Title

Digitizing and Generating Social Conflict Data with Artificial Intelligence

Journal

Chinese Political Science Review

ISSN

2365-4244

Author

Steinert-Threlkeld, Zachary C

Publication Date

2026

DOI

10.1007/s41111-026-00340-7

Digitizing and Generating Social Conflict Data with Artificial Intelligence

Zachary C. Steinert-Threlkeld[1]

[1]Luskin School of Public Affairs, University of California, Los Angeles, zst@luskin.ucla.edu, ORCID: 0000-0002-5122-9660

Abstract

The difficulty of collecting social conflict data has caused its study to focus on recent events predominantly in the West. This paper shows how to use artificial intelligence to generate social conflict data regardless of language, location, or date. Digitizing tertiary books, lightly cleaning their text representation, and submitting that text to a large language model (LLM) produces accurate date, location, and event descriptions. These capabilities and results are demonstrated with books on Latin America after 1492, Imperial Russia, and Tokugawa Japan. Using LLMs requires a larger fixed cost than working a team of research assistants, but it produces results more quickly for most sources. It is less accurate for the Tokugawa Japan source; whether it or human translation is cheaper depends on the cost of correcting the LLM's work. These results represent the floor of accuracy for artificial intelligence, suggesting researchers should soon use it for creating social conflict data from tertiary sources.

Keywords: social conflict; event data; large language models; optical character recognition; historical data

Acknowledgments

Thanks are owed to Rock Vitale and Easie for AI, cloud, and Python support generating data from Japan; Kao Kawasumi, Risako Takamura, and Harun Vermupalli for invaluable research assistance; and Cédric Chambru and Yuri Zhukov for feedback.

Competing Interests

There are no competing interests to declare.

1 Introduction

Social conflict — petitions, protests, revolts, riots, strikes — is a core mechanism of political feedback. Even in democracies, collective action enables marginalized actors to shape agendas, communicate policy discontent, and contest the boundaries of legitimate authority. Yet the empirical study of social conflict remains uneven: extensive datasets and robust literatures exist for well-documented episodes and favored regions. Most data and studies focus on Europe and North America after 1789, with years such as 1848, 1968, 1989, and 2011 receiving the preponderance of quantitative attention. Except for recent datasets that extract data from newspapers or social media (Raleigh et al., 2010, Boschee et al., 2015, Weidmann and Rod, 2018), social conflict data outside of Europe and North America is hard to find. Social conflict data from earlier periods or outside those regions are even harder to find.

This article proposes using artificial intelligence for recovering social conflict from eras and regions neglected in the study of social conflict. Recent advances in hardware, optical character recognition, and artificial intelligence (large language models) make it easy to digitize tertiary sources, translate them, and extract meaning with much less effort than visiting archives, searching catalogues, scanning, and coding from reading. This article demonstrates the broad applicability of this methodology by generating social conflict data from post-contact Latin America (1492-1936), Imperial Russia (1796-1917), and Tokugawa Japan (1590-1877). The tertiary sources are in French, Russian, and Japanese, respectively, but the resulting data are in English.

The large language model results are validated using manual review from individuals who can read the source material. The artificial intelligence is accurate enough on the French and Russian sources that its output can be used without further review. The Japanese results are not. Despite extra data preparation given to the Japanese scans, the translated results would still require manual review before using in research. A comparison of the cost curves for creating data manually with research assistants versus artificial intelligence shows that scanning, prompt engineering, and large language models is always cheaper for easier sources

such as Latin America and Imperial Russia, but for more complex sources like Tokugawa Japan, human coding is cheaper when the source is between 32 and 309 pages. For hard texts like the Japanese source, whether or not this paper's methodology is cheaper depends on the cost of correcting the artificial intelligence's output. For the Japanese source this paper uses, using artificial intelligence is only cheaper when the time to correct the artificial intelligence translation is less than 72% the time of directly translating. The research assistant used for translating the Japanese source requires 30%.

This paper makes two contributions to the literatures on social conflict event data and the use of large language models in research. First, it demonstrates a scalable, accurate approach that takes advantage of recent hardware and artificial intelligence advances to digitize social conflict event data from books. When tertiary sources catalogue events in a structured format, converting the prose into accurate social data is now possible. While specific event details depend on the book, those in this paper allow for the creation of event data recording date, location (to at least the province level), event type, and target. Second, it calculates cost curves for the artificial intelligence workflow and teams of research assistants manually reading the sources and creating event data. These curves are based on the Tokugawa Japan book, the source for which the artificial intelligence workflow requires the most configuration, testing, and validation. Comparing the fastest human coding to the slowest artificial intelligence, the artificial intelligence is cheaper for a book longer than 75 pages, about 595 events.

2 Motivation

Any dataset represents one possible sample from all potential data sources. This sample is not problematic if the created datasets are a random sample of all possible datasets. There are reasons to believe they are not.

The leading approach to making social conflict data is to rely on newspapers (Demarest

and Langer, 2022), which introduces two concerns. The first is recency bias. For example, the Mass Mobilization in Autocracies dataset focuses on the 20th century, and the Integrated Conflict Early Warning System starts in 1995 (Boschee et al., 2015, Weidmann and Rod, 2018). Even datasets that extend further back do not extend far. The Armed Conflict Location and Event Data (ACLED) stops in 1900, and the majority of their observations start after the end of the Cold War (Raleigh et al., 2010). There are also datasets that cover older, specific events or movements such as the 1830-1832 Swing Riots, the 1871 Paris Commune, America's Women's Suffrage movement of the late 19th and early 20th century, American urban unrest in the 1960s, and so on (Snyder and Kelly, 1977, Gould, 1991, McCammon et al., 2001, Holland, 2005, Aidt, Leon-Ablan and Satchell, 2021).

The second concern is selection bias. Decades of research have shown that newspapers miss a substantial number of events that are not violent, near major cities, are ongoing, or are not perceived as relevant to the newspapers' readers (Myers and Caniglia, 2004, Herkenrath and Knoll, 2011, Hellmeier, Weidmann and Geelmuyden Rød, 2018). Reliance on them also reinforces a Western-centric study of social conflict (Nam, 2006). A growing body of research has therefore started to build social conflict data from other sources such as social media, grassroots organizations' reports, and state, often police, archives (Davenport and Ball, 2002, Sullivan, 2016, Zhang and Pan, 2019, Göbel and Steinhardt, 2022, Steinert-Threlkeld and Joo, 2022).

This paper introduces a new type of source, tertiary books which catalog social conflict based on a compendium of primary and secondary sources such as diaries, monographs, and, most commonly, state records from internal communication and security archives.[1] Because the widespread circulation of newspapers is a 19th century creation, the class of sources used in this paper enable the study of events that predate newspapers. For example, this paper includes data from Latin America from 1492, Columbus' first contact, Russia from 1796, and Japan starting in 1590. Using artificial intelligence therefore extends the study of event data

[1]A primary source for China, Japan, South Korea, and Vietnam is each country's Veritable Records, which are discussed later.

several centuries beyond the current horizon.[2]

Tertiary resources ameliorate selection bias since they rely on primary and secondary sources. While individual sources can contain a bias, on average the use of heterogenous sources should lead to less bias than that from relying on newspapers. Tertiary resources often include events from state archives. If states are more likely to record certain types of events, such as those above a size or disorder threshold, then the resulting catalog will also include a biased record. For most of recorded history, however, the state was the only consistent recorder, so researchers need to be cognizant of this potential bias but not let it stop the use of these resources.

Aside from reducing recency and selection bias, the use of tertiary resources enables advancements in several research domains.

A concern with relying on newspapers to study social conflict is that the media are a primary channel for the diffusion of information that helps spread events. Relying on newspapers therefore makes it difficult to disentangle the effect of media coverage, social networks, and infrastructure on the spread of conflict. It is likely, for example, that events diffuse more slowly before the advent of the railroad or telegraph, and they are likely to be more constrained by waterways then than now. The difficulty of diffusion suggests that protests were less likely to spread, but whether or not that is the case is not testable without the construction of social conflict datasets from pre-newspaper eras.

The temporal limitation of existing datasets likely also biases researchers' understanding of the issues over which people protest. For many events of interest, such as the United States Civil Rights movement or anti-regime protests in the Union of Soviet Socialist Republics, the issue is clear and therefore unquestioned. In cases where datasets collect information that can suggest the issue being protested, the issues are limited to modern ones (Salehyan et al., 2012). It remains unknown, however, the extent to which political, religious, economic, or identity issues animate in the pre-modern or non-Western world.

[2]The Historical Conflict Event Dataset codes military conflict since 1468 BC but is not included in the following review because it focuses on interstate conflict (Miller and Bakar, 2023).

In addition, the repertoires of social conflict change over time. Certain tactics, like work stoppages or mass demonstrations, can only come into being with industrialization and urbanization (Beissinger, 2022). Others, like sit-ins, holding white signs, or removing hijabs, are particularistic. Rural conflict often focuses on granaries or estates. It is therefore conceivable that the manifestation of discontent in pre-modern or non-Western contexts takes forms not observed in existing datasets or is distributed differently.

3 Historic Social Conflict Data

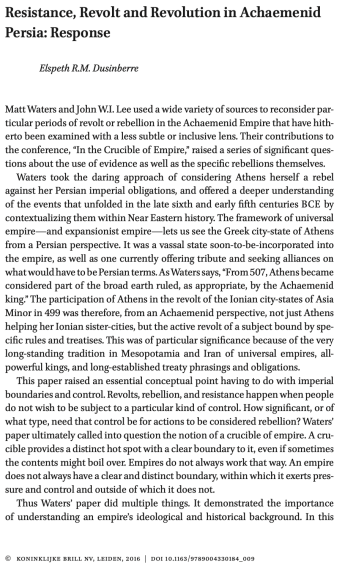
There are two datasets that record social conflict in old or non-European contexts. The Seshat project brings together dozens of anthropologists, archeologists, and historians to produce datasets on economic, political, and social features of societies from as far back as 10,000 years ago (Turchin et al., 2013). Several types of conflict are tracked; the five most closely related to social conflict are internal wars, intra-elite wars, popular uprisings, militant revolt, and separatist rebellions. The data are sparse, however. Each event comprises one row, even if the event itself lasted decades. Details are contained in narrative entries as opposed to columned data, and the sampling criteria for societies and events is not clear. The Seshat project is awe-inspiring in its scope but is ultimately too broad for the study of the dynamics of social conflict.

