly within the context of historical data. While the quality of the generated data appears sufficient to improve performance on certain benchmark tasks, such data may not faithfully represent historical realities. As a result, including these datasets into pretraining for training large language risks reinforcing existing biases and may ultimately undermine model robustness.

2. Motivation for synthesizing historical genres

Our central hypothesis is that genre distributions across time reflect evolving communicative strategies within our source materials. Thus, genre classification may yield richer historical insights than semantic segmentation alone.

The unlabeled data for classification originates from the ActDisease project (ERC-2021-STG 10104099), which investigates the modern history of European medicine. The dataset consists of periodicals published by ten patient organizations across Europe, spanning the 20th century and covering four languages: Swedish, German, French, and English (Aangenendt et al., 2024). It has only recently been digitized, is not publicly available, subject to copyright restrictions, and not included in the pretraining data of existing LLMs.

The periodicals are diverse in terms of text and layout types, with individual pages often combining multiple genres. The quality of OCR hinders the reliable extraction of distinct texts with their correct reading order. As a result, term frequency counts and topic models tend to be biased toward dominant genres, complicating accurate historical interpretation (Danilova and Söderfeldt, 2025).

Working with the historians in the project, we identified 23 distinct genres: *academic, administrative, advertisement, announcement, appeal, opinion, patient_organization_report, QA, news, legal, invitation, interactive (crossword or puzzle sections), guidance, fiction, personal, medical_case, humour, recipe, medical_recipe, biographical, interview, letter_of_thanks, complaint*. These genres are based on previous work on genres of historical journalism (Harbers, 2014) and genres in the history of medicine (Pomata, 2014), as well as web genre classification (Kuzman and Ljubešić, 2023). However, suitable training and test data is scarce, and for 14 genres, no annotated data is available for training and testing the models: *administrative, announcement, appeal, biographical, complaint, humour, interview, invitation, letter_of_thanks, medical_case, medical_recipe, news, personal, recipe*. This limitation motivated us to investigate the use of language models for synthetic data generation. Rather than producing data solely for classifier training, we systematically examine the generated content across different generation scenarios and evaluate its diversity and faithfulness, which is critical for ensuring the reliability and representativeness of synthetic historical texts.

3. Methodology

The content of domain-specific historical corpora is inherently tied to particular historical timelines. This study explores the ability of LLMs to reflect as many features of these timelines as possible and to not introduce distortions. We use a zero-shot setup where the models are expected to use linguistic and contextual features consistent with the relevant historical timeline, drawing on knowledge acquired during pre-training about general history, the history of medicine, journalistic practices, linguistics, and both historical and modern text genres.

We introduce several generation scenarios (prompt configurations) to assess how they influence both the diversity of generated outputs and their historical reliability. To our knowledge, no existing automatic metrics can adequately assess the accuracy of representations of historical realities—particularly in such a specific domain as patient organization periodicals. Consequently, historians of medicine play a central role in the evaluation process, being able to detect major historical distortions in the representations produced by the models.

In the following, we provide details on the evaluation criteria. Since LLMs have been shown to exhibit various forms of inherent bias, including primacy, recency, and centrality biases (Yin et al., 2025), we additionally assess how effectively the models can produce structured JSON outputs and whether their generated content displays a biased preference for certain years depending on contextual cues. Such analysis provides insights into the biases in model behavior, which may impact both diversity and faithfulness. Furthermore, we outline how the evaluation criteria are applied to address the research questions and describe the rationale for model selection and the design of the generation scenarios.

3.1 Evaluation criteria

Faithfulness. As defined by Long et al. (2024), *faithfulness* refers to outputs that are logically and grammatically coherent, free from factual inaccuracies or irrelevant content. In our case, this is particularly challenging due to the domain-specific and historically grounded nature of the data. Since our downstream goal involves training models on historical materials, it is critical to maintain historical accuracy: real events must not be conflated or misrepresented, and anachronisms must be avoided. In genre classification specifically, such historical inaccuracies may introduce topical shifts that distort the training signal. Previous work (Petrenz, 2011) has shown that shifts in topic distribution can change the behavior of genre classification models. Hence, faithfulness evaluation is helpful for both analyzing and avoiding potential classification errors, as well as for preventing the propagation of historical inaccuracies in future datasets and computational tools trained on them.

The evaluation of faithfulness addresses RQ1 through analyzing 1.A: alignment with factual historical timeline, 1.B: semantic similarity to original dataset, 1.C: accurate reflection of genres specified in input prompt, 1.D: historically plausible language.

The factual accuracy of the text (1.A) is verifiable by reference to the most literal, and most widely agreed on, understanding of "what happened" in the history of

medicine—the passing of specific legislation, the development of medical technologies, the social and institutional histories of physician-patient relationships, etc. This information is available in textbooks or encyclopedias as historical "facts" that have been communicated by socially accredited historical authorities who have in principle verified the facts in an archive and communicated their historical significance. The facticity of these facts is thus a function of some level of concreteness–whether founding years, major events, and personnel of groups and institutions, the testing, patenting, and circulating of technologies, or less concrete facts such as the emergence or transformation of particular ideas.

Historical plausibility (1.D) concerns whether a text feels believable or authentic within its claimed historical context — even if it's not factually verified. This evaluation focuses on style, tone, worldview, language use, genre conventions, and attitudes that fit the period.

Assessing the plausibility of texts is challenging, since historical sources often defy plausibility—and this may, in fact, characterize what makes them valuable as primary sources. A primary source is an authenticated object from the past that reveals what actually happened, rather than what seems likely or expected to have happened. For instance, a reader letter published in the *Swedish Diabetics Association's* periodical in 1978 describes life with diabetes in strikingly negative terms and even includes curse words—a tone that runs counter to what one might expect from a patient organization's publication and its public messaging at the time (*Från läsekretsen*, *Diabetes*, vol. 28, issue 1, 1978: 24).

In the present experiment, the experts are fully aware that the synthetic texts are not authentic primary sources. Yet many of these generated outputs could easily be imagined as plausible historical documents—the kind of materials one might encounter in an archive. Importantly, the assessment of their plausibility cannot be separated from the context of the evaluation itself.

The evaluation of model outputs is performed by two historians of medicine from the ActDisease project who specialize in European patient organizations. In the preliminary evaluation phase, the historians were given outputs from LLaMA 3.1 8B, Qwen 2.5 7B, and Qwen 2.5 14B and were asked the following questions: 1) How well do the generated texts reflect the historical genres identified in the periodicals?; 2) How diverse (lexically, structurally) are the generated samples within each genre?; and 3) How well does the generated data fit the selected historical period? Based on these questions, the historians provided a summary of the historical genre fit in the produced synthetic data.

In the main evaluation phase, the historians addressed RQ1 in detail. Given the high cost of expert evaluation at the level of each generated text, the RQ1 sub-questions were provided to historians as a questionnaire to allow for an efficient yet thorough assessment of overall quality. Higher synthetic data quality is defined as a closer alignment between the outputs and the historians' expectations. Each response is given as an integer score within the interval [0, 10], where 0 indicates no alignment with expectations, and 10 indicates the highest possible alignment.

Diversity. The diversity metric captures the variation among the generated data, reflecting differences in text length, topic, or writing style. Higher synthetic data diversity is reported to mitigate distribution collapse and positively influence the performance of fine-

tuned models (Schaffelder and Gatt, 2025). In previous literature, it has been observed that instruction-tuned conversational models widely used for synthetic data generation produce outputs with limited semantic diversity. Base models, although not as good at following complex instructions, have been observed to be more semantically diverse in outputs (Zhu et al, 2025; West and Potts, 2025). Evaluating diversity across models and generation scenarios offers valuable insights into the lexical repetitiveness of model outputs and the breadth of topics they cover when they are conditioned on a historical context.

The evaluation of diversity addresses RQ2 through analyzing 2.A: Does diversity differ across generation scenarios as expected from the prompt design? 2.B: Do base models produce more diverse outputs according to the metrics, and what insights emerge from manual inspection?

