

Can a Language Model Read Russian for You? Auditing a Leader-Visit Dataset with an LLM Research Assistant

20 October 2026

Ali Balci, Sakarya University

alibalci@gmail.com

Abstract: Researchers who cannot read the languages of the countries they study must rely on international media or hire native-speaking collectors. Scholars in the Global South often can afford neither, so data collection itself becomes a source of methodological inequality. This article reports an experiment in closing that gap. Using a large language model as an agentic research assistant, I audited a dataset of Russian leaders' foreign visits (1991–2023) that my co-author and I had built from English-language sources without reading Russian. The model read every issue of the daily *Kommersant* from 1992 to 2023, checked the dataset against the Kremlin's official trip archive, and logged each proposed change with a verifiable source, while all coding decisions remained mine. The audit recovered 134 missing visits and removed or corrected 76 records. The gaps were concentrated in the 1990s and among prime ministers, where international coverage was thinnest. Re-estimating the models of the original article with the revised data left its main finding intact: domestic instability raises the probability of a Russian official visit to autocracies but not to democracies. I draw lessons about where English-language media suffice, where they do not, and how LLMs can lower the cost of multilingual data collection for under-resourced scholars.

Researchers who study countries whose languages they do not read face a basic problem in collecting data. They can rely on international media, or they can hire collectors who know the

language. The first route requires institutional subscriptions to news databases such as Factiva or LexisNexis. The second requires funding. Scholars in the Global South often have access to neither (Tickner 2013; Medie and Kang 2018; Öztğ 2022). Libraries, particularly at peripheral universities, rarely subscribe to the major news databases (UNESCO and ISSC 2010). Research grants are scarce, and when they are secured, they pay research assistants very little (Steinert et al. 2021). Data collection itself thus becomes a source of methodological inequality in global academia (Pasha and Murphy 2002). My co-author and I ran into these barriers when we built a dataset of Russian leaders' foreign visits (Balci and Aras 2025). My co-author solved the staffing problem by recruiting graduate students to collect the data. A second problem remained: none of us read Russian. We assumed that high-level visits are among the most visible acts of a state, so international English-language media would capture them. We ran several reliability checks, but the data were still incomplete.

Large language models now give us a way to test that assumption. I used Claude Opus 5.5, a large language model from Anthropic, as research assistants that work directly in Russian. The model did not answer from memory. It worked as an agent, running a web browser on my computer and writing its own code to collect, compare and record evidence. It went through the Kommersant archive issue by issue, from the weekly editions of 1991-1992 to the daily paper through 2023. That covered more than 8,700 issues and the headlines of roughly 400,000 articles. It flagged every report of a Russian president, prime minister or foreign minister traveling abroad and checked each report against our dataset by leader and date. Where no matching record existed, it read the full Russian text to establish where the leader went, when, and why. It then checked the dataset against official Russian sources, including the Kremlin's trip archive for 2000–2023 and the archive of Putin's time as prime minister for 2008–2012. Every change it proposed was logged with a link to the source, so I could check it, and every coding decision remained mine: whether to count trips that were cancelled or cut short, trips on the sidelines of summits, or visits outside a leader's term. In effect, the model removed the language barrier and most of the labor cost.

The rest of this article proceeds in four steps. First, I describe how the LLM-assisted audit worked: how the model searched Russian-language newspapers and official archives, how it matched what it found against our dataset, and how I decided which records to add, correct or remove. Second, I report what the audit produced, namely the missing visits it recovered, the errors it corrected, and where in time and across leaders those gaps cluster. Third, I re-estimate the models from our original article with the revised data. This tests our core assumption that international media capture high-level visits well enough for this research design. Finally, I discuss what this exercise implies for scholars who work across language barriers with limited resources, and what LLM-assisted data collection can and cannot do for them.