The Historical Social Conflict Database, HiSCoD, records social conflict spanning 1090-1890 from 32 historical entities in 25 contemporary countries (Chambru and Maneuvrier-Hervieu, 2024). Most of the countries are Central and Western European, and those that are not are colonies (Egypt, Gambia, Haiti, Mozambique). Just over 75% of events, however, are from France; over 95% are from France, England, and Italy; and almost half of events are from after 1789. HiSCoD brings some of the past into the study of social conflict, for three countries.

To create historic social conflict data across multiple countries first requires identifying

tertiary sources. These sources have performed the laborious work of extracting social conflict from archives. These tertiary resources fall into three types. Narratives are historical accounts. Sometimes they are about a social conflict, but they can also be biographies or histories that coincidentally contain conflict data. Maps show events spatially. Chronologies are books with temporally ordered entries that follow a consistent structure. Chronology entries can be in paragraph form or as tables. Figure 1 shows an example of each.

Figure 1: Three Types of Tertiary Sources



CHAPTER 8

Resistance, Revolt and Revolution in Achaemenid Persia: Response

Elspeth R.M. Dusinberre

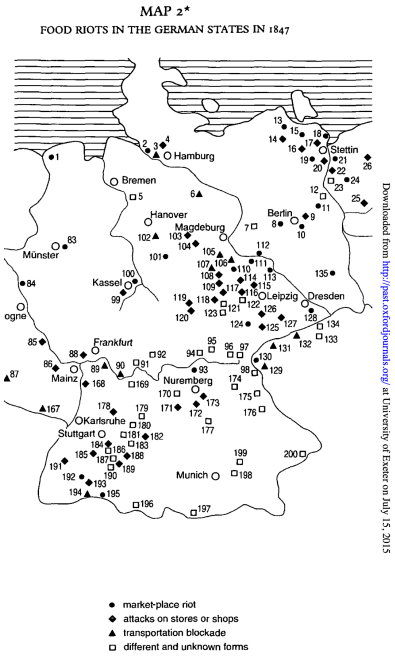
Matt Waters and John W.I. Lee used a wide variety of sources to reconsider particular periods of revolt or rebellion in the Achaemenid Empire that have hitherto been examined with a less subtle or inclusive lens. Their contributions to the conference, "In the Crucible of Empire," raised a series of significant questions about the use of evidence as well as the specific rebellions themselves.

Waters took the daring approach of considering Athens herself a rebel against her Persian imperial obligations, and offered a deeper understanding of the events that unfolded in the late sixth and early fifth centuries BCE by contextualizing them within Near Eastern history. The framework of universal empire—and expansionist empire—lets us see the Greek city-state of Athens from a Persian perspective. It was a vassal state soon-to-be-incorporated into the empire, as well as one currently offering tribute and seeking alliances on what would have to be Persian terms. As Waters says, "From 507, Athens became considered part of the broad earth ruled, as appropriate, by the Achaemenid king." The participation of Athens in the revolt of the Ionian city-states of Asia Minor in 499 was therefore, from an Achaemenid perspective, not just Athens helping her Ionian sister-cities, but the active revolt of a subject bound by specific rules and treatises. This was of particular significance because of the very long-standing tradition in Mesopotamia and Iran of universal empires, all-powerful kings, and long-established treaty phrasings and obligations.

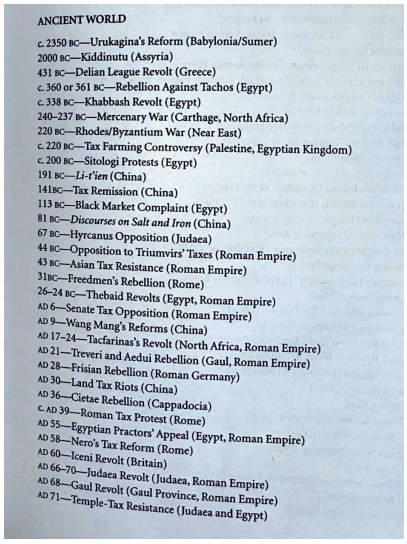
This paper raised an essential conceptual point having to do with imperial boundaries and control. Revolts, rebellion, and resistance happen when people do not wish to be subject to a particular kind of control. How significant, or of what type, need that control be for actions to be considered rebellion? Waters' paper ultimately called into question the notion of a crucible of empire. A crucible provides a distinct hot spot with a clear boundary to it, even if sometimes the contents might boil over. Empires do not always work that way. An empire does not always have a clear and distinct boundary, within which it exerts pressure and control and outside of which it does not.

Thus Waters' paper did multiple things. It demonstrated the importance of understanding an empire's ideological and historical background. In this

(a) Narrative



(b) Map



ANCIENT WORLD

c. 2350 BC—Urukagina's Reform (Babylonia/Sumer)
2000 BC—Kiddinutu (Assyria)
431 BC—Delian League Revolt (Greece)
c. 360 or 361 BC—Rebellion Against Tachos (Egypt)
c. 338 BC—Khabbash Revolt (Egypt)
240–237 BC—Mercenary War (Carthage, North Africa)
220 BC—Rhodes/Byzantium War (Near East)
c. 220 BC—Tax Farming Controversy (Palestine, Egyptian Kingdom)
c. 200 BC—Sitologi Protests (Egypt)
191 BC—*Li-t'ien* (China)
141BC—Tax Remission (China)
113 BC—Black Market Complaint (Egypt)
81 BC—*Discourses on Salt and Iron* (China)
67 BC—Hyrcanus Opposition (Judaea)
44 BC—Opposition to Triumvirs' Taxes (Roman Empire)
43 BC—Asian Tax Resistance (Roman Empire)
31BC—Freedmen's Rebellion (Rome)
26–24 BC—Thebaid Revolts (Egypt, Roman Empire)
AD 6—Senate Tax Opposition (Roman Empire)
AD 9—Wang Mang's Reforms (China)
AD 17–24—Tacfarinas's Revolt (North Africa, Roman Empire)
AD 21—Treveri and Aedui Rebellion (Gaul, Roman Empire)
AD 28—Frisian Rebellion (Roman Germany)
AD 30—Land Tax Riots (China)
AD 36—Cietae Rebellion (Cappadocia)
c. AD 39—Roman Tax Protest (Rome)
AD 55—Egyptian Practors' Appeal (Egypt, Roman Empire)
AD 58—Nero's Tax Reform (Rome)
AD 60—Iceni Revolt (Britain)
AD 66–70—Judaea Revolt (Judaea, Roman Empire)
AD 68—Gaul Revolt (Gaul Province, Roman Empire)
AD 71—Temple-Tax Resistance (Judaea and Egypt)

(c) Chronology

Narratives provide the most detail about an event but are not standardized, so extracting data from them is not efficient. Maps present more events at once, but they often lack detail, including labels for points that contain events. Extracting more detail requires consulting the sources used to make the map, putting the researcher back into the narrative realm. Chronologies are the most amenable to generating social conflict because they focus on specific types of events or historical periods. While they are as long as narratives, each page contains multiple events, which is not guaranteed in narratives. Crucially, each chronology standardizes the presentation of an event, facilitating the programmatic extraction of data.

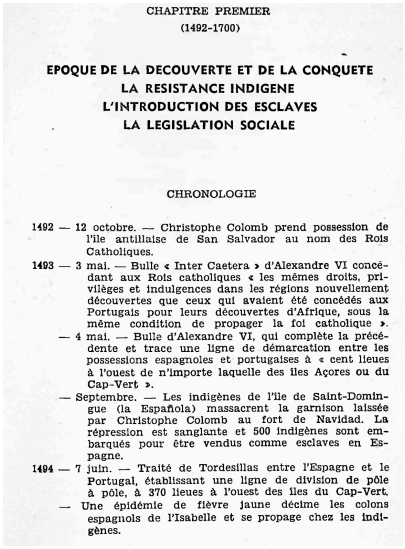
Having identified chronologies as best suited for creating historic social conflict data at scale, it is necessary to determine which chronologies to use. To demonstrate the broad applicability of artificial intelligence in the creation of historic social conflict data, three chronologies from different regions and eras and written in a language other than English are used. In the New World, very few documents survived contact with Europeans, so no suitable documents before 1492 appear to exist. Instead, a 1958 French chronology, *Mouvements Ouvriers et Socialistes (Chronologie et Bibliographie): L'Amérique Latine (1492-1936)*, chronicles political developments, including social conflict, since October 12, 1492 (Rama, 1959). Closer to Europe, the Russian series *Krestianskoe dvizhenie v Rossii v XIX-nachale XX veka* chronicles peasant unrest in Eurasian Russia from 1796-1917 (Shapkarin A. V., 1959, Nifontov A. S., 1960, Predtechenskiĭ and Kudriavtseva, 1961, Valk S. N., 1961, Okun S. B., 1962, 1963, Ivanov L. M., 1964, Shapkarin, 1966, Zaionchkovskiĭ and Paina, 1968, Anfimov, 1998). Compiled by Soviet historians, the books rely on several state archives and the archive of the Russian Orthodox Church, as well as tertiary sources; for more information, see Finkel, Gehlbach and Olsen (2015). Finally, Koji Aoki has published a book in Japanese of peasant unrest and urban riots in Japan from 1590-1877, a little more than the Tokugawa (Edo) period (Aoki, 1981). For more information, see White (1995).[3]

Figure 2 shows these three tertiary datasets in their raw chronological form. Each clearly shows events demarcated by dates, but that is where the similarity ends. Each chronology uses a different language and different format. The Latin America and Russia chronologies consist primarily of prose, whereas the Japanese chronology is a large table. A Japanese text, it is also oriented differently than Western tables: each event is a column, not row.