The following proxy metrics are used to quantitatively assess the diversity of model outputs:

1) pairwise cosine distances between embeddings generated using the Sentence Transformers[1] framework. Each generated text in each language is embedded using the Qwen3-Embedding-4B model which supports multiple languages and long input context (32k tokens). Cosine distances are in range [0,2] (calculated as 1 – cosine similarity on normalized embeddings using sklearn.pairwise_distances). Distributions concentrated near zero indicate lower semantic diversity (i.e., outputs are more similar), while distributions with values farther from zero suggest higher semantic diversity. We visualize these distributions to evaluate whether semantic similarity increases or decreases as a function of specific model or prompt modifications.
2) self-BERTScore (Zhang et al., 2024) where the BERTScore F1 metric (Zhang et al., 2020) is computed separately for each language subset within every model and prompt configuration, across both disease and genre groups. Contextual embeddings were obtained using the bert-base-multilingual-cased model. To achieve this, the data were first grouped by language, model, genre, and disease, and all corresponding text samples were aggregated into lists. Groups containing fewer than two texts were excluded. For each remaining group, pairwise BERTScores were calculated among all texts within that group to assess internal semantic similarity. A score closer to 1.0 indicates greater semantic overlap—i.e., more tokens from the candidate and reference share similar meanings in embedding space. Distributions of pairwise scores concentrated near zero indicate higher semantic diversity. To quantify and compare these patterns we summarize each distribution using its mean and standard deviation.
3) the proportion of unique topics relative to the total number of generated texts for each model and prompt condition. To measure the proportion of unique topics, topic modeling is performed using the BERTopic framework (Grootendorst, 2024), applied to the same set of embeddings as in 1), with GPT-4o serving as the topic representation model.

1 https://huggingface.co/sentence-transformers

Generation Performance. We measure how often the models fail to generate a valid JSON entry across prompt configurations. Systematic failure to generate data for specific years or contexts might indicate underlying biases in the model's training data or limitations in its ability to generalize beyond dominant temporal or linguistic patterns. It may also suggest that certain combinations of input conditions fall outside the model's learned distribution, resulting in non-responses, irrelevant completions, or format violations.

Year and Digit Preference. We hypothesize that the model's year selection patterns will reflect either inherent positional biases, domain-specific knowledge biases (e.g., the model's awareness of events associated with particular years in the context of a given disease), or both. The models are prompted to select a year from the 20th century based on the provided context. The evaluation examines whether the models' year preferences shift depending on the breadth of the context and the presence of references to medicine, disease, or patient organizations. We analyze plots showing the distribution of selected years in the model outputs, where counts are normalized by the total number of outputs for each model and prompt configuration. These distributions are then compared across different models and prompt versions.

To determine whether year preferences remain consistent despite prompt modifications, we introduce several controlled variations:

1. **Removal of domain-specific instructions:** The model receives no guidance to emulate patient organization periodicals for specific diseases. If the year preference is strongly influenced by this instruction, we expect a marked deviation from the baseline distribution.
2. **Temporal range restriction:** The available timeframe is reduced from the entire 20th century to only its first half. This allows us to observe whether the model's preferences are positionally anchored or adapt to the narrower temporal scope.
3. **Repositioning of reasoning summary:** Initially, the chain-of-thought summary appears at the end of the generated output. We move it to the beginning and explicitly instruct the model to reflect on its choices regarding year and topic selection. This variation tests whether encouraging explicit reasoning affects the model's year preferences.

Given the observed numerical preferences of LLMs[2], we hypothesize they will manifest in the model's selection of year during synthetic data generation, especially when the generation is constrained to a predefined value range and conditioned on specific input variables in the prompt.

3.2 Models

To address the research questions outlined above and perform evaluation, we conduct a series of comparisons across two pairs of models with different sizes. The first pair con-

2 E.g., https://sanando.github.io/llmrandom/

sists of Qwen 2.5 7B Instruct and LLaMA 3.1 8B Instruct. Since a LLaMA model of comparable size to Qwen 2.5 14B Instruct is not available, we instead use Mistral Nemo 2407 12B Instruct as the second comparison point. This setup enables us to examine differences both across model architectures and across model sizes within the same architecture.

Since we generate not only text but other information, such as chain of thought (CoT), rationale behind the choice of time period features, year, etc., we use instruct models that are able to consistently generate JSON outputs. Additionally, Instruct-tuned models have been shown to perform better in zero-shot settings across various NLP downstream tasks (Wei et al., 2021), making them particularly suitable for this study.

For the diversity evaluation purposes, we include base models that have not undergone further training to align with human preferences in instruction-following tasks. Specifically, we compare the Qwen 2.5 14B Base model with its corresponding Instruct version. This comparison is intended to investigate whether the base model produces considerably more diverse outputs. This model is capable of producing valid JSON outputs, however, it does so at a considerably lower rate and approximately four times slower than the Instruct version. The Mistral and LLaMA Base models were unable to generate the minimum required number of valid JSON outputs.

3.3 Prompt configurations

We employ six zero-shot prompting scenarios in these initial experiments, applying minimal, targeted modifications to the prompt — specifically avoiding few-shot generation — since our experimental setup involves a long tail of underrepresented classes for which no real examples exist and no representative seeds are available to support few-shot prompting. Another reason is that the models can be severely influenced by the semantics and order of few-shot examples (Lu et al., 2022). The few-shot setup will be explored in the future work.

Baseline Prompt. The baseline prompt that will be further modified is constructed as follows. The system part of the prompt focuses the model on the historical content, specifically, on patient organization magazines and the period of their publication, and stresses the importance of strictly following the instructions:

> You are an AI specializing in historical content synthesis. Your purpose is to generate structured JSON data that emulates authentic documents from 20th-century patient association magazines. Adhere strictly to the provided historical context and output format.

The user prompt further provides the input parameters (contextual variables, such as type of disease, country, etc.), steps of JSON creation and an example of output JSON structure, clearly separating them with formatting attributes. In the baseline prompt, it introduces the following input parameters: *genre, text types associated with it, target text type, country, disease, language*.

A dataset of prompts is generated by randomly sampling input parameters (genre, disease, country, language). The model is instructed to select a plausible year and to gen-

erate JSON entries with the following fields in the following order: *genre, text type, year, text, topic, country, disease, language, time period features in text,* and *the CoT*. Then the generation is performed sampling parameters including *tau* = 0.8 and *top_p* = 0.95 fixed across generations. Each model generates, in each run, 100 sample entries per language and prompt configuration.

This baseline addresses annotators' tendency to disproportionately focus on specific aspects of a genre's definition, leading to inconsistent classifications. Rather than relying on explicit, potentially prescriptive, textual definitions, we propose to implicitly define each genre by a set of exemplary text types. This strategy aims to mitigate the risk of prioritizing certain definitional components over a holistic understanding. Furthermore, this baseline leverages only the inherent, pre-trained knowledge embedded within the model.

The overview of baseline prompt modifications is given in Table 1. Within each modification, the prompt and sampling parameters are fixed to be able to perform comparison across models. Across modifications, sampling parameters are the same.