Auditing the Data with an LLM Research Assistant

The audit had three stages, each built on the one before. In the first, the model reviewed the dataset period by period (1991–97, 1998–2004, 2005–13 and 2014–23). It flagged records with impossible dates, visits falling outside a leader’s term, duplicate entries and inconsistent coding. In the second, it searched *Kommersant*, a leading Russian business daily whose full archive is online. It covered 62 weekly issues from July 1991 to September 1992 and then the 8,760 daily issues from October 1992 to December 2023. The search ran in passes organized by foreign minister: Kozyrev, Primakov, Ivanov and Lavrov, each together with the prime ministers who served in his period. For each issue, the model collected every headline that named a sitting leader alongside words for travel, talks or summits. It compared each one with the dataset, allowing a few days’ difference between the report and the recorded visit. Where no record matched, it read the full Russian article to establish who traveled, where, when and for what purpose. In the third stage, it checked the dataset against official lists: the Kremlin’s archive of 455 presidential trips abroad from 2000 to 2023, and the archive of Vladimir Putin’s trips as prime minister from 2008 to 2012.

The model proposed; I decided. Every proposed change entered a review report with its category, its reasoning and a link to the source, and I approved or rejected it. The categories were addition, deletion, correction, flag or unconfirmed candidate. Ambiguous cases forced coding rules that

the original project had left implicit, so I set them explicitly: Visits before a leader took office or after he left it are excluded, trips that were cancelled or cut short are excluded, trips on which a leader accompanied another official are counted, private trips are counted when found but were not searched for. The most consequential rule followed directly from our theory in the original article. Since an official visit signals support only when it is aimed at the host, a meeting with the host leader on the sidelines of a multilateral summit is not counted as an official visit. Neither is a meeting with a third country's leader that merely happens to be held in the host country, such as the Putin–Biden summit in Geneva. Rows removed under these rules were not deleted. They were moved to a separate sheet with the reason for each, so other researchers can restore them under different assumptions.

Two features made the process verifiable rather than a black box. First, the model did not work from memory. It wrote code to fetch and parse each source, and every claim in the final dataset traces back to a specific newspaper article or official web page. Second, this transparency made errors visible, including the model's own. Partway through the audit, it found that its scanner read only the first of the archive's batches of articles, roughly half of each issue. It reported the defect, corrected the scanner and re-read all 8,760 issues in full, adding about 200,000 articles to the search. Access, not language, set the remaining limits. The Russian government's website (government.ru) blocked automated access, so I could not check Dmitry Medvedev's trips as prime minister (2012–20) against an official list. The foreign ministry's site (mid.ru) also put up a CAPTCHA, an automated "are you a robot" check, when the requests came too fast.

What the Audit Found

The original dataset rested on two kinds of sources. For the period from 2000 onward, it drew mainly on official Russian websites: the Kremlin's record of presidential trips, the foreign ministry's pages for foreign-minister visits, and the government's site and the archived site of the prime minister for prime-ministerial trips. These official sources account for about 70 percent of the 1,915 records. For the 1990s, when no such websites existed, the data came from English-language news archives: LexisNexis searches, targeted Google searches and, for cross-checking,

outlets such as Radio Free Europe/Radio Liberty, UPI, CNN and the Washington Post. Each record carries a link or a headline, and a secondary column records a second source where one was found, from partner-country foreign ministries to photo archives that confirm a leader's presence. Collection took place between October 2023 and February 2024.

The audit, conducted in October 2026, grew the dataset from 1,915 to 2,010 visits. It recovered 134 visits that were missing, removed 39 records that did not meet the coding rules, and corrected dates or purposes in 37 more. The removed records include visits outside a leader's term, aborted trips, sideline meetings and summits held in a third country. Each record was checked against a Russian-language or official source. The gaps were not random. They clustered by period and by office, in a pattern that tells us where English-language media coverage of Russian diplomacy was thin. Most missing visits date from the 1990s and early 2000s. Viktor Chernomyrdin's trips as prime minister in 1993–95 were the largest single gap. By office, gaps were largest for prime ministers and smallest for foreign ministers in recent years. Sergey Lavrov's record was almost complete: the audit added only three visits to his 718. Before 2004, by contrast, prime ministers' trips to regional councils and to neighboring capitals were regularly missed.