This paper excludes two promising source of historical social conflict data in East Asia, Veritable Records (*shilu*) and Chinese gazetteers. Veritable Records refer to official histories compiled by Chinese imperial courts from an emperor's communication with subordinates

[3]A third motivation of this project is to preserve social conflict data for future generations. White (1995) created a dataset on magnetic tape from Aoki (1981). When the author contacted White to request the data, he was informed it was lost. A smaller subset of the data has been digitized for (Steele, Paik and Tanaka, 2017).

Figure 2: Post-Contact Latin America, Imperial Russia, and Tokugawa Japan Chronologies



CHAPITRE PREMIER
(1492-1700)

EPOQUE DE LA DECOUVERTE ET DE LA CONQUETE
LA RESISTANCE INDIGENE
L'INTRODUCTION DES ESCLAVES
LA LEGISLATION SOCIALE

CHRONOLOGIE

1492 — 12 octobre. — Christophe Colomb prend possession de l'île antillaise de San Salvador au nom des Rois Catholiques.

1493 — 3 mai. — Bulle « Inter Caetera » d'Alexandre VI concédant aux Rois catholiques « les mêmes droits, privilèges et indulgences dans les régions nouvellement découvertes que ceux qui avaient été concédés aux Portugais pour leurs découvertes d'Afrique, sous la même condition de propager la foi catholique ».

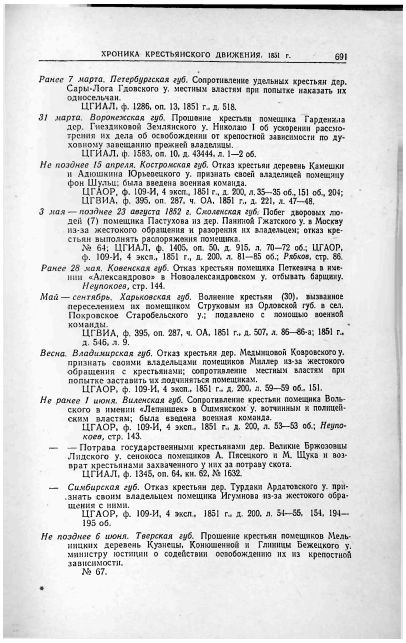
— 4 mai. — Bulle d'Alexandre VI, qui complète la précédente et trace une ligne de démarcation entre les possessions espagnoles et portugaises à « cent lieues à l'ouest de n'importe laquelle des îles Açores ou du Cap-Vert ».

— Septembre. — Les indigènes de l'île de Saint-Domingue (la Española) massacrent la garnison laissée par Christophe Colomb au fort de Navidad. La répression est sanglante et 500 indigènes sont embarqués pour être vendus comme esclaves en Espagne.

1494 — 7 juin. — Traité de Tordesillas entre l'Espagne et le Portugal, établissant une ligne de division de pôle à pôle, à 370 lieues à l'ouest des îles du Cap-Vert.

— Une épidémie de fièvre jaune décime les colons espagnols de l'Isabelle et se propage chez les indigènes.

(a) Latin America Tertiary Data



ХРОНИКА КРЕСТЬЯНСКОГО ДВИЖЕНИЯ. 1851 г. 691

Ранее 7 марта. Петербургская губ. Сопротивление удельных крестьян дер. Сары-Лога Гдовского у. местным властям при попытке наказать их односельчан.
ЦГИАЛ, ф. 1286, оп. 13, 1851 г., д. 518.

31 марта. Воронежская губ. Прошение крестьян помещика Гардениля дер. Гнездиковой Землянского у. Николаю I об ускорении рассмотрения их дела об освобождении от крепостной зависимости по духовному завещанию прежней владелицы.
ЦГИАЛ, ф. 1583, оп. 10, д. 43444, л. 1—2 об.

Не позднее 15 апреля. Костромская губ. Отказ крестьян деревень Камешки и Адюшкина Юрьевецкого у. признать своей владелицей помещицу фон Шульц; была введена военная команда.
ЦГАОР, ф. 109-И, 4 эксп., 1851 г., д. 200, л. 35—35 об., 151 об., 204; ЦГВИА, ф. 395, оп. 287, ч. ОА, 1851 г., д. 221, л. 47—48.

3 мая — позднее 23 августа 1852 г. Смоленская губ. Побег дворовых людей (7) помещика Пастухова из дер. Панютой Гжатского у. в Москву из-за жестокого обращения и разорения их владельцем; отказ крестьян выполнять распоряжения помещика.
№ 64; ЦГИАЛ, ф. 1405, оп. 50, д. 915, л. 70—72 об.; ЦГАОР, ф. 109-И, 4 эксп., 1851 г., д. 200, л. 81—85 об.; *Рябков*, стр. 86.

Ранее 28 мая. Ковенская губ. Отказ крестьян помещика Петкевича в имении «Александрово» в Новоалександровском у. отбывать барщину.
Неупокоев, стр. 144.

Май — сентябрь. Харьковская губ. Волнение крестьян (30), вызванное переселением их помещиком Струковым из Орловской губ. в сел. Покровское Старобельского у.; подавлено с помощью военной команды.
ЦГВИА, ф. 395, оп. 287, ч. ОА, 1851 г., д. 507, л. 86—86-а; 1851 г., д. 546, л. 9.

Весна. Владимирская губ. Отказ крестьян дер. Медниковой Ковровского у. признать своими владельцами помещиков Миллер из-за жестокого обращения с крестьянами; сопротивление местным властям при попытке заставить их подчиняться помещикам.
ЦГАОР, ф. 109-И, 4 эксп., 1851 г., д. 200, л. 59—59 об., 151.

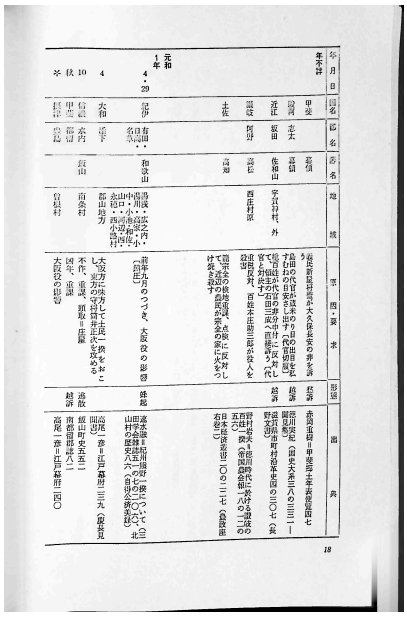
Не ранее 1 июня. Виленская губ. Сопротивление крестьян помещика Вольского в имении «Лепишнех» в Ошмянском у. вотчинным и полицейским властям; была введена военная команда.
ЦГАОР, ф. 109-И, 4 эксп., 1851 г., д. 200, л. 53—53 об.; *Неупокоев*, стр. 143.

— — Потрава государственными крестьянами дер. Великие Брмозовцы Лидского у. сенокоса помещиков А. Пясецкого и М. Щука и возврат крестьянами захваченного у них за потраву скота.
ЦГИАЛ, ф. 1345, оп. 54, кн. 62, № 1632.

— *Симбирская губ.* Отказ крестьян дер. Турдаки Ардатовского у. признать своим владельцем помещика Игумнова из-за жестокого обращения с ними.
ЦГАОР, ф. 109-И, 4 эксп., 1851 г., д. 200, л. 54—55, 154, 194—195 об.

Не позднее 6 июня. Тверская губ. Прошение крестьян помещиков Мельницких деревень Купецы, Конюшенной и Глинцы Бежецкого у. министру юстиции о содействии освобождению их из крепостной зависимости.
№ 67.

(b) Russia Tertiary Data



(c) Japan Tertiary Data

after the passing of that emperor. Versions exist for the Ming and Qing dynasties (1368-1912). Electronic text versions are available from Harvard's Chinese Text Project as well as from Chinese websites (Sturgeon, 2021). For an explanation of Veritable Records and an example of their use in the study of social conflict, see Hung (2011). Because of China's influence, there are also Veritable Records in Chinese for Korea's Joseon Dynasty (1392-1897, Kang and Suh (2023)), Japan (858-887), and Vietnam (1545-1909).

Veritable Records hold immense promise for scholars of social conflict. Already digitized, they do not require scanning copies of books. They cover periods longer than any European dynasty. This coverage enables research designs not possible using European event data. For example, records exist spanning effectively complete isolation to integration within an international system. The two biggest drawbacks to Veritable Records are their focus on royal court communication, likely biasing records, and use of Classical Chinese, a language on which current large language models do not perform well.

In China, local gazetteers (*difangzhi*) offer an alternative. Often semi-official, they are

written geographically closer to the information they contain. In addition to political history, they contain literature, religious, and geographic data. They extend back at least as far as the Tang Dynasty (618 CE). Many research libraries offer access to scanned gazetteer collections; for more detail about them, see Chen et al. (2023).

4 Making Data

Having identified sources, there are three steps to transform them into digital data.

The first step is to digitize the physical documents. Recent hardware developments have transformed the process from one requiring expensive, specialized hardware to create images and then slow software to apply optical character recognition (OCR) to the images into a much simpler process. The CZUR ET24Pro book scanner fits on a normal size desk, generates high-resolution scans; automatically removes the distortion from a book's spine, fingers captured in the scan, and tilt; uses ABBYY FineReader, the leading OCR software, to convert the images to documents; and can output scans as images, text files, .pdfs, or Microsoft Word and Excel files. Once past the initial learning curve, a user can digitize a page every 5-10 seconds. At the end of scanning, the books are searchable files.

Word, .txt, and .pdf scans of the books were tested. The .pdfs produced the best results because they preserve the original formatting of the books. Instead of submitting a .pdf, however, the `PyPDF2` library extracts the text, cleaning is performed to remove page numbers, headers, and footers, and then the text is submitted to a large language model. In the winter of 2025, Google's Gemini and OpenAI's ChatGPT 4o were tested, and ChatGPT 4o produced better results. Initially, each submission contained 10 pages of text, but that quantity led to hallucinations and timeouts. Submitting one page at a time significantly increased the accuracy and speed of ChatGPT 4o. Scans with a resolution of 400 dots per inch were used.