Table 1: Prompt configurations

prompt_id	Defined by prompt	Selected by LLM	Difference from the baseline prompt
v0 (baseline)	genre, text types, text type, disease, country, language	year, topic	-
v1	genre, country, language	year, topic	excluded information about medicine and patient organization from system and user prompt (**some text types and genres are related to medicine and are not excluded**)
v2	year, genre, text types, text type, disease, country, language	topic	year is defined in the prompt
v3	genre, text types, text type, disease, country, language	year, topic	changed the time period: 20th century substituted by 1900–1955
v4	genre, disease, country, language	year, topic	excluded information about text types
v5	genre, text types, text type, disease, country, language	year, topic	CoT generated first

4. Synthetic dataset description

The dataset includes 18,389 entries: 4 runs per model were performed for the baseline prompt and 2 runs per model – for the rest of the prompt configurations, with varying percentage of successful JSON generations. The dataset covers all 23 genres with the average number of entries per genre equal to 799 and standard deviation of 38. An example entry generated by Qwen 2.5 14B is provided below:

Genre: Opinion
Text type: arguing for a treatment
Year: 1968
Text: 'As a sufferer of rheumatism, I have personally witnessed the transformative impact of corticosteroids on my condition. These powerful medications have drastically improved my quality of life and allowed me to manage symptoms that once severely limited my activities. However, despite their benefits, there remains a hesitancy among some physicians and patients to embrace corticosteroids fully due to lingering concerns about side effects. It is critical to emphasize that corticosteroids are not only safe when properly administered but also represent a vital tool in managing severe rheumatism effectively. We should continue to advocate for their use while working closely with healthcare providers to monitor and mitigate any adverse effects.'
Topic: Advocating for the Use of Corticosteroids
Country: United Kingdom
Disease: rheumatism
Language: English
CoT: 'The selected year is 1968 due to the availability of corticosteroids as a treatment for rheumatism, which was considered groundbreaking at the time. This advancement led to numerous discussions within patient communities regarding its efficacy and safety. Additionally, corticosteroids were not without controversy due to potential side effects, making them a topic of debate among patients and physicians alike.'

5. Evaluation results

To contextualize the evaluation of faithfulness and diversity, we first examine the models' JSON generation performance and the observed year and digit biases.

JSON Generation Performance. Tab. 2 shows the accuracy of the models in producing valid JSON outputs. On average, LLaMA produces the fewest number of successful outputs across prompt scenarios. Larger models are on average almost 100% accurate when they autonomously select the year. However, when the year is included in the prompt, larger models generate almost 50% less successful formats, while LLaMA 3.1 8B and Qwen 2.5 7b – 26% and 42% less, respectively. Qwen 2.5 14B Base was used only for generation with the baseline prompt and produced 35.6 % of valid outputs. Also, it took 4 times longer for the Base version to complete output generation.

Table 2: Accuracy of models in producing structured JSON (averaged across runs)

Average across Prompts v0, v1, v3, v4, v5 - year is selected by the LLM		**Prompt v2 - year is pre-defined in the prompt**	
Model	**JSON (%)**	**Model**	**JSON (%)**
LLaMA 3.1 8B-Instruct	59.0	LLaMA 3.1 8B-Instruct	33.0
Mistral Nemo Instruct 2407	99.0	Mistral Nemo Instruct 2407	49.3
Qwen 2.5 14B-Instruct	97.6	Qwen 2.5 14B-Instruct	48.6
Qwen 2.5 7B-Instruct	88.5	Qwen 2.5 7B-Instruct	46.9

Table 3: Average percentage of years covered by the models per language using Prompt v2

Language|Model	**LLaMA 3.1 8B**	**Qwen 2.5 7B**	**Qwen 2.5 14B**	**Mistral Nemo**
English	58.24 %	89.56 %	98.90 %	99.45 %
French	75.27 %	99.45 %	99.45 %	98.35 %
German	81.87 %	95.05 %	95.05 %	97.25 %
Swedish	48.90 %	91.76 %	97.25 %	97.80 %

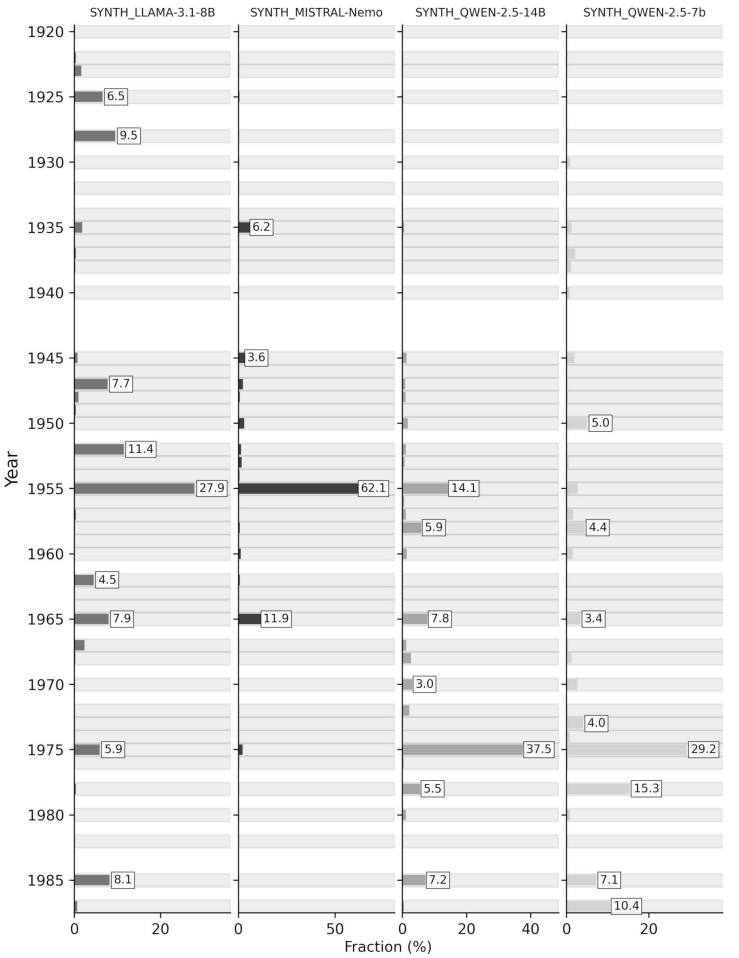
As expected, larger models tend to follow our complex instructions more effectively. However, even when the prompt explicitly specifies the years, the accuracy in generating valid outputs decreases and models do not fully cover the intended year range. Coverage also varies by language, with smaller models like LLaMA and Qwen 7B covering fewer years in English compared to other languages (Tab. 3).

Year and Digit Preference. The models were prompted to autonomously select a year within the 20th century, following exposure to contextual information: patient organization, disease, genre, and text types. The distribution of year selections for the baseline prompt is presented in Fig. 1. Mistral exhibits a pronounced bias toward the year 1955, with over 62% of its selections falling on this specific year (on average per run with standard deviation – SD = 2.8 across runs).

We compare the coefficients of variation (CV = SD / mean) across models to assess the relative variability in year generation. LLaMA exhibits the highest variability (CV = 0.009), while Mistral shows the lowest (CV = 0.004). LLaMA's relative spread is approximately

114% greater than Mistral's. The Qwen models fall in between, with CV values ranging from 0.006 to 0.007, suggesting a moderate level of dispersion.

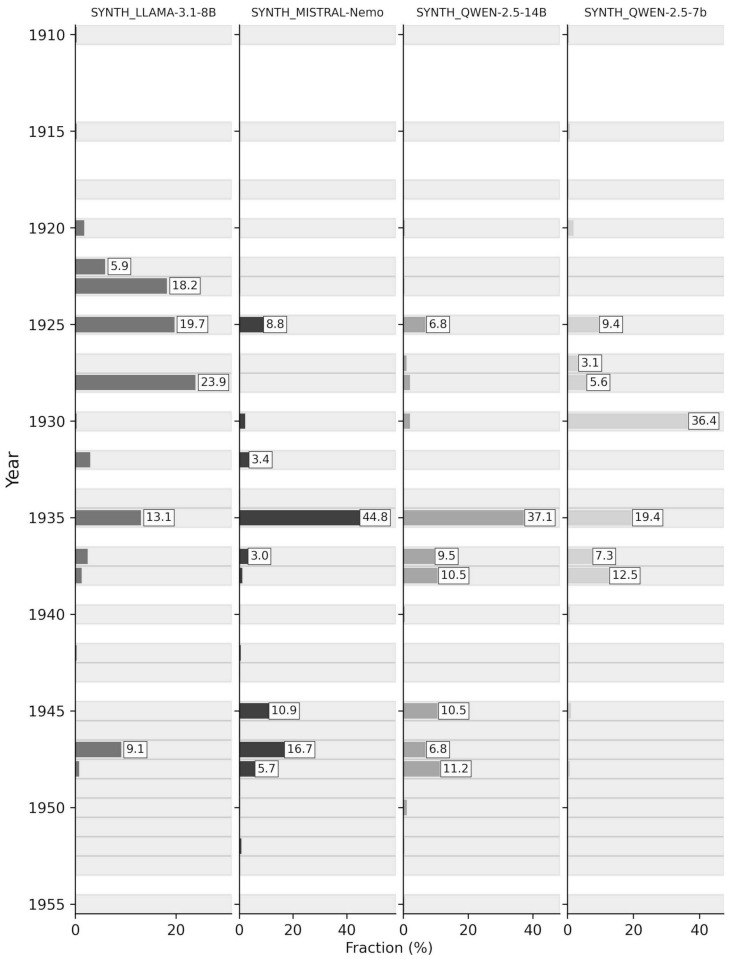
Figure 1: Prompt v0 – baseline: year autonomously selected by the LLM