Change	Rows	Main effect on the outcome
Missing visits added	134	Mostly 1992–2006: Chernomyrdin, Kozyrev, Putin's early presidency
Moved to Deleted Rows	39	Visits outside a leader's term, aborted trips, sideline meetings, third-country summits
Purpose or date corrected	37	15 Official rows recoded as Multilateral; dates fixed

The recoding rules mattered as much as the new visits. Twenty-seven records originally coded as official visits turned out to be meetings with the host leader on the sidelines of multilateral summits. Another fifteen described summits, conferences or three-way meetings rather than bilateral visits. Under our theory, neither kind signals support for the host, yet both had inflated

the official-visit count. For the period of our original analysis, 2009–2020, the picture is different. The dependent variable is binary: did a country receive at least one official visit in a given year? Of the 1,963 country-years in that sample, only one changed its value. Our assumption that international media capture high-level visits was thus largely right for the post-2008 period and for the most senior leaders. It was weakest exactly where coverage was weakest: the 1990s, and the prime minister's office.

Re-estimating the Original Models

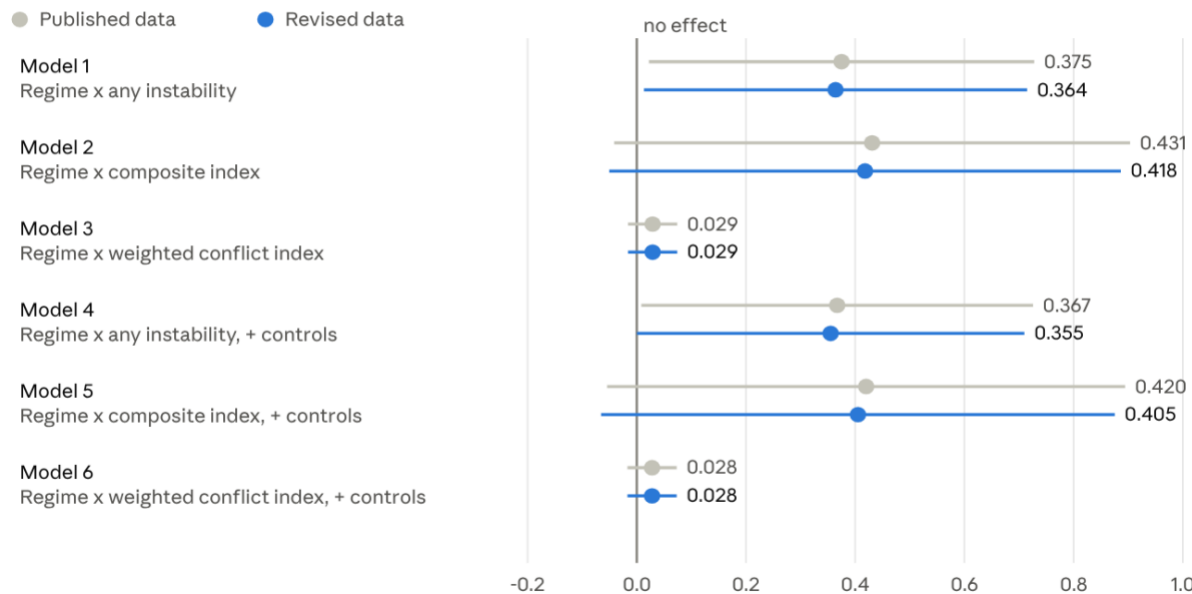
I re-ran the models from our original article with the revised data, keeping everything else unchanged. The original article, Balci and Aras (2025), used probit models of whether a country received an official visit from the Russian president, prime minister or foreign minister in a given year. Its key term interacted domestic instability with regime type for 2009–2020. I used the same specification, controls, year fixed effects and country-clustered standard errors. The main finding holds. In Model 1, instability in autocracies increased the probability of a Russian official visit, with an interaction coefficient of 0.375 in the original estimate and 0.364 with the revised data, both significant at $p < 0.05$. The joint test of instability's effect in autocracies moved from $p = 0.056$ to $p = 0.058$. Across all six interaction models, no coefficient moved by more than 0.02. Every estimate kept its sign and significance level. Population, GDP per capita and membership in the Commonwealth of Independent States remained strong predictors, and the other variables barely changed.