Aoki (1981) was scanned in the same manner as the other two sources. Submitting one page at a time produced very poor results, however, even when the page was rotated 90

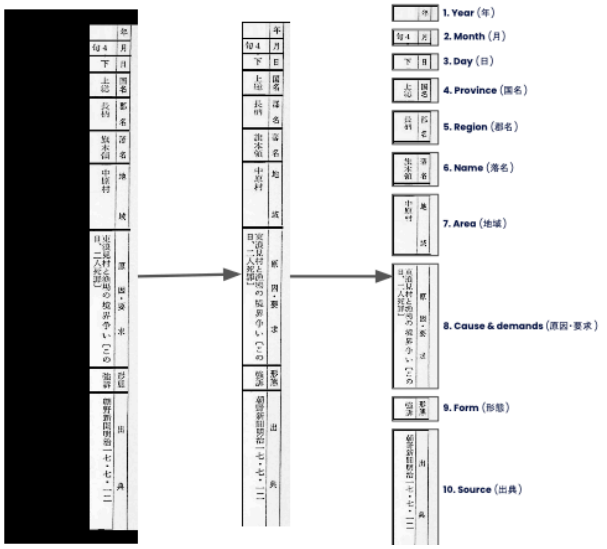
degrees so that a column of data becomes a row. The Japanese source therefore requires different, more intensive processing before submitting to a large language model. The following work was performed by Easie, a technology services company, in consultation with the author.

After receiving the one page scans from the CZUR ET24Pro, Python converts them to .png files and removes a slight tilt that the scanner sometimes did not remove. A human then manually crops each .png image so that each column (event) becomes its own file. Each of these image strips is then added to a slide template to create uniform dimensions and enhance the black separator between columns, both of which improve readability for the later steps. The slides, one per strip, are saved as a .pdf and converted back to .png. A script then splits each image strip into ten substrips, one per row.[4] Next, each image is enhanced to sharpen the image and remove artifacts such as blotches, dust, halos, and paper texture. Specifically, bilateral filtering reduces noise while preserving character boundaries, Lanczos interpolation increases the image resolution, adaptive histogram equalization optimizes contrast to maximize character visibility, and binarization converts the grayscale scans to black and white. These enhancements use deterministic mathematical formulas with known behaviors, in contrast to a multimodal artificial intelligence model that could introduce its own distortions. Figure 3 shows an image strip split into ten substrips and one substrip before and after enhancement.

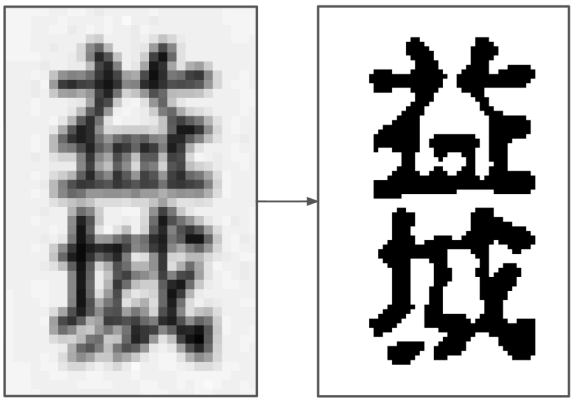
The second step is prompt engineering, recursively editing prompts until satisfactory results are returned. The Latin America and Russian texts are submitted to ChatGPT 4o *in their original language.* The ability of large language models to understand languages different from the researcher's, and of ABBYY FineReader to recognize 180 languages, means the language of a text is a much lower barrier to entry to acquiring data than ever. Some language knowledge is still necessary to identify sources and verify large language model interpretations of the source text, but a researcher does not need expertise, or even proficiency,

[4]The tables are vertically oriented, so a column is the event and each row of the column is a variable. The ten variables are year, month, day, province, region, name, area, cause/demand, form, and source.

Figure 3: Aoki (1981) Image Strip Split Into Substrips (left) and Before and After Enhancement (right)



(a) One image per event-field



(b) Event-field after deterministic enhancement

in the source language.

The Japanese data are handled differently at this stage as well. The enhanced substrips are first uploaded to EasieOps, Easie's cloud analytics platform that includes computer vision models. EasieOps extracts the Japanese text from each substrip and submits it to Google's Gemini 2.0-flash with a temperature of .1.

Section S1 shows the three prompts and discusses the iterative process of the prompt engineering in more detail. Section S2 provides the ChatGPT 4o settings used. Section S3 describes the Easie relationship and Japanese pipeline in more detail.

For Rama (1959) and the Russian scans, minimal post-processing is required. For Rama (1959), ChatGPT 4o did not return complete dates for about 5% of entries. An incorrect entry is one whose date column does not contain two '|' separators. In those cases, a missing separator and 'NA' value are added. For the Russia scans, post-processing was required to clean the names. Despite instructions to the contrary, many place names were kept in Russian. Place name spelling often varied by source, such as "Kiev", "Kyiv", " ", and other variants. The LLM also interchangeably used oblast, province, region, and Russian

spellings, so those were standardized. 1,082 unique place names were reduced to 74. All corrections were addressed with Python scripts.

Because Aoki (1981) was processed one field at a time, there were no location formatting errors.

All accuracy results are for the processed location fields.

In addition, a team of native and second-generation Japanese research assistants translated Aoki (1981) into an Excel spreadsheet. Manual creation was performed for two reasons. The author reads French and the research assistant working on the Russian data reads Russian, but neither read Japanese. A ground truth was therefore needed to evaluate the large language model's output. In addition, manual creation provides insight into when researchers should prefer to use artificial intelligence or humans to create data. The former has higher fixed but lower marginal costs while the latter has lower fixed but higher marginal costs.

5 Results

5.1 Accuracy

To determine the accuracy of the LLM output, a subset of events for each source is manually evaluated. For Latin America, the first 46 events are evaluated. For Japan, the first 100. For Russia, a random sample of 100. Researcher availability determined the size of each sample.

For each event, the date, description, and location of the LLM response is compared to the original text. Match values are "no", "yes", and "partial". The translation does not have to be perfect to receive a "Yes" label, but the translation's meaning has to match. "Yes" are translations which preserve the details of the event, such as actors' names, content of the event, and numbers (60 killed, 2 deaths, and so on). "Partial" matches are situations such as a date with the correct month, date, or year but not all three, a hierarchical location with at least one correct level, or a description whose translation captures some of the original meaning but loses some details. Japan does not have a date entry because Aoki (1981)

provides dates in the imperial date system and post-processing is used to convert those to Roman dates. Aoki (1981) provides a field called "Action" that is the type of event, such as "flee" or "petition". For all sources, the accuracy is for output from the LLM without subsequent corrections.

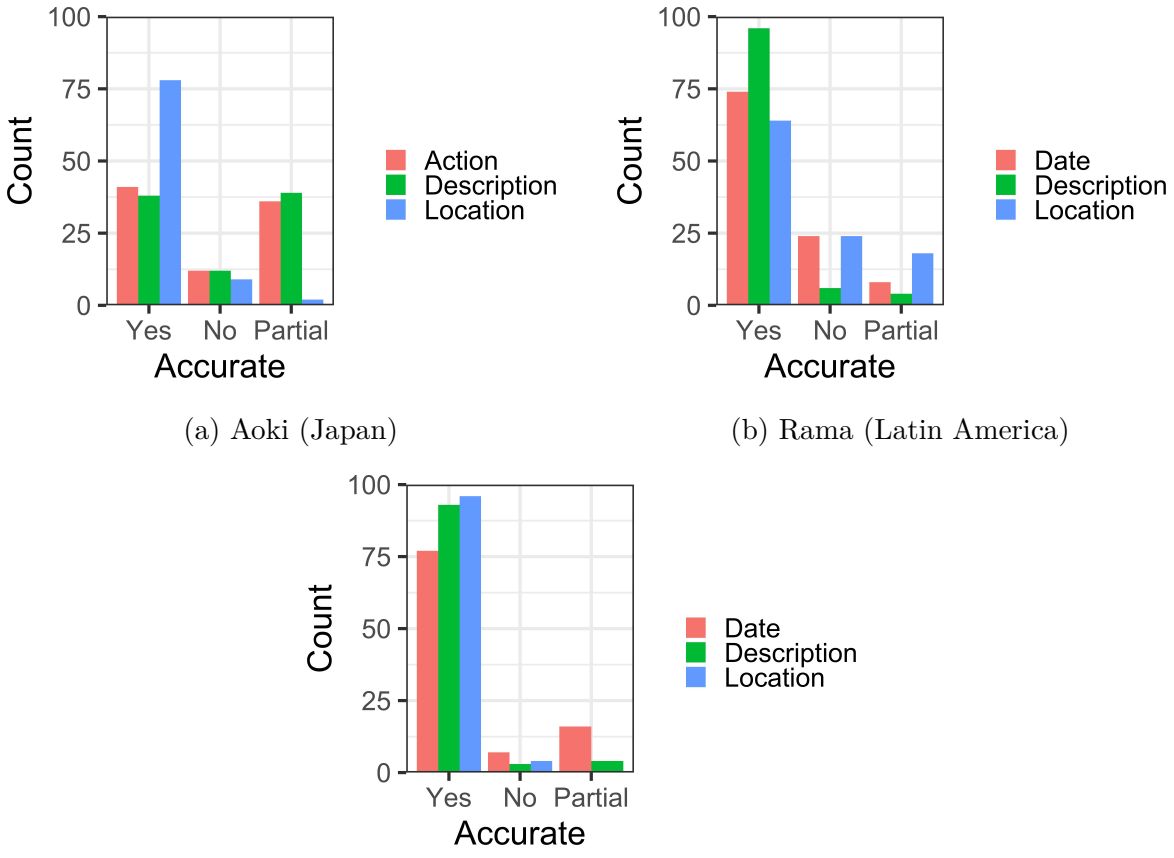
Each source is evaluated by one rater who reads that language. Aoki (1981) was evaluated by a native Japanese speaker. Rama (1959) was evaluated by someone who has studied in two Francophone countries and has reading proficiency. The Russian series was evaluated by an undergraduate research assistant who minors in Russian. Given these readers' experience with their respective language, it is unlikely the human validation is wrong when the artificial intelligence is correct.