While LLaMA demonstrates the highest degree of variability in its year selections, its most frequent choice also corresponds to 1955 in 28% of generations on average (SD=3.06 across runs). Qwen 2.5 7B and 14B models tend to concentrate on the year 1975, which is consistent across runs for both models (CVs are 7 and 4.6, respectively). Surprisingly, all models in this context completely avoid the interval 1941–1944, as well as other intervals (1926–1927, 1983–1984) and years (1921, 1924, 1929, 1931, 1933, etc.).

Notably, the digit 5 appears more frequently in the model-generated outputs compared to its expected frequency under a uniform distribution of years. We only consider the last two digits in each year for the calculation.

Figure 2: Prompt v3 – "20th century" substituted by the range 1900–1955



We measure the relative overrepresentation (%) of digits as follows:

$$(Observed - Expected)/Expected \times 100$$

where *observed* is the observed proportion of a digit under a given model and *expected* is the proportion of the digit's frequency under a uniform distribution of years. On average, the overrepresentation of 5 is approximately 370% with the highest for Mistral – 648% and the lowest for Qwen 7B – 186%. LLaMA avoids the digits 0, 1, and 9 with an occurrence rate of only 0.05%, 0%, and 0%. Similarly, Qwen Instruct models show a marked underrepresentation of the digits 2, 3, 4 appearing in 1% of cases on average. The digits 1 and 9 are also completely avoided. In contrast, these models display a heightened preference for the digit 7, with proportions of 26% (14b) and 33% (7b), respectively, compared to 9.6% in the Qwen Base model.

To further assess whether the selection of certain years reflects a general positional bias or a consistent preference independent of the interval, we restrict the original "20th

century" range to a narrower window: 1900–1955. As illustrated in Fig. 2, the models generate almost no selections for 1955 within this interval. Instead, the distribution shifts toward the central portion of the range. Notably, the larger models converge on a single year – 1935 – as their primary point of preference. When compared to the results from v0, similar distributional patterns emerge within the 1900–1955 interval, with the exception of the previous peak at 1955. Additionally, a small number of selections now appear around 1942–1943 and other years that were entirely absent in case of the earlier prompt configuration.

The absence of 1955 preference in the output for prompt v3 and the shift of the peak to the central part indicate potential positional bias. To determine whether the models' limited focus on specific years is influenced by the medical context in the prompt, we removed this contextual constraint and analyzed the resulting distribution (see Fig. 3). Without the medical framing, the models exhibited a broader and more diverse range of year selections.

The digit 5 remains overrepresented, although its proportion decreases across models. On average, the relative overrepresentation of 5 is 122%. The digits 1 and 9 are consistently underrepresented, each appearing in less than 1% of outputs across all models. LLaMA shows a strong preference for the digit 2, which is overrepresented by 276%, while all Qwen models overrepresent the digit 3 by an average of 204% (SD = 42%).

When the prompt is modified to generate the CoT reasoning prior to generating the content (Fig. 4) of the other fields and the model is explicitly instructed to reflect on an appropriate year and topic within the given context – the observed imbalance is substantially reduced, particularly in the case of Mistral. This effect is evidenced by the increase in entropy of the digit preference distribution under prompt version v5 relative to version v0, as shown in Tab. 4. The maximum Shannon entropy for a uniform distribution of digit preference is 2.3. It is also notable that for prompt versions v0 and v1 – where the CoT reasoning is generated after the content of all other fields – Qwen exhibits higher entropy in its digit preference distribution compared to Mistral. The Qwen 2.5 14B Base model achieves the highest entropy under prompt v0.

In summary, this analysis demonstrates that adding, removing, or repositioning contextual variables within the prompt affects both the accuracy of JSON generation and the models' preferences or aversions toward specific years and digits. The models produce substantially more valid JSON outputs when they are allowed to select the year themselves. Under the baseline prompt—which includes both the disease and patient organization context—stronger biases toward central years and particular digits emerge, especially in larger models. Conversely, removing medical context from the user and system prompts, as well as repositioning the CoT, increases digit entropy, most notably for the larger models.

Figure 3: Prompt v1 – medical framing excluded from system and user prompts

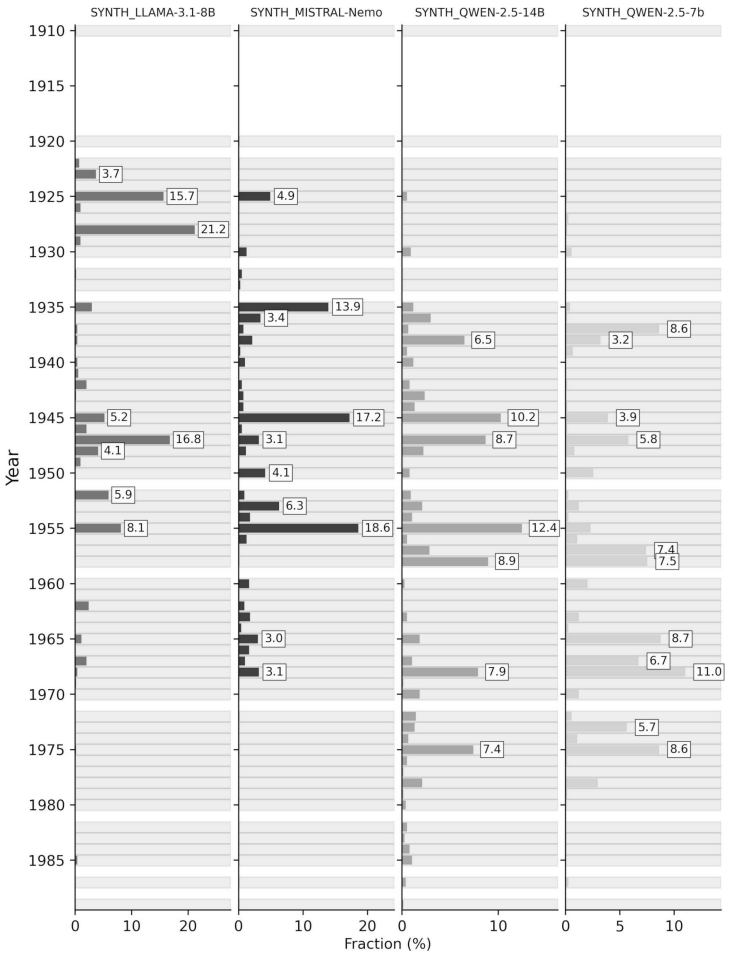
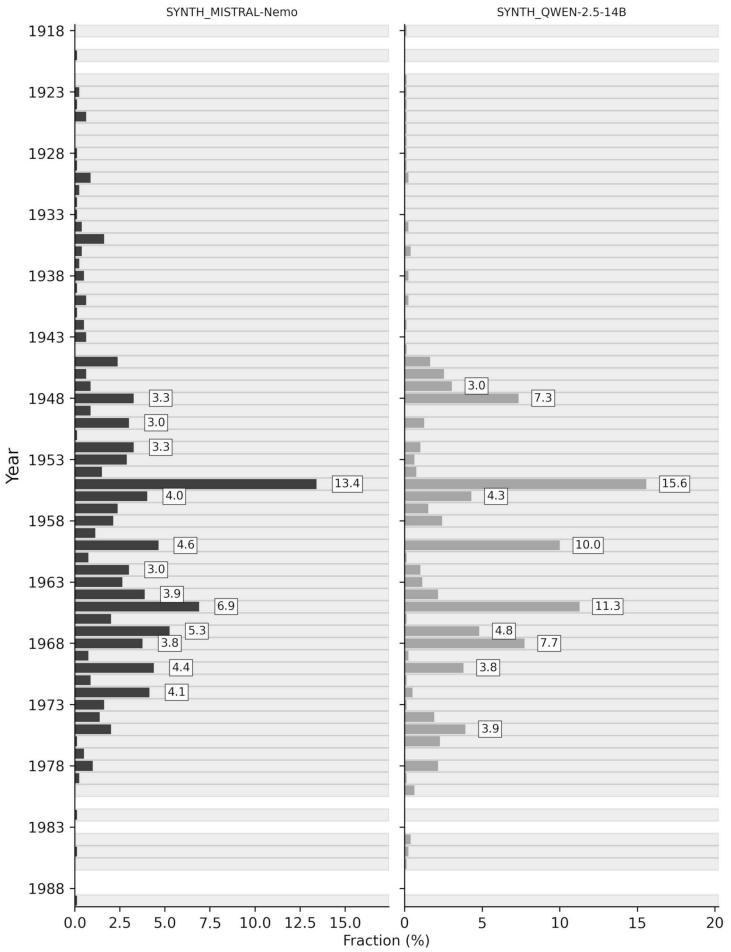


