A longitudinal database of agricultural indicators from 1930 to 1960

Darya Teplykh[1], Kerstin Forster[1,2], Alessandro Schioppa[3], David Wuepper[3], and Stefan Feuerriegel[*,1,2]

[1]LMU Munich, Munich, Germany

[2]Munich Center for Machine Learning, Munich, Germany

[3]University of Bonn, Bonn, Germany

[*]Corresponding author: feuerriegel@lmu.de

Abstract

Long-term agricultural indicators are important for understanding long-term changes in farming systems. However, historical data, such as agricultural reports, are typically contained in non-machine-readable documents and are unavailable in a structured format, which makes their analysis difficult. Here, we present a new dataset covering 130 countries and 275 harmonized agricultural indicators from 1930 to 1960. Our dataset includes 10,983 unique country–year observations and includes key indicators of farm structure, land use, agricultural production and input use, such as farm population, number and area of holdings, and crop area and production. We created the dataset using a large language model (LLM), which we used to extract structured data from archived FAO World Census of Agriculture (WCA) reports. We validated the LLM-based pipeline with manual validation, where the LLM pipeline achieves an accuracy of 80.3%. We further compared our LLM pipeline against external databases, which are less comprehensive and are often derived from secondary sources rather than the raw country reports. Our dataset fills important gaps in existing historical data, as many values were previously missing or unstructured. Overall, the result provides a new resource for long-term analyses of agricultural change, enabling comparisons across countries and improving understanding of agricultural dynamics that played out beforethe start of most currently available datasets.

Introduction

Access to historical agricultural data can be important for researchers and policy makers alike. It is, e.g., regularly a pivotal input for understanding long-run economic and political development [1,2], and it can be relevant for the implementation of current sustainability policies to consider long-term trends and historical contexts, e.g., within global policy frameworks such as the Post-2020 Global Biodiversity Framework [3] and the UN Decade on Ecosystem Restoration (2021–2030) [4, 5]. Historical circumstances can also frequently explain current behaviors of farmers and other land users that would otherwise appear puzzling [6].

However, existing data needed to track agricultural development, such as land use, agricultural structures, and crop diversity, are typically contained in large archival documents (e.g., scanned images) [7] and are therefore not readily usable for downstream analysis. Hence, there is a growing call to make historical records available as structured, machine-readable data sources to support long-term research and evidence-based policy development [8].

Here, we create a structured dataset of historical agricultural indicators, covering 130 countries (see Supplementary Table S1) and consisting of 10,983 unique country-year observations across 275 harmonized long-term indicators (see Supplementary Table S2), which cover key categories such as as: holding and tenure, land utilization, crops, livestock and poultry, employment in agriculture, farm population, agricultural technology, irrigation and drainage, fertilizers and soil dressings (Fig. 1a). We specifically extract data for indicators covering various dimensions of farm structure, land use, agricultural production, and input use documented in the FAO World Census of Agriculture (WCA) reports. Our dataset focuses on the 1930, 1950, and 1960 census rounds, which track agricultural characteristics before most production statistics became annual. To construct the dataset, we developed an automated machine learning (ML) framework that extracts indicators from historical agricultural documents based on retrieval-augmented generation (RAG) (Fig. 1b).