The human validation is also unlikely to be incorrect because the LLM is largely translating. In traditional event data creation, a newspaper article is read and pieces of information are extracted. The three texts for this paper, however, are already event datasets. Their authors have combed primary sources and compiled them into structured event data. This paper then takes those structured event data and extracts the date, location, and translates the description. Human validation is required because the three texts are formatted differently and contain idiosyncrasies, so the human validation ensures the LLM knows where to look in the document for the date and location and then translates the description accurately. Since the human validators can easily identify the date and time and have language skills ranging from proficient to fluent, multiple validations were not performed.

Multiple validations per source are also not used because this paper's goal is to measure the accuracy of artificial intelligence coding. When the artificial intelligence translation is wrong, it is glaringly wrong. For example, it will hallucinate sentences from a previous entry, split one event into two, or literally translate an entry instead of using natural-sounding English. In these cases, the artificial intelligence is clearly wrong. Whether or not two validators agree on the exact translation is less interesting, for this project, than whether or not the LLM is wrong.

Figure 4 shows the performance for each source. Artificial intelligence is accurate for all sources, though to varying amounts. It is least accurate on the Japanese data. Despite the significant amount of time committed to producing high-fidelity scans and prompt engineering, the most common event description and action result is partial accuracy. Translations coded as partial capture a combination of actors and actions but miss enough detail, such as the initiating and targeted party or the order of events, that a researcher should prefer a human translation. Actor names rarely match the manual translation. Many of the Japan actions coded as partial are consistently mistranslated, such as "flight and dispersal; desertion" for "flee", so with minimal post-LLM processing, the action field would become reliably accurate.

Figure 4: Accuracy Results per Source



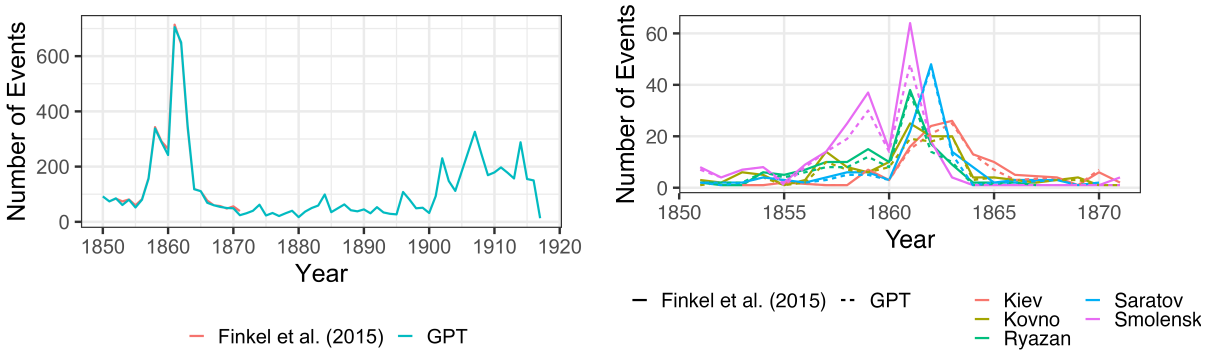
(a) Aoki (Japan)

(b) Rama (Latin America)

(c) Various (Russia)

The Latin America results are the second most accurate. The most common result for

Figure 5: Russia LLM Results Match Their Handcoded Equivalent



the date, description, and location is output that is correct without further processing.

The most accurate results are from Russia. The LLM records date, description, and location correctly for a minimum of 75% of entries (date) and almost 100% for location. Date inaccuracies arise from non-standard entries such as ranges, missing days or months in the original data, or entries that span multiple pages.

The Russia LLM is further validated using a dataset built from the same sources for the years 1851-1871 (Finkel, Gehlbach and Olsen, 2015). A team of Russian experts read and coded *Krestianskoe dvizhenie v Rossii v XIX-nachale XX veka*, so their data should be 100% accurate. Figure 5(a) shows that the LLM data match the national level time series. Figure 5(b) shows the result for the five most common provinces, and the LLM data closely match for each province, though not as well as at the national level.

As a robustness check, the same Aoki (1981) scans were also sent to Qwen3-VL Plus, a model released in September 2025 (Bai et al., 2025). Qwen is a family of artificial intelligence models from Alibaba. Qwen3 is the most recent iteration of this family, as of January 2026, and "VL" is the vision language version. Qwen3-VL Plus was chosen for three reasons. First, its weights are open source, meaning a researcher can use the model on their computer, avoiding API costs and, most importantly, ensuring consistency over time and across researchers. Second, Qwen3-VL Plus is the best performing open source model on Japanese text (Onohara, Baek and Aizawa, 2025). Third, since it is a vision model, it does not require

OCR pre-processing. Submitting scans to it is therefore much simpler than for ChatGPT 4o because the OCR steps are not necessary. The model is run only on the Japanese data, for two reasons. First, the accuracy rate is sufficiently high for the Latin America and Russia data. Second, the Japanese data require significant OCR pre-processing, so a vision language model that obviates those steps streamlines the data generating process. For the Latin America and Russia data, the CZUR machine performs OCR as it scans, so a vision language model would not quicken the data generating process.

Figure A1 shows the Japan results using Qwen3-VL Plus: it outperforms ChatGPT 4o. It records correctly 4 extra locations, 5 actions, and 21 descriptions. The increased performance for descriptions is especially important because those contain event details and are harder to translate than the action and location fields. The number of incorrect records also declines by 2, 7, and 11 for location, action, and description respectively.

Overall, these results show that LLMs can code structured event data with high levels of accuracy.

5.2 Cost

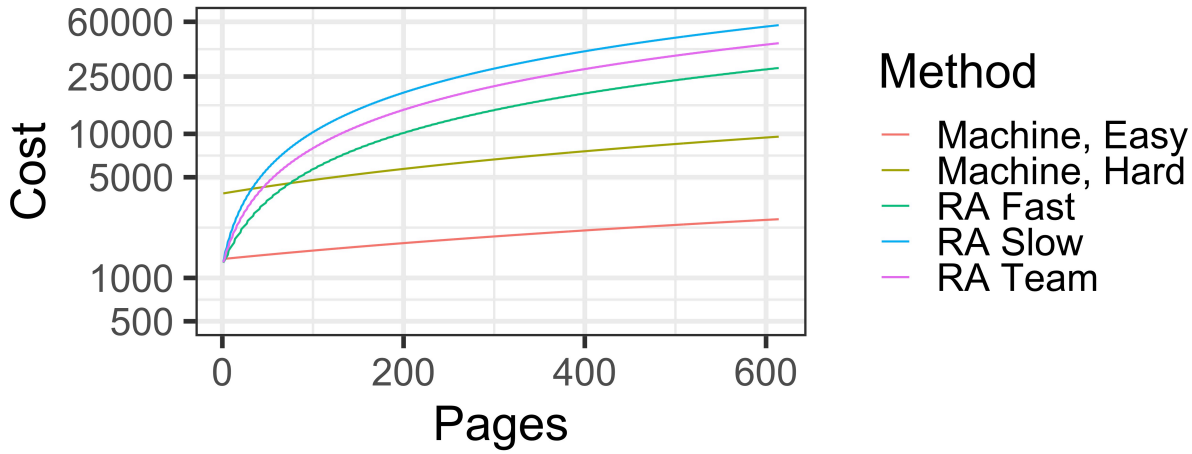
This accuracy does not mean, however, that they automatically dominate human approaches. Scanning documents and refining prompts takes more time than training human coders, and human coder output requires less review. Whether humans, LLMs, or a hybrid approach should be preferred will therefore depend on the cost and speed of humans.

Two sets of costs are calculated, one for human-only coding and the other for machine coding. Both sets include a fixed cost of the CZUR machine, training an RA to scan pages, and scanning pages. The human-coding includes a fixed cost for training research assistants. Three cost curves for human-coding are calculated. "RA Fast" refers to a team of research assistants who read the text quickly, "RA Slow" is the opposite, and "RA Team" means a mixture of the two speeds. Research assistant marginal cost is calculated at $26 an hour, requiring two hours of training, and coding one page every 1.2 (fast) or 2.7 (slow) hours.

The professor's wage for meeting times is calculated as $75 per hour, with a meeting held for every 8 hours of research assistant work. "Machine, Easy" refers to texts such as the Latin America and Russia ones that do not require the extra pre-processing employed for the Japan data. To the human fixed cost, a fixed cost representing determining proper scans, prompt engineering, and professor time is included. The marginal cost consists of writing code to submit to the LLM, evaluating results, iterating, and meeting time.[5] "Machine, Hard" is the cost curve for creating the Japan data. Its fixed costs include half of the engagement cost of hiring Easie, the vendor. Its marginal cost consists of Easie's time to create the image strips from the provided scans and the other half of the engagement cost. Figure 6 shows these five cost curves for generating social conflict data; the x-axis range is determined by the number of pages scanned for Aoki (1981).

Whether or not a large language model should be used depends on the difficulty of the book. For books like Rama (1959) and the *Krestianskoe dvizhenie v Rossii v XIX-nachale XX veka* series, using a large language model is quicker for any source longer than 2 pages, which is probably any source. For a difficult book, the large language model cost is less than the team after 45 pages. For a fast research assistant, 75 pages, and for a slow one 32.

Figure 6: Comparing Human and Machine Costs



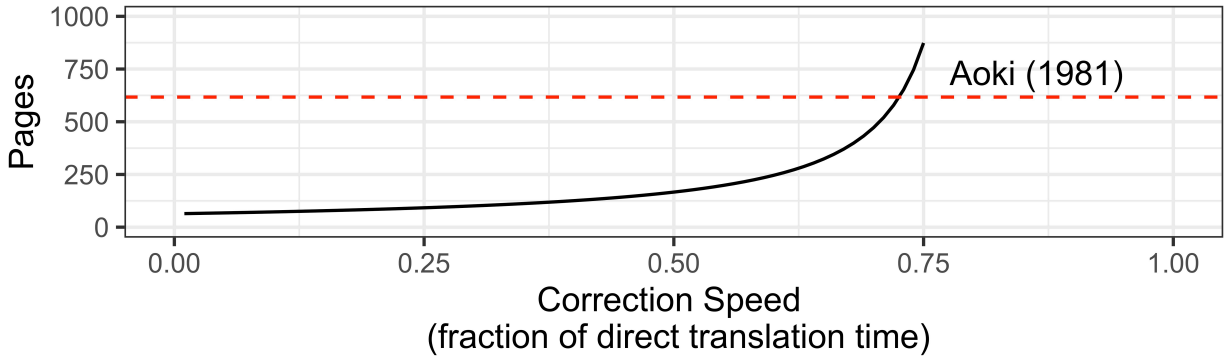
[5]These costs could be considered fixed costs. The true marginal costs are the API costs and time to submit code to the API, all of which are less than $.05 per page.