Table 4: Shannon entropy of digit preference distribution

Model\Prompt version	prompt v0	prompt v1	prompt v5
Mistral Nemo Instruct 2407	0.9	1.65	1.97
Qwen 2.5 14B-Instruct	1.4	1.84	1.80
Qwen 2.5 14B	1.76	-	-

Figure 4: Prompt v5 – LLM is instructed to generate the CoT first



5.1 Faithfulness

Preliminary Evaluation. Domain experts—historians of medicine—observe that SLMs achieve a higher fit to historical genre and time-period characteristics within the fiction genre, but perform less well in humor, interactive, and news genres.

In the patient-organization periodicals, fiction exhibits greater syntactic and semantic variation—in grammar, tone, style, voice, and format—than other genres. Although it contains less historically specific content, its themes related to disease and patient life overlap with those found in historical periodicals, likely explaining its resemblance to historical fiction.

By contrast, texts generated in the humor, interactive, and news genres tend to diverge considerably from their historical counterparts. Their tone, diction, and syntax often sound too contemporary, and the news outputs, in particular, sound too sensationalist.

Historians note that the perceived historical fit of the generated texts—aside from factual accuracy—largely depends on how effectively the model reproduces the genre conventions and content patterns of the periodicals, and how well it generalizes time-period features. However, these generalized features often yield vague or anachronistic depictions of medical knowledge and practice within the historical context of the *ActDisease* dataset.

Main evaluation. Historians completed the questionnaire as shown in Tab. 5. Their scores are compared across models and two prompt versions: v0 (SLM autonomously selects a plausible year in the 20th century) and v2 (SLM is explicitly instructed what year to produce data for). LLaMA consistently received lower scores across both prompts and all questions. Qwen 14B and Mistal-Nemo-Instruct-2407 achieved identical scores on most items, with the exception of RQ1.A, where historians rated Qwen slightly higher. For the smaller models (LLaMA and Qwen), historians noted differences between the responses generated under prompt versions v0 and v2. In contrast, the outputs of the larger models (Qwen and Mistral) appeared similar across prompt conditions. All models scored higher in RQ1.B and RQ1.C—questions that address genre fit and semantic similarity to the dataset. For the questions related to historical accuracy and plausibility (RQ1.A and RQ1.D), scores are at or below the midpoint of the scale [0,10].

Table 5: Responses to the Faithfulness Questionnaire

Model	**Llama 3.1 8B**		**Qwen 2.5 7B**		**Qwen 2.5 14B**		**Mistral 12B**	
RQ1\prompt configuration	**v0**	**v2**	**v0**	**v2**	**v0**	**v2**	**v0**	**v2**
A. To what extent do the generated outputs align with the factual historical timeline?	*2*	*2.5*	*4*	*4.5*	*5*	*4.5*	*4*	*4*
B. To what extent are the generated outputs semantically similar to the dataset they are intended to mimic?	*4*	*4.5*	*6*	*7*	*7*	*7*	*7*	*7*
C. How accurately do the genres of the outputs reflect the genres specified in the input prompt?	*5*	*6*	*6*	*7*	*7*	*7*	*7*	*7*
D.To what degree is the language of the outputs historically plausible?	*2*	*2.5*	*4*	*4.5*	*5*	*5*	*5*	*5*

Through the detailed analysis of the outputs, the historians identified three main kinds of distortion in the synthetic data: 1) literal hallucinations and errors of technolo-

gies, legislation, organizations, and disease concepts; 2) anachronisms of tone and style; and 3) overgeneralizations in time-period features and chains of reasoning.

Straightforward factual inaccuracies, especially in the Swedish data, include references to websites from 1955 (Mistral Nemo Instruct 2407) or references to "shellinkmeats-allergy" (Qwen 2.5 14B Instruct). In the English material, Qwen 2.5 7B hallucinates UK legislation in the form of a supposed 1985 'Patient Protection Act' – which, according to the LLM, was 'a comprehensive piece of legislation enacted in 1985 to protect the rights of individuals suffering from allergies and other chronic illnesses'. No such Act of Parliament was ever passed, although the hallucinated name, interestingly enough, alludes to the official title of the 2010 US legislation commonly known as 'Obamacare', suggesting something of what material the model may have been trained on, and the UK Parliament did approve a 'Food and Environment Protection Act' in 1985.

The model's representations of time periods and its reasoning chains were not factually incorrect, but they tended to be overly broad and only loosely connected to the accuracy of the details in the generated text or to the historical record. For example, Qwen 2.5 14B Instruct offers the following time period feature justification for a hypothetical medical case in a French diabetes journal describing insulin treatment in 1955: "The text mentions the common use of insulin in the treatment of diabetes and potential complications such as vision damage and kidney disease, which were common during this period when the long-term effects of the disease were just beginning to be understood". Long-term complications of diabetes had been discussed in medical literature as early as Avicenna's *Canon of Medicine* (1025), making this statement false if "diabetes" is taken as a set of related disease entities, symptoms, and illness experiences over time. If the reader takes a narrower view of a historically specific disease entity in 1955, however, they might defend it as accurate. By the time Sanger sequenced the chemical structure of insulin in 1955, industrial biomedicine (principally Novo Nordisk in Denmark) had developed purified, long-acting insulins that required fewer injections and allowed for better control of blood glucose levels. Here, "disease" would mean something like the biochemical understanding of diabetes used by pharmaceutical and biomedical researchers working on insulin purification: as the effects of the lack of a specific protein hormone and, after Sanger's sequencing in 1955, the lack of a 51 amino acid sequence in two polypeptide chains.

In addition, it is noteworthy that the patient-organization narratives produced by the considered SLMs frequently emphasize themes of continuity, consensus, and progress. For instance, in the British case of the 1980s, the decade is portrayed primarily as a period of rising attention to patient rights and the chronically ill, rather than one marked by financial constraints in the health service, reductions in disability benefits, and heightened activism from external pressure groups operating outside official structures.