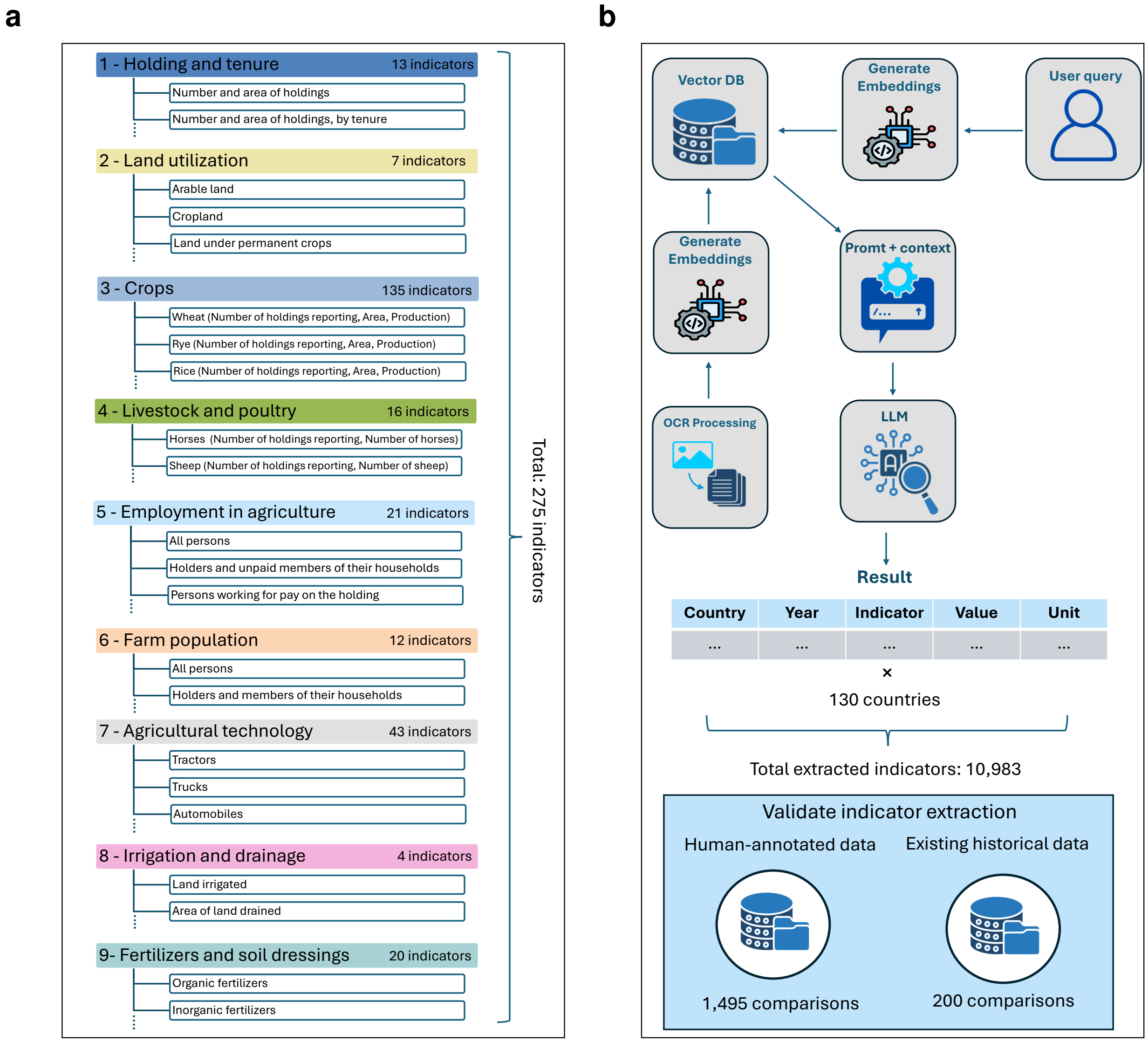

Figure 1: **Machine learning framework for extracting agricultural indicators.** **a**, Our framework aims to extract data on 275 harmonized long-term indicators (see Supplementary Table S2), distributed across 9 FAO World Census of Agriculture (WCA) subject categories for the major census rounds of 1930, 1950, and 1960. **b**, The figure presents a machine learning framework architecture designed to transform archival reports into a structured dataset. The process results in a structured dataset covering 130 countries and containing 10,983 unique extracted observations. To validate the reliability of the extracted data, the observations are compared against manual annotation (i.e., based on 1,495 manually-annotated observations). Additionally, a comparison is conducted with existing databases, but which are often limited in coverage.

Results

Dataset Description

Here, we created a dataset based on historical records from the FAO World Census of Agriculture (WCA), focusing on the 1930, 1950, and 1960 censuses. Our process resulted in a structured dataset containing 10,983 unique country-year observations across 130 countries. These observations capture 275 harmonized long-term indicators along key dimensions of agricultural structural change. The final dataset adheres to a standardized schema (country, year, indicator, value, unit). Example data is provided for Austria (Table 1), Japan (Table 2), Kenya (Table 3), and the United States (Table 4). The full dataset is available as a CSV file; see Data Availability statement.

Country participation in census rounds demonstrates uneven coverage. Only 17.5% of the countries provided data for all three census rounds, establishing a robust foundation for long-term analysis. Moreover, most territories participated in only one or two rounds, and some regions remained outside the scope of the censuses (Fig. 2a). A more detailed breakdown of indicators by region and category reveals heterogeneity in the dataset (see Supplementary Fig. S1). Europe shows a fairly comprehensive and balanced coverage. Latin America shows the highest number of indicators related to crop production. Africa and Asia show extensive data on crop production, but fewer indicators on land use, farm structure, and technology. North America and Oceania show lower overall indicators, mainly because fewer countries provide data for these regions than for Europe or Latin America, which naturally leads to fewer indicators being extracted, rather than gaps in individual reports.

The extracted indicators show a strong historical focus on economic performance (Fig. 2b). The crops category dominates, accounting for 42.7% of all extracted data. Together, the crops and livestock categories account for almost 60% of the data corpus. This reflects a systematic historical focus on key economic outputs rather than structural or social indicators.

The distribution of extracted indicators by category remains relatively stable throughout all three census rounds (Fig. 2c). The crop category dominates in all periods (33-50%), confirming the consistent priority given to production indicators. Data on agricultural technology increased from 9% in 1930 to 11.2% in 1960, illustrating the growing attention to mechanization processes. The percentage of data on land utilization has also increased, reflecting the importance of accounting for land structure in understanding agricultural transformations.