The intersection points for the "Machine, Hard" and research assistant curves can reasonably be considered to be further out. Given the large number of partial accuracy results for Japan, the "Machine, Hard" category represents a lower bound on the cost. The research assistant curves represent upper bounds since research efficiency improves with practice. For any source for which scanning and translation proves difficult, human review of output is extensive enough to call into question the LLM approach. Since the translations cannot be completely trusted, unlike in the "Machine, Easy" category, they each should be read and corrected when necessary.

When human correction is necessary to fix machine translation, whether or not to proceed with correction or only use human translation depends on how quickly a researcher can correct the machine translation. Figure 7 shows the page length at which it is quicker to use human translation as a function of the speed of correction relative to the speed of translation, using the costs from the "RA Team" and "Machine, Hard" calculations. For example, a value of .5 means that it takes half as long to correct a machine translation as it does to translate directly, so a book of more than 166 pages should use humans to translate directly. Figure 7 shows that if correction takes up to 72% as long as human translation, artificial intelligence is cheaper. The research assistant who translated Aoki (1981) requires 30% as much time to correct it as they do to translate it directly, so after 101 pages the artificial intelligence approach is faster.

Every row of the LLM output is manually reviewed, but not every one requires correction. Figure 7 assumes every LLM response is read, so the x-axis is an average of the entries that do and do not require correction. For Aoki (1981), every entry description requires some correction. For the Russian texts, 10%. For Rama (1959), 5%.

Figure 7: Minimum Source Length for LLM Translation to Be Cost-Effective as a Function of Correction Speed



6 Discussion

These results demonstrate that with appropriate hardware, pre-processing, and prompt engineering, social conflict event data stored in books can be digitized accurately. This paper has shown that social conflict data about post-contact Latin America, Imperial Russia, and Tokugawa Japan stored on the printed page are amenable to conversion into rectangular datasets. This ability pushes the horizon of social conflict studies back in time and widens it to geographies not commonly represented in current datasets.

Despite this advance, researchers should not immediately conclude that scanning and large language models should be used for all projects. The results suggest that languages that do not use Roman or Cyrillic characters require more researcher intervention. Nonetheless, the main lesson from the cost calculations is that large language models will almost always produce digital data from analog sources at a lower cost, in time and money, than human teams.

The cost calculations and results should be taken as guidance rather than final determinations. If research assistant or research time costs more than those used for this project, large language models are better for even shorter projects. The opposite is true if they are less expensive. If research assistants can code a book particularly quickly, they should be preferred. For hard source books, two considerations require project specific modeling.

The fixed cost becomes especially important. More importantly, if inspecting large language model results causes one to have research assistants correct the output, then the manual approach will beat the large language model approach for books hundreds of pages long.

These results represent the floor of what is achievable with artificial intelligence. Fixed costs will decrease. Hardware prices should decline. As researchers and assistants become more proficient with artificial intelligence, less training and prompt engineering are needed to start generating results, reducing cost. Accuracy should also increase. As scanning hardware improves, less processing before submitting scans to a large language model will be required. Artificial intelligence is also evolving much faster than the pace of academic research. Since the start of the work, large language models have advanced two generations, likely improving translation. More prosaically, the accuracy results do not include any post-processing of the large language model output. Even if hardware and software ossify, programmatically correcting mistakes in dates, locations, and event translations will improve accuracy.

The results for the Japanese data demonstrate the rapidly changing capabilities of artificial intelligence. Using Qwen3-VL Plus, an open-source model released 16 months after ChatGPT 4o, the model used for the first set of results, generates a large improvement in description accuracy. This improvement reduces the average correction time for artificial intelligence outputs, further reducing the situations in which direct translation outperforms artificial intelligence translation. Given the replicability and transparency advantages of Qwen3-VL Plus, its lower cost, and its improved accuracy, it, or at least other open source models, should be the default choice for researchers.

Developing this paper's techniques also generated several insights for future procedures. Large language models prefer as small a number of events per submitted scan as possible. For Aoki (1981), each submission was one event, but for Rama (1959) and the *Krestianskoe dvizhenie v Rossii v XIX-nachale XX veka* series, one page per submission was as accurate as one event. After determining the best submission size, iteration through prompts is necessary. A prompt should be treated more like a codebook than oral instructions, and more detail

is usually better. In addition, a prompt should focus on translation and not event detail extraction. The two tasks are distinct, and mixing them leads to worse translation and detail extraction than using separate prompts. Extracting details such as number of participants, names of leaders, types of grievances, e.g. taxes or forced labor, is best handled by subsequent work. Working with LLM memory and different formats or quantities of scanned text may also improve results (Timoneda and Vallejo Vera, 2026).

Researcher intervention is still necessary. Most importantly, a researcher should always review the LLM output. Even for texts where tests show high accuracy, human review should not be skipped. For those texts, the small number of entries needing correction mean minimal researcher time will be devoted to correcting the output. Researcher intervention is needed for edge cases. For example, when an event spans multiple scans, the LLM will struggle with the fragment at the start of a scan, which often results in hallucination. ChatGPT 4o would sometimes be unable to read a scan, a failure that would only become apparent when trying to merge outputs. If a subsequent rescan and LLM call does not solve the issue, then one has to manually code that scan. Researchers may also consider modifying the LLM's translations, which tend to be literal as opposed to idiomatically correct in English. Literal translation did not change meaning in any of the texts used in this paper, but it expresses the machine behind the magic.

The approach has only been tested on structured sources such as chronologies or tables. Future testing should explore the extent to which data is available from unstructured sources like travelogues, narrative histories, and monographs.

Future work should also consider why accuracy varies across source type. The three sources for this paper vary in geographic focus, language, structure, paper and ink quality, and artificial intelligence model used. With only three sources, the degrees of freedom do not exist to determine the cause of the difference in accuracy. Part of the explanation is likely that the training data for ChatGPT 4o and Gemini contain more Spanish text than Russian and more Russian than Japanese. If the language use of user queries reflects this inequality,

the accuracy difference will persist. One could interrogate this possibility by translating the three sources into English, having artificial intelligence code the English, and seeing if the accuracy difference persists. If it does, it suggests the difference is due to non-linguistic differences, such as the formatting of the source.

This paper joins a growing body of political science work that exploring how large language models can augment researchers' data generation process. LLMs can read recent texts from American media to extract content, while this paper shows they can do the same for old texts (Li and Jiang, 2026). In contrast, other work has found that LLMs are too optimistic when asked to evaluate how new data supports a hypothesis (Paci, 2026). It is also possible that the texts could be used to train agents in future simulations to recreate historical dynamics (Bojić et al., 2025, Wang et al., 2025)

The past offers a rich future for the study of social conflict. Events from centuries ago, outside of Europe, or both have not received scholarly attention in proportion to their frequency. Changes in the last three years ameliorate this situation. The advent of low-cost hardware and large language models means these data can be resuscitated relatively inexpensively, quickly and, most of the time, accurately. Given the extensively developed tools and methodologies for generating current social conflict data from newspapers (Weidmann, 2015, Wang et al., 2016), future research should focus on digitizing and generating social conflict data from tertiary sources using artificial intelligence.

References

Aidt, Toke, Gabriel Leon-Ablan and Max Satchell. 2021. "The Social Dynamics of Collective Action: Evidence from the Diffusion of the Swing Riots, 1830–1831." *The Journal of Politics* 84(1):000–000.

Anfimov, A M. 1998. *Krest ianskoe dvizhenie v Rossii v 1901-1904 gg. : sbornik dokumentov [The peasant movement in Russia, 1901-1904: Collection of documents].* Krest ianskoe dvizhenie v Rossii v XIX-nachale XX veka Moskva: Nauka.

Aoki, Koji. 1981. *Hyakusho Ikki Sogo Nempyo.* Tokyo: Sanichi.

Bai, Shuai, Yuxuan Cai, Ruizhe Chen and Ke ... Zhu. 2025. "Qwen3-VL Technical Report.".

Beissinger, Mark R. 2022. *The Revolutionary City: Urbanization and the Global Transformation of Rebellion.* Princeton University Press.

Bojić, Ljubiša, Velibor Ilić, Veljko Prodanović and Vuk Vuković. 2025. "An Agent-Based Simulation of Politicized Topics Using Large Language Models: Algorithmic Personalization and Polarization on Social Media." *Chinese Political Science Review* pp. 1–31.

Boschee, Elizabeth, Jennifer Lautenschlager, Sean O'Brien, Steve Shellman, James Starz and Michael Ward. 2015. "ICEWS Coded Event Data.".
URL: *http://dx.doi.org/10.7910/DVN/28075*

Chambru, Cédric and Paul Maneuvrier-Hervieu. 2024. "Introducing HiSCoD: A New Gateway for the Study of Historical Social Conflict." *American Political Science Review* 118(2):1084–1091.

Chen, Shih-Pei, Calvin Yeh, Sean Wang and Qun Che. 2023. "Treating a genre as a database: a digital research methodology for studying Chinese local gazetteers." *International Journal of Digital Humanities* 4(1-3):171–193.
URL: *https://doi.org/10.1007/s42803-022-00048-5*

Davenport, Christian and Patrick Ball. 2002. "Views to a Kill: Exploring the Implications of Source Selection in the Case of Guatemalan State Terror, 1977-1995." *Journal of Conflict Resolution* 46(3):427–450.

Demarest, Leila and Arnim Langer. 2022. "How Events Enter (or Not) Data Sets: The Pitfalls and Guidelines of Using Newspapers in the Study of Conflict." *Sociological Methods and Research* 51(2):632–666.