Finally, the synthetic data assessment asks the historian to invert the sensibility they bring to the historical archive of trying to explain often implausible facts. In this case, the historians in this experiment know the synthetic data are not authentic primary sources, but it is easy to imagine many of the outputs as sources in historical archives. The assessment of their plausibility, however, cannot be independent of the context of the assessment (an experiment on synthetic data in which the historians know the source of the data). The historian's question is not, as the as the media scholar Roland

Meyer has described generative AI works, what is 'latent[ly] possib[le]' in the past at a particular historical moment, but rather what historical actors thought and did given what they left behind (Meyer, 2023). The contemporary trust in historical factchecking in the most literal sense of events in the past is an effect of the emergence and centralization of state archives as the centers of historical truth in the nineteenth century by the oft-cited founder of history as an academic discipline, Leopold von Ranke: archives were the places where the historian discovered primary sources that allowed them to write history "how it actually was" (Ranke, 1824; Eskildsen, 2008).

5.2 Diversity

The design of prompt configurations addresses the evaluation of diversity in the following way. We hypothesize that text types that define genres in the our baseline prompt contribute to diversifying the outputs and that omitting them will lead to more semantically similar responses (v4). Moreover, the removal of patient organization and disease context is expected to lead to more broad themes (v1) and potential decrease of semantic and vocabulary diversity than in the baseline. The specification of year in the prompt (v2) is expected to lead to covering a greater number of health-related themes. Prompt v5 allows us to explore the effect of prompt structure – specifically, the placement of reasoning – on diversity.

Pairwise Cosine Distances. Fig. 5 presents the probability density functions (PDFs) of pairwise cosine distance distributions across various language models and prompt configurations. Distributions concentrated toward the left imply that the generated texts are more semantically similar, while rightward concentrations reflect greater variability in content.

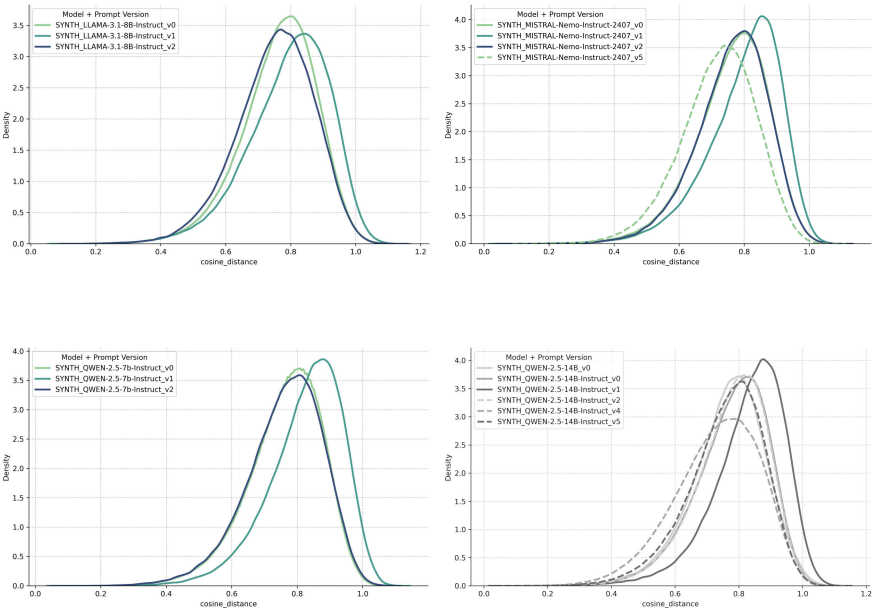
Across all models, prompt version v1 consistently yields a noticeable rightward shift in the cosine distance distribution indicating greater diversity. Prompt versions v0 and v2 tend to produce similar distributions with peaks near a mean distance of approximately 0.8. For the Qwen 2.5 14B Instruct model, we additionally compare with the Base model version and prompt v4. The Base model's distribution aligns closely with those of prompt v0 and v2, and is slightly shifted leftwards. The distribution for v4 shifts leftward, indicating a higher semantic similarity between outputs. A similar pattern is observed for prompt version v5 in the Mistral model.

Overall, the generated texts exhibit semantic diversity, as reflected by the distribution of pairwise cosine distances, which serve as a proxy for measuring variation in content.

Pairwise BERTScore. The overall trends in the PDFs of the self-BERTScore metric (Fig. 7, Appendix) are generally consistent with those observed for pairwise cosine distances, reinforcing our earlier findings regarding prompt versions v1 and v4. To further explore intra-model variability, we examined boxplots of self-BERTScore distributions across languages (Fig. 8, Appendix) and across genres under different prompt configurations in the larger models (Fig. 9, Appendix). Slightly higher diversity is observed for Swedish and German texts, suggesting greater semantic variability in these language subsets. Across

genres, Mistral shows a more pronounced variation in self-BERTScore than Qwen 14B (average coefficients of variation = 0.06 (SD=0.004) and 0.05 (SD=0.006), respectively), with the greatest dispersion occurring under prompt v5 for Mistral.

Figure 5: Probability density functions of pairwise cosine distance distributions across various language models and prompt configurations



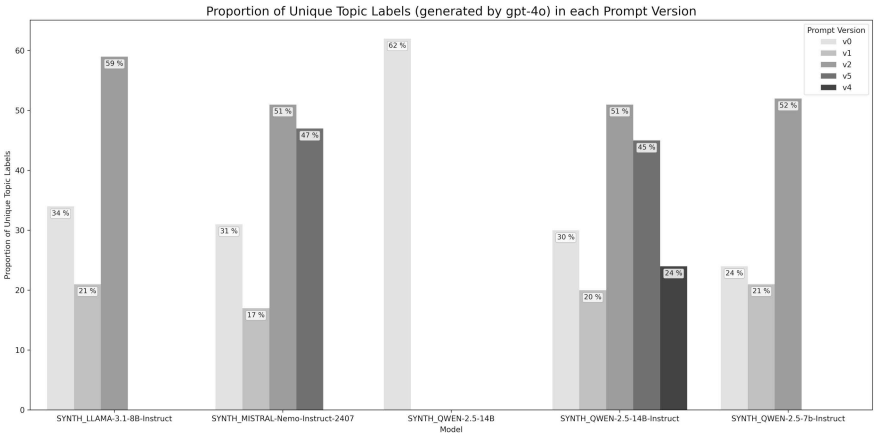
Topic Modeling. Fig. 6 presents the proportion of unique topics relative to the total number of generated texts for each model and prompt configuration. BERTopic produced a total of 549 topics for the entire synthetic dataset. Topic –1, typically representing outliers or unclassified data, was excluded from the plot.

The Qwen 14B Base model yields the highest proportion of unique topics, reflecting a greater degree of topic diversity. Prompts v2 and v5 also lead to relatively high levels of topic variation, indicating that these prompt configurations encourage more semantically varied generations. Prompt v4 results in a lower proportion of unique topics, consistent with the earlier analysis. Notably, prompt v1 produces the lowest proportion of unique topics across all models. Intuitively, we expect this since the models now lack specific medical context. However, it contrasts with the earlier cosine distances analysis.

Within the topic model, the top 10 topics account for a large number of generated texts and represent the broadest thematic groupings. The -1 topic is also assigned to a large portion of the dataset (21%). To examine whether v1 contributes disproportionately to these general categories, we calculated the deviation between the average proportion of texts assigned to the top 10 topics and outliers (-1) across all prompts and the proportion specifically associated with prompt v1. The results indicate that Prompt v1 drives an above-average shift toward general category assignments, yielding increases of approx-

imately 6% in LLaMA, 3% in Mistral, 2% in Qwen 14B, and a notable 11% in Qwen 7B In Qwens and LLaMA, all other prompts consistently yield lower-than-average allocations to these general categories. An exception to this trend is observed in the Mistral model, where prompt v5 shows a 5% increase in this share. Interestingly, earlier analysis of cosine distances indicates that v5 produces less semantically diverse outputs than v1 in Mistral. However, topic modeling results suggest greater topical diversity for v5 compared to v1. This apparent contradiction may indicate that Mistral, under prompt v5, generates a larger number of narrowly defined topics that are nonetheless semantically similar to one another. In contrast, prompt v1 may elicit broader, high-level themes that are more semantically distinct in the embedding space, resulting in greater pairwise cosine distances despite lower topic diversity. Importantly, in the computation of pairwise cosine distances, texts assigned to the -1 topic are also included; because this topic captures outliers or weakly structured content, its inclusion can inflate average pairwise distances and is therefore a likely contributing factor to the higher distances observed for prompt v1.