Why did a substantial audit change so little? Most errors sat in the 1990s and early 2000s, outside the main sample. The dependent variable records only whether a country received at least one official visit in a year, so a missed second visit or a misdated trip does not change it. The audit thus vindicates our original assumption, but only within limits. International media were good enough for a binary measure of visits by Russia's top leaders after 2008. They were not good enough for counting visits, for the 1990s, or for prime ministers. The re-analysis also caught an error in our own reporting. The figure we published as the "sum of coefficients" for instability in

autocracies (3.644) was in fact the test statistic for that sum; the sum itself is 0.254. The p-value we reported was correct.

The interaction estimates barely move with the revised data

Probit coefficients with 95% CIs, 2009–2020, 1,963 country-years. Model 1: 0.375 → 0.364



The robustness checks from the original article tell the same story. Restricting the sample to Africa, to presidential and prime-ministerial visits only, or to the pre-2009 period, and replacing the V-Dem regime measure with Polity or the Machine Learning Democracy Index, left every conclusion in place. No estimate changed sign, and only two crossed a significance threshold, in opposite directions: the MLDI interaction in Model 1 slipped from just below to just above $p = 0.1$, while in the president-and-prime-minister-only models the interaction in Model 4 moved from just above to just below it. The pre-2009 interaction terms, non-significant in the original, moved closer to zero with the revised data, so the contrast between the two periods on which the article rests became slightly sharper rather than weaker.

Implications and Discussion

Park and Rains (2026) conclude that the role of LLMs in political science data collection is conditional: they expand access for clearly structured tasks, while human judgment remains essential for context-dependent coding. Our experience confirms their division of labor and

extends it in one direction. In their study, the models worked within documents handed to them and were instructed not to look anything up. In ours, the model went out and found the documents. The task was not to extract fields from a known text but to search an archive in a language the researchers could not read, decide which of some 400,000 articles mattered, and trace each candidate back to its source. That is the step on which language and money had constrained us, and it is the step the model could take over. This gain has two limits of its own. Access to sources is no more evenly distributed than language skills: Kommersant's archive was open, the Russian government's website was not, and a model that cannot reach a source cannot read it. And our test was favorable to the original data. The dependent variable was binary and the main sample began in 2009, when official websites already existed; a study of visit counts, or of the 1990s, would have reached a different verdict. Those who rely on English-language sources for earlier periods or lower-ranked officials should assume gaps of the size I found, around a quarter of prime-ministerial visits before 2004.

The gain also depended on the kind of data we were collecting, and this is the main caution. The literature now converges on a pattern: LLMs do well where the answer is in the text and poorly where it must be inferred. Park and Rains found their models reliable on organization names and submission status but not on headquarters locations the documents did not state. Halterman and Keith (2025) show that even careful codebook prompting leaves models struggling with concepts such as protest or political violence, whose boundaries require judgment. Weidmann, Faulborn and García (2026) find that models asked to code V-Dem democracy indicators track the experts but carry systematic biases of their own, one model rating countries as less democratic than the experts do, the other as more. A leader visit sits at the easy end of this range. It is a visible act with a name, a date and a place, reported in the same words whoever reports it. Whether Chernomyrdin went to Copenhagen in November 1993 is a question with one answer, and the newspaper states it. Whether a regime is an electoral autocracy, or whether a demonstration counts as instability, is not.

Heseltine and Clemm von Hohenberg (2024) showed that LLMs can annotate political text across languages at expert level; our case shows they can also find that text when the researcher cannot. The audit cost no subscriptions and no research assistants. It required a laptop, an account, and a few days of supervision. That supervision was not incidental. Zhu, Wang and Zhang (2026) find that the reliability of LLM-assisted research depends less on the model than on where human decisions sit in the workflow, and the same model that fails in most unconstrained runs rarely fails when humans hold the gates and computation is deterministic. Our audit had that shape by design: the model searched and retrieved, code did the matching, and every coding rule and every change was a researcher's decision. The researcher still has to know the theory well enough to tell the model what counts, and read the sources well enough to check its claims. Within those limits, the gain is large. The study of leader visits is young and its datasets are few (Balci et al. 2025), each built by one team in the languages it happens to read. If enough researchers apply this method to their own regions, a global dataset of leader visits, collected from local sources rather than from the international press, is within reach.