Finkel, Evgeny, Scott Gehlbach and Tricia D. Olsen. 2015. "Does Reform Prevent Rebellion? Evidence From Russia's Emancipation of the Serfs." *Comparative Political Studies* 48(8):984–1019.

Göbel, Christian and H. Christoph Steinhardt. 2022. "Protest Event Analysis Meets Autocracy: Comparing the Coverage of Chinese Protests on Social Media, Dissident Websites, and in the News." *Mobilization: An International Quarterly* 27(3):277–295.

Gould, Roger V. 1991. "Multiple Networks and Mobilization in the Paris Commune, 1871." *American Sociological Review* 56(6):716–729.

Hellmeier, Sebastian, Nils B. Weidmann and Espen Geelmuyden Rød. 2018. "In The Spotlight:Analyzing Sequential Attention Effects in Protest Reporting." *Political Communication* 35(4):587–611.

Herkenrath, M. and A. Knoll. 2011. "Protest events in international press coverage: An empirical critique of cross-national conflict databases." *International Journal of Comparative Sociology* 52(3):163–180.

Holland, Michael. 2005. *Swing unmasked: the agricultural riots of 1830-1832 and their wider implications*. Family and Community Historical Research Society.

Hung, Ho-fung. 2011. *Protest with Chinese Characteristics: Demonstrations, Riots, and Petitions in the Mid-Qing Dynasty*. New York City: Columbia University Press.

Ivanov L. M., Guzunov M A. 1964. *Krest ianskoe dvizhenie v Rossii v 1861-1869 gg.; sbornik dokumentov [The peasant movement in Russia, 1861-1869: Collection of Documents].* Krest ianskoe dvizhenie v Rossii v XIX-nachale XX veka Moskva: Mysl .

Kang, Hyeok Hweon and Michelle Suh. 2023. "Korean Chronicles Under a Macroscope: Towards a Digital Infrastructure in Premodern Korean Studies." *Korean Studies* 47:8–33.

Li, Rende and Ying Jiang. 2026. "LLM-Based Viewpoint Mining in the "Blame Game": How US Media Frame China's Debt Debate." *Chinese Political Science Review* pp. 1–23.

McCammon, Holly J., Karen E. Campbell, Ellen M Granberg and Christine Mowery. 2001. "How Movements Win: Gendered Opportunity Structures and U.S. Women's Suffrage." *American Sociological Review* 66(1):49–70.

Miller, Charles and K. Shuvo Bakar. 2023. "Conflict Events Worldwide Since 1468BC : Introducing the Historical Conflict Event Dataset." *Journal of Conflict Resolution* 67(2-3):522–554.

Myers, Daniel J. and Beth Schaefer Caniglia. 2004. "All the Rioting That's Fit to Print: Selection Effects in National Newspaper Coverage of Civil Disorders, 1968-1969." *American Sociological Review* 69(4):519–543.

Nam, Taehyun. 2006. "What You Use Matters: Coding Protest Data." *PS: Political Science & Politics* 39(2):281–287.

Nifontov A. S., Zlatoustovskiĭ B V Syromiatnikova Mariia Alekseevna. 1960. *Krest ianskoe dvizhenie v Rossii v 1881-1889 gg. : sbornik dokumentov [The peasant movement in Russia, 1881-1889: Collection of Documents].* Krest ianskoe dvizhenie v Rossii v XIX–nachale XX veka Moskva: Izd-vo sotsial no-ėkon. lit-ry.

Okun S. B., Paina E S Sorokina N F. 1962. *Krest ianskoe dvizhenie v Rossii v 1850-1856 gg. : sbornik dokumentov [The peasant movement in Russia, 1850-1856: Collection of*

Documents]. Krest ianskoe dvizhenie v Rossii v XIX-nachale XX veka Moskva: Izdatel stvo sotsial no-ėkonomicheskoĭ literatury.

Okun S. B., Sivkov K V Kudriavtseva Z I. 1963. *Krest ianskoe dvizhenie v Rossii v 1857-mae 1861 gg. : sbornik dokumentov [The peasant movement in Russia, 1857-May 1861: Collection of documents]*. Krest ianskoe dvizhenie v Rossii v XIX–nachale XX veka Moskva: Izd-vo sotsial no-ėkon. lit-ry.

Onohara, Shota, Jeonghun Baek and Kiyoharu Aizawa. 2025. "JMMMU-Pro: Image-based Japanese Multi-discipline Multimodal Understanding Benchmark via Vibe Benchmark Construction.".

Paci, Simone. 2026. "Is Research Safe in the AI Revolution? LLMs Fall Short of Expert Benchmarks in Scientific Evidence Evaluation." *Chinese Political Science Review* .

Predtechenskiĭ, A V and Z I Kudriavtseva. 1961. *Krest ianskoe dvizhenie v Rossii v 1826-1849 gg. : sbornik dokumentov [The peasant movement in Russia, 1826-1849: Collection of Documents]*. Krest ianskoe dvizhenie v Rossii v XIX-nachale XX veka Moskva: Izd-vo sotsial no-ėkonomicheskoĭ lit-ry.

Raleigh, Clionadh, Andrew Linke, Havard Hegre and Joakim Karlsen. 2010. "Introducing ACLED: An Armed Conflict Location and Event Dataset: Special Data Feature." *Journal of Peace Research* 47(5):651–660.

Rama, Carlos M. 1959. "Mouvements Ouvriers et Socialistes (Chronologie et Bibliographie): L'Amérique Latine (1492-1936).".

Salehyan, Idean, Cullen Hendrix, Jesse Hammer, Christina Case, Christopher Linebarger, Emily Stull and Jennifer Williams. 2012. "Social Conflict in Africa: A New Database." *International Interactions* 38(4):503–511.

Shapkarin, A V. 1966. *Krest ianskoe dvizhenie v Rossii, iiun 1907 g.-iiul 1914 g. : sbornik dokumentov [The peasant movement in Russia, June 1907- July 1914: Collection of documents].* Krest ianskoe dvizhenie v Rossii v XIX-nachale XX veka Moskva: Izdatel stvo "Nauka".

Shapkarin A. V., Zlatoustovskiĭ B V Pishvanova N I. 1959. *Krest ianskoe dvizhenie v Rossii v 1890-1900 gg : sbornik dokumentov [The peasant movement in Russia, 1890-1900: Collection of documents].* Krest ianskoe dvizhenie v Rossii v XIX-nachale XX veka Moskva: Izdatel stvo sotsial no-ėkonomicheskoĭ literatury.

Snyder, David and William R Kelly. 1977. "Conflict Intensity, Media Sensitivity, and the Validity of Newspaper Data." *American Sociological Review* 42(1):105–123.

Steele, Abbey, Christopher Paik and Seiki Tanaka. 2017. "Constraining the samurai: Rebellion and taxation in early modern Japan." *International Studies Quarterly* 61(2):352–370.

Steinert-Threlkeld, Zachary and Jungseock Joo. 2022. MMCHIVED: Multimodal Chile and Venezuela Protest Event Data. In *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 16 pp. 1332–1341.

Sturgeon, Donald. 2021. "Chinese Text Project: A dynamic digital library of premodern Chinese." *Digital Scholarship in the Humanities* 36(August 2019):I101–I112.

Sullivan, Christopher M. 2016. "Political Repression and the Destruction of Dissident Organizations." *World Politics* 68(4):645–676.

Timoneda, Joan C and Sebastián Vallejo Vera. 2026. "Rolling Memory: A New Approach to Annotation with Generative LLMs in Social and Political Research." *Chinese Political Science Review* pp. 1–15.

Turchin, Peter, Thomas E Currie, Edward A.L. Turner and Sergey Gavrilets. 2013. "War, space, and the evolution of Old World complex societies." *Proceedings of the National*

Academy of Sciences 110(41):16384–16389.

URL: *http://www.pnas.org/content/110/41/16384.short*

Valk S. N., Kudriavtseva Z I. 1961. *Krest ianskoe dvizhenie v Rossii v 1796-1825 gg. : sbornik dokumentov [The peasant movement in Russia, 1796-1825: Collection of documents]*. Krest ianskoe dvizhenie v Rossii v XIX–nachale XX veka Moskva: Izd-vo sotsal no-ėkon. lit-ry.

Wang, Wei, Ryan Kennedy, David Lazer and Naren Ramakrishnan. 2016. "Growing pains for global monitoring of societal events: Automated event coding raises promise and concerns." *Science* 353(6307):1502–1504.

Wang, Zhenyu, Dequan Wang, Yi Xu, Lingfeng Zhou and Yiqi Zhou. 2025. "Intelligent computing social modeling and methodological innovations in political science in the era of large language models." *Journal of Chinese Political Science* pp. 1–36.

Weidmann, N. B. 2015. "On the Accuracy of Media-based Conflict Event Data." *Journal of Conflict Resolution* 59(6):1129–1149.

Weidmann, Nils B. and Espen Geelmuyden Rod. 2018. Coding Protest Events in Autocracies. In *The Internet and Political Protest in Autocracies*. Oxford University Press chapter Chapter 4.

White, James W. 1995. *Ikki: Social Conflict and Political Protest in Early Modern Japan*. Cornell University Press.

Zaionchkovskiĭ, P and E Paina. 1968. *Krest'ianskoe Dvizhenie v Rossii v. 1870- 1880 gg.: Sbornik Dokumentov [The peasant movement in Russia, 1870-1880: Collection of documents]*. Krest ianskoe dvizhenie v Rossii v XIX-nachale XX veka Moskva: Nauka.

Zhang, Han and Jennifer Pan. 2019. "CASM: A Deep-Learning Approach for Identify-

ing Collective Action Events with Text and Image Data from Social Media." *Sociological Methodology* 49(1):1–57.

Supplementary Materials for

Digitizing and Generating Social Conflict Data with Artificial Intelligence

S1 Prompts

This section explains the iterative nature of the prompt engineering and shows the prompts submitted with each scan.

S1.1 Prompt Engineering

The prompts displayed in the next subsection were created after several rounds of iteration. Broadly, I would start with a simple prompt, test it on three scanned pages, and evaluate the returned data. Across all three prompts, the common lessons are to be as specific as possible and provide examples in the prompt. It is also useful to give the model an identity, e.g. "careful translator" or "helpful assistant".