Figure 6: Proportion of unique topics in generated outputs across prompts versions and models



The explicit inclusion of the year in the prompt (v2) yields the most effective balance between generating a high proportion of unique topics and maintaining semantic diversity, based on cosine distance metrics.

Under prompt version v0, the Base model produces the highest proportion of unique topics—nearly double that of the other models. On the one hand, this finding adds to our earlier observation of its more uniform year selection (i.e., greater temporal diversity) and is consistent with prior literature highlighting the model's capacity to generate more diverse outputs. On the other hand, a manual review of the results revealed that the model frequently generated fabricated words for certain time periods, and even cited these invented forms as evidence in its explanations of temporal features as in the following example: "*This text reflects the 1950s through its use of the word 'invitaera' (to invite) instead of the modern Swedish 'invitera'. The use of 'Havsbadet' (seaside resort) indicates that this is an invitation to a seaside location.*" Such cases suggest that the model's output diversity may come

at the expense of increased hallucination. Addressing this trade-off should be a priority for the approaches that use base models for synthetic data generation.

In summary, the observed diversity patterns largely align with expectations based on the prompt design. When text-type specifications are omitted, the outputs become more semantically similar, and the absence of medical framing reduces topic coverage and leads to broader themes. When the prompt includes the full set of contextual variables—including the year—the outputs display greater topical diversity.

6. Conclusion

This study focuses on synthetic data generation for training LLMs to analyze historical documents. It evaluates the quality of zero-shot generated historical synthetic data intended to replicate genres found in twentieth-century European patient organization periodicals. The analysis shows that when models autonomously select a year based on the provided context, they exhibit clear year and digit preference biases, the strength of which varies across prompt configurations. These preferences are most pronounced under the baseline prompt with medical framing. Additionally, the models tend to omit certain historical periods—most notably, all models completely exclude the years 1941–1944 under the baseline prompt. The accuracy of valid JSON generation also varies depending on whether the year is model-selected or predefined in the prompt, with the highest accuracy observed when the models choose the year themselves.

The faithfulness evaluation suggests that none of the evaluated models are rated as historically accurate or plausible, as indicated by the low scores (≤5) for the questions directly addressing these criteria (RQ1.A and RQ1.D). The main problems in the outputs include literal hallucinations and errors of technologies, legislation, organizations, and disease concepts, as well as anachronisms of tone and style, and overgeneralizations both in the main texts and the CoT. Higher scores are observed primarily for questions related to aspects related to semantic proximity to the source material (RQ1.B), as well as the correspondence of the generated genre to the prompted one (RQ1.C). Among the larger models, Qwen 14B and Mistral were observed to have similar outputs across prompt configurations, whereas, for the smaller models (Llama 8B and Qwen 7B), prompt configuration v2 (year is pre-defined in the prompt) resulted in more convincing outputs on average, according to the questionnaire. Overall, Llama ranks lower relative to other models, while Qwen 14B is rated slightly higher than Mistral.

Semantic diversity varies considerably across prompts and its patterns generally align with what we expect from the prompt configurations. Excluding the medical context in prompt v1 shifts the output focus toward broader, less domain-specific topics. Interestingly, while these topics may appear generic, pairwise cosine distance metrics reveal that they remain semantically distant from one another, indicating retained diversity at a global level. Prompt v2, which includes explicit year control and extensive parameter specification, yields the best balance of semantic diversity and structural coherence. Additionally, prompt version v5, where CoT reasoning is generated first, also performs well. It demonstrates high semantic diversity and favorable entropy in

digit distribution, suggesting that early CoT placement may help scaffold more varied outputs.

According to the metrics, Qwen 2.5 14B Base model exhibits higher entropy in digit preference distributions than its instruction-tuned counterpart, as well as generates a larger proportion of unique topics. Manual analysis further reveals that the diversity observed in the outputs may partly reflect hallucinations, which were considerably less prominent in the outputs of Instruct models, highlighting an important direction for future research.

In conclusion, the results raise important concerns regarding the suitability of the evaluated models for generating synthetic datasets that include historical, temporal, and numerical information. Numerical biases are particularly pronounced under the baseline prompt with explicit medical framing, suggesting a need for further investigation—especially given the growing use of LLMs in the medical and healthcare domains for synthetic data generation.

Although the base model appears more diverse according to automated metrics, manual inspection reveals frequent hallucinations and fabricated vocabulary, which may account for this apparent diversity. This finding underscores the importance of complementing automatic evaluations with human assessment, as prior studies have often relied primarily on machine-based metrics when comparing base and instruction-tuned models.

The synthetic dataset produced in this study was judged by historians as relatively plausible in terms of genre conventions and semantic similarity to original patient organization periodicals, but less so in terms of historical accuracy and factual plausibility. Consequently, it may serve as a valuable resource for augmenting training data in genre classification tasks, but is not suitable for post-training of large language models.

Overall, the generation and use of synthetic datasets with these models should be accompanied by rigorous human and automated evaluations incorporating multi-dimensional metrics tailored to the research domain. Such an approach is essential to prevent the propagation of historical inaccuracies into future datasets and computational tools.

Acknowledgements

Co-funded by the European Union (ERC, ActDisease, ERC-2021-STG 10104099). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them. Co-funded by the Wallenberg AI, Autonomous Systems and Software Program – Humanity and Society (WASP-HS Research Group 807).

References

Aangenendt G, Skeppstedt M and Söderfeldt Y (2024) Curating a Historical Source Corpus of 20th Century Patient Organization Periodicals. In *Proceedings of the Huminfra Conference (HiC 2024)*: 76–82. DOI: https://doi.org/10.3384/ecp205011.

Barr AA, Quan J, Guo E and Sezgin E (2025) Large Language Models Generating Synthetic Clinical Datasets: A Feasibility and Comparative Analysis with Real-World Perioperative Data. *Frontiers in Artificial Intelligence*: 8. DOI: https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2025.1533508.

Binici K, Kashyap A, Schlegel V, Liu A, Dwivedi V, Nguyen T-T, Gao X, Chen N and Winkler S (2025) MEDSAGE: Enhancing Robustness of Medical Dialogue Summarization to ASR Errors with LLM-Generated Synthetic Dialogues. *Proceedings of the AAAI Conference on Artificial Intelligence* 39: 23496–23504. DOI: https://doi.org/10.1609/aaai.v39i22.34518.

Danilova V and Söderfeldt Y (2025) Classifying Textual Genre in Historical Magazines (1875–1990). In *Proceedings of the 9th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2025)*. Albuquerque, NM, USA: Association for Computational Linguistics, pp. 160–71. DOI: https://doi.org/10.18653/v1/2025.latechclfl-1.15.

Eskildsen, KR (2008) Leopold Ranke's archival turn: location and evidence in modern historiography. *Modern Intellectual History* 5(3): 425–453.

Gayu Z and Yong L (2025) Towards a Theoretical Understanding of Synthetic Data in LLM Post-Training: A Reverse-Bottleneck Perspective. In: *The Thirteenth International Conference on Learning Representations*, Singapore, April 2024. DOI: https://openreview.net/forum?id=UxkznlcnHf.

Grootendorst M (2022) BERTopic: Neural Topic Modeling with a Class-Based TF-IDF Procedure. arXiv. https://arxiv.org/abs/2203.05794.