References

- Balci, Ali, James D Kim, Jonathan D Moyer, Collin J Meisel, Kylie McKee, Alexander Baturo, Byungwon Woo, Seulah Choi, Minseon Ku, Eric Van Rythoven, and Marcus Holmes. 2025. "High-Level Leader Visits: A Promising Area of Study in IR." *International Studies Perspectives*: ekaf008. <https://doi.org/10.1093/isp/ekaf008>.
- Balci, Ali, and Bülent Aras. 2025. "Giving a Hand to Autocrats: Are Russia's High-Level Visits for Authoritarian Durability?" *Journal of Peace Research* 62 (7): 2307–21. <https://doi.org/10.1177/00223433251318911>.
- Halterman, Andrew, and Katherine A. Keith. 2026. "Codebook LLMs: Evaluating LLMs as Measurement Tools for Political Science Concepts." *Political Analysis* 34: 188–204. <https://doi.org/10.1017/pan.2025.10017>.
- Heseltine, Michael, and Bernhard Clemm von Hohenberg. 2024. "Large Language Models as a Substitute for Human Experts in Annotating Political Text." *Research & Politics* 11 (1). <https://doi.org/10.1177/20531680241236239>.

- Medie, Peace A., and Alice J. Kang. 2018. "Power, Knowledge and the Politics of Gender in the Global South." *European Journal of Politics and Gender* 1 (1–2): 37–53. <https://doi.org/10.1332/251510818X15272520831157>.
- Öztiğ, Lacin Idil. 2022. "The Global North/South Inequalities in the IR Discipline: Some Reflections and Insights." *Alternatives* 47 (2): 123–27. <https://doi.org/10.1177/03043754221095304>.
- Park, Chun-Young, and Matthew Rains. 2026. "Expanding Access, Preserving Judgment: Large Language Models in Social Science Data Collection." *International Studies Quarterly* 70 (4): sqag084. <https://doi.org/10.1093/isq/sqag084>.
- Pasha, Mustapha Kamal, and Craig N. Murphy. 2002. "Knowledge/Power/Inequality." *International Studies Review* 4 (2): 1–6. <https://doi.org/10.1111/1521-9488.00251>.
- Steinert, Janina I., David A. Nyarige, Milan Jacobi, Jana Kuhnt, and Lennart Kaplan. 2021. "A Systematic Review on Ethical Challenges of 'Field' Research in Low-Income and Middle-Income Countries: Respect, Justice and Beneficence for Research Staff?" *BMJ Global Health* 6 (7): e005380. <https://doi.org/10.1136/bmjgh-2021-005380>.
- Tickner, Arlene B. 2013. "Core, Periphery and (Neo)imperialist International Relations." *European Journal of International Relations* 19 (3): 627–46. <https://doi.org/10.1177/1354066113494323>.
- UNESCO and International Social Science Council. 2010. *World Social Science Report 2010: Knowledge Divides*. Paris: UNESCO. Available at: <https://unesdoc.unesco.org/ark:/48223/pf0000188333>.
- Weidmann, Nils B., Mats Faulborn, and David García. 2026. "Large Language Models Are Democracy Coders with Attitudes." *PS: Political Science & Politics* 59 (1): 17–23. <https://doi.org/10.1017/S1049096525101248>.
- Zhu, Chen, Xiaolu Wang, and Weilong Zhang. 2026. "(Human) Attention Is (Still) All You Need: Human Oversight Makes AI-Assisted Social Science Reliable." arXiv:2606.12848, preprint, not peer reviewed. <https://arxiv.org/abs/2606.12848>.