Latin America The first prompt for Rama (1959) was,

```
Code this chronology and return comma separated columns for each event.

Here are the columns: Date (Year|Month|Day), Country, Region, City,
   Town, Actor, Cause, Actor Size, Target. If any information is
   missing, use "NA". If there is a year missing for the event, assume
   that it is from the same year as the event before.

Do not miss any events. Give me all of the output in one textbox.
```

The returned text had several problems. Not every row contained nine columns. Either columns were missing or extra ones were created because a column would contain commas. It struggled with dates because not every entry provided a year, month, or day. Subsequent prompts telling the model to be a helpful assistant still had issues. The final prompt became three times as long because it is much more specific. It explains what each column is, how to handle non-standard data in each column, and provides examples. It also stipulates not to miss any events and to use semicolon separation. The final Latin America prompt is,

```
You are a careful translator whose main goal is to be accurate.
   You are translating a book that is a timeline of events.
   Complete the following tasks.
```

```
First, translate this chronology from French to English.  Then,
   return each translated event as a row of 9 columns.  Use a
   semicolon separator.

The first column is year, the second is day, and the third is month
   . Year should be returned with 4 digits, in the YYYY format.
   Month and day, two digits; if the actual digit is one, add a
   leading 0.  For example, the date "1493 - 3 mai" is 1493  03
   05.  If there is a year missing for the event, assume that it is
    from the same year as the event before.

The fourth column is the translated text.

Add additional columns for Event (what the entry is about), Actor (
   who initiates the event), Cause (cause of the event), Actor size
    (number of participants), and Target (who the event targets).

Here is an example of three events:

1493 - 4 mai - Bulle d ' Alexandre VI, qui complète la précédente et
    trace une ligne de démarcation entre les possessions espagnoles
    et portugaises à << cent lieues à l ' ouest de n ' importe
   laquelle des îles Açores ou du Cap-Vert >>.

Septembre - Les indigènes de l ' ile de Saint-Domingue (la Espanola)
    massacrent la garnison laissée par Christophe Colomb au fort de
    Navidad. La répression est sanglante et 500 indigènes sont
   embarqués pour être vendus comme esclaves en Espagne.

- Une épidémie de fièvre jaune décime les colons espagnols de l '
   Isabelle et se propage chez les indigènes.

The first event occurs on 1493;04;05.  The second event occurs on
   1493;NA;09.  The third event occurs on 1493;NA;NA.  When no
   dates are given, assign the previous year and "NA" for the day
   and month.

If you do not know any information for a column, write "NA".

Return all events in the order you read them. Do not miss any
   events.

Separate the columns with a semicolon.
```

Russia The first prompt for Russia was,

```
First, translate this chronology from Russian to English. Then, return
   each translated event as a row. From each event, pull and return
   three comma-separated columns: Location, Date (YYYY/MM/DD), Source.
   For context, there is a bibliography under each event.
```

Initial prompt versions instructed the model to translate the Russian chronologies and extract locations, dates, and sources. While translation quality was generally high from the beginning, early attempts revealed a few systematic problems. First, the model frequently misinterpreted complex or nonstandard date expressions common in the source material, e.g., seasonal references, phrases such as "no later than," or date ranges spanning calendar years. For example, the prompt had said, "If the day is not provided, use this day rule: (Unknown: 00). For example, if an event took place in January of year (the day is unknown), the format should be: (year/01/00). If the month is not provided, use this month rule: (Unknown: 17). For example, if an event took place in year (no other date information is listed), the format should be: (year/17/00)." This complexity caused hallucinations, so the final prompt states, "If any part of the date is not available, use NA." Second, it struggled with indented or hyphenated entries that implicitly shared a date with a preceding event. In these cases, the model often coded indented events as having missing dates. Explicit, rule-based instructions to solve such problems. The model performs more reliably when given explicit time context (specifying the base year for each batch of chronologies) rather than inferring year transitions mid-page. Bibliographic sources are kept in Russian because the model encountered difficulty translating large blocks of text. The final prompt reflects lessons learned from previous failures: specificity outperformed generality, examples were useful to guide the model, and rules were simplified to avoid model hallucination.

The final prompt is,

```
You are a helpful assistant. First, translate this chronology from
   Russian to English. Then, return each translated event as a row.
   From each event, pull and return four comma-separated columns:
   Location, Date (YYYY/MM/DD), Source, Event Description. For context,
    there is a bibliography under each event. Give me all of the output
    in one textbox.
```

Use these rules when coding the chronology:

The Source (bibliography) column should remain in Russian; all other columns should only be in English. The location should be only in English.

Only provide one location in the Location column. Any locations other than the province should go in the Event Description column.

If there is a date range, provide a start and end date.

This chronology only contains entries that start in the year {year}, but may have end dates outside of {year}.

If the event only refers to a season, code the event according to this month rule: (Winter: 13, Spring: 14, Summer: 15, Fall: 16). For example, if an event takes place during the Spring of {year}, the format should be: ({year}/14/00).

If any part of the date is not available, use NA.

Use semicolons to separate columns. Do not use semicolons for any other purpose.

Coded events should be in the same order in the textbox as given in the chronology.

Phrases such as " no later than " or " no earlier than " should be returned as the date given with the phrase. For example, " no later than April 15th " should be coded as ({year}/04/15).

A hyphen or em dash between two dates should return a date range. For example, {year}; May 3rd — no later than August 23rd {int(year)+1} should return: ({year}/05/03 - {int(year)+1}/08/23).

If there is a hyphen or em dash with an indentation before an entry with no date, assume that the entry has the same date as the previous event. For example, within the following structure, entries 2 and 3 should be output with the date August 20th, {year}:

August 20 {year}, Entry 1
- Entry 2
- Entry 3

You must follow each rule, do not ignore any rules.

Japan The prompt engineering process involved iterative refinement of two distinct prompts: one for extraction and one for translation. Both used structured JSON outputs with embedded examples. The final prompts reflect lessons learned from systematic failures. For extraction, specificity outperformed generality, where detailed instructions about headers, reading order, and multi-column detection reduced confusion. For translation, simplicity reduced hallucinations. In both cases, concrete examples and structured outputs prevented formatting inconsistencies.

For extraction, the initial prompt instructed the model to extract table data from images with separate strips of data in Japanese, specifying field names with a JSON schema example. Testing revealed consistent failures. The model frequently extracted column headers (年, 月, 日, 国名, 郡名, 落名, 地域, 原因·要求, 形☒, 出典) as data content and struggled with reading order in multi-column cells. The final extraction prompt became substantially longer and more prescriptive. It includes explicit header detection instructions, listing all column headers to avoid extracting and describing the two-box structure of each cell (header box on the right, content box on left). It also specifies detailed reading order instructions: top to bottom, right to left by column, starting with the rightmost text. The prompt also includes system-level instructions preventing refusals for lower-quality images and a multi-step verification checklist. It uses a temperature of 0.0 for deterministic output.

For translation, the initial prompt requested accurate, natural translations with a JSON output example. A second iteration added field-specific context (distinguishing year values from protest descriptions) and historical framing about civil unrest documentation. However, this change led to increased hallucinations, with the model over-interpreting context and inventing details. Easie switched to a simpler prompt, removing all contextual complexity and adding only a concrete Kanji number example (一二三四五六七八 for 12345678) and an instruction to provide exactly one translation per field. Temperature of 0.5 provided the highest translation accuracy.

The final prompt is,

```
Translate the following Japanese text to English. \{context\}. This
   text is from historical Japanese protest documentation written in
   Kanji, Hiragana, and Katakana scripts from historical records of
   civil unrest and social movements.

Japanese text: \{japanese_text\}

Provide an accurate, natural English translation that preserves the
   historical and cultural context. For historical terms, geographical
   names, and period-specific references, prioritize accuracy over
   modern equivalents. If numbers appear in Kanji format, convert them
   to Arabic numerals in the translation.
If you see special symbols like approximately symbols (~,  ) or
   asterisks (*), preserve and return those symbols exactly as they
   appear in your translation.
Return only the English translation without additional explanation or
   commentary.
```

S2 LLM Settings

For Latin America and Russia, the following settings were used. Temperature was set at 1. Top-p is also 1. The frequency penalty and presence penalty are 0. There is no upper bound for response tokens. The work was performed on model gpt-4o-2024-08-06 with a seed value of 07042025. All interaction was handled through version 1 of the chat completion API endpoint.

S3 Japan Pipeline Configuration

The Aoki (1981) data present the greatest difficulty for extraction. After spending two months iterating with a research assistant and making poor progress, professional help was sought. The author had worked with Easie on several data extraction projects since 2020 and was happy with their work, so they were the only vendor approached. Scans of the book and a summary of the work attempted to that point were provided. Three meetings were held to clarify scope. The services detailed below were purchased for a flat fee of $10,000.

The work for the paper's initial submission took two months, while that for the revision took one and a half.

After receiving the per page scans from the author, Easie creates image strips. An image strip is a section of a scanned page that corresponds to one event, a column in the Aoki tables. This process is described in Section 4. Then, each strip is split into 10 substrips, one per row (variable) of the Aoki table. Enhancement occurs with denoising, upscaling, increasing contrast, and binarization. Denoising smooths negative space, the space away from borders and text, using a bilateral filter of diameter 5 pixels, a color sigma of 80, and space sigma of 80. Upscaling doubles the image size by using the Lanczos-4 interpolation algorithm to add pixels. Contrast is increased using contrast limited adaptive histogram equalization (CLAHE) using a clip limit of 1.2 and tile grid of eight by eight. Binarization converts the image to black and white using a dual-threshold process with an adaptive block of 11x11 pixels, adaptive contrast of 2, global threshold of 200 (of 255) and a cleanup kernel that is a two by two ellipse. This work was completed using Python's OpenCV, Pillow, and NumPy libraries.

The EasieOps portal is only available to EasieOps clients. Once the dataset is completed, each image strip (event scan) will be released.

S4 Robustness Check

Figure A1: Japan Robustness Check with Qwen3-VL Plus