Kessler B, Nunberg G and Schutze H (1997) Automatic Detection of Text Genre. In: *35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*, Madrid, Spain: Association for Computational Linguistics, pp. 32–38. https://doi.org/10.3115/976909.979622.

Li Z, Zhu H, Lu Z and Yin M (2023) Synthetic Data Generation with Large Language Models for Text Classification: Potential and Limitations. In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Singapore: Association for Computational Linguistics, pp. 10443–61. DOI: https://doi.org/10.18653/v1/2023.emnlp-main.647.

Long L, Wang R, Xiao R, Zhao J, Ding X, Chen G and Wang H (2024) On LLMs-Driven Synthetic Data Generation, Curation, and Evaluation: A Survey. In: *Findings of the Association for Computational Linguistics: ACL 2024*, Bangkok, Thailand: Association for Computational Linguistics, pp. 11065–82. DOI: https://doi.org/10.18653/v1/2024.findings-acl.658.

Lu Y, Bartolo M, Moore A, Riedel S and Stenetorp, P (2022) Fantastically Ordered Prompts and Where to Find Them: Overcoming Few-Shot Prompt Order Sensitivity. In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Vol-*

ume 1: Long Papers), Dublin, Ireland: Association for Computational Linguistics, pp. 8086–98. https://aclanthology.org/2022.acl-long.556.

Meyer R (2023) The New Value of the Archive: AI Image Generation and the Visual Economy of 'Style'. *IMAGE: Zeitschrift für interdisziplinäre Bildwissenschaft*. 19(1): 100–111. http://dx.doi.org/10.25969/mediarep/22314.

Nikolenko S (2021) *Synthetic Data for Deep Learning*. New York: Springer. DOI: https://doi.org/10.1007/978-3-030-75178-4.

Park PS, Schoenegger P and Zhu C (2024). Diminished Diversity-of-Thought in a Standard Large Language Model. *Behavior Research Methods* 56: 5754–70. DOI: https://doi.org/10.3758/s13428-023-02307-x.

Petrenz P and Webber B (2011) Squibs: Stable Classification of Text Genres *Computational Linguistics* 37 (2): 385–93. DOI: https://doi.org/10.1162/COLI_a_00052.

Ranke Lv (1824) *Geschichten der romanischen und germanischen Völker von 1494–1535*. Reimer.

Schaffelder M and Gatt A (2025) Synthetic Eggs in Many Baskets: The Impact of Synthetic Data Diversity on LLM Fine-Tuning. *arXiv*. DOI: https://arxiv.org/abs/2511.01490.

Wagner S, Behrendt M, Ziegele M and Harmeling S (2024) The Power of LLM-Generated Synthetic Data for Stance Detection in Online Political Discussions. DOI: https://doi.org/10.48550/arXiv.2406.12480.

Wang, F, Lin M, Ma Y, Liu H, He Q, Tang X, Tang J, Pei J and Wang S (2025) A Survey on Small Language Models in the Era of Large Language Models: Architecture, Capabilities, and Trustworthiness. *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, Toronto, Canada, pp. 6173–6183. DOI: https://doi.org/10.1145/3711896.3736563.

Wei J, Bosma M, Zhao V, Guu K, Wei Yu A, Lester B, Du N, Dai AM and Le QV (2021) Finetuned Language Models Are Zero-Shot Learners. *arXiv*. DOI: https://arxiv.org/abs/2109.01652.

West P and Potts C 2025 Base Models Beat Aligned Models at Randomness and Creativity. *arXiv*. DOI: https://arxiv.org/abs/2505.00047.

Zhang T, Peng B and Bollegala D (2024) Improving Diversity of Commonsense Generation by Large Language Models via In-Context Learning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, Miami, Florida, USA, pp. 9226–9242.

Zhang T, Kishore V, Wu F, Weinberger KQ and Artz Y (2020) BERTScore: Evaluating Text Generation with BERT. *arXiv*. DOI: https://arxiv.org/abs/1904.09675.

Zhu A, Asawa P, Quincy Davis J, Chen L, Hanin B, Stoica, I, Gonzalez JE and Zaharia M (2025) BARE: Combining Base and Instruction-Tuned Language Models for Better Synthetic Data Generation. *arXiv preprint arXiv:2502.01697*.

Appendix

Figure 7: KDE plots of pairwise BERTScore

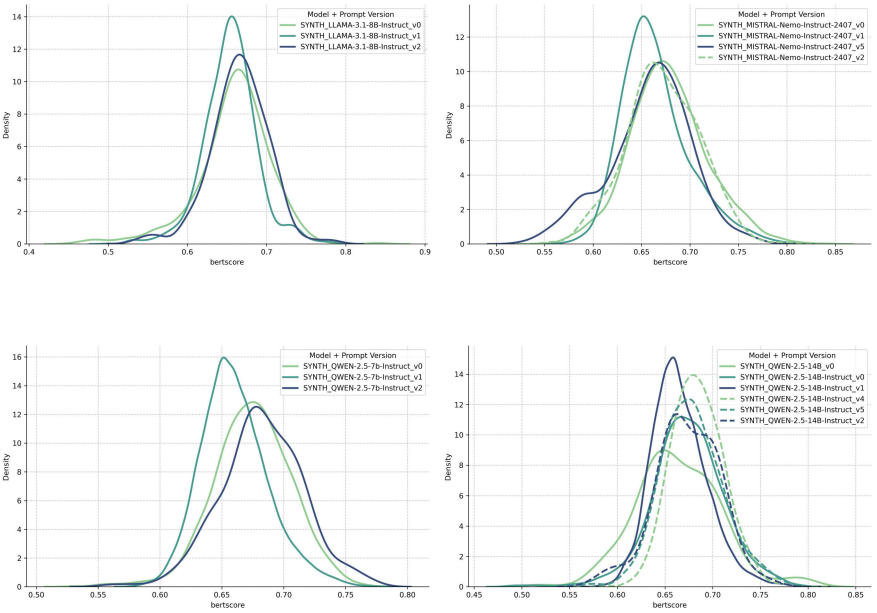


Figure 8: Comparison of pairwise BERTScore distributions across languages

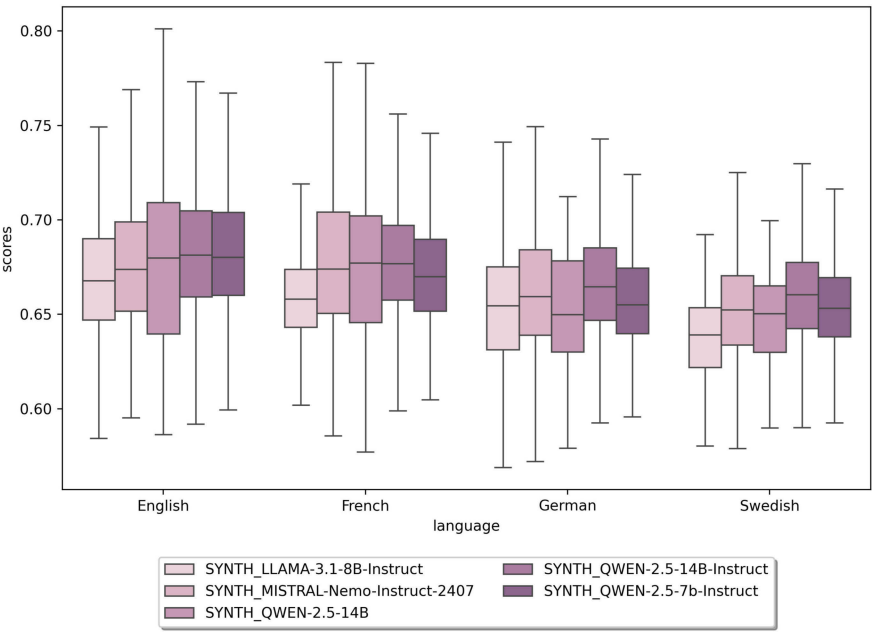


Figure 9: Pairwise BERTScore distributions across genres and prompt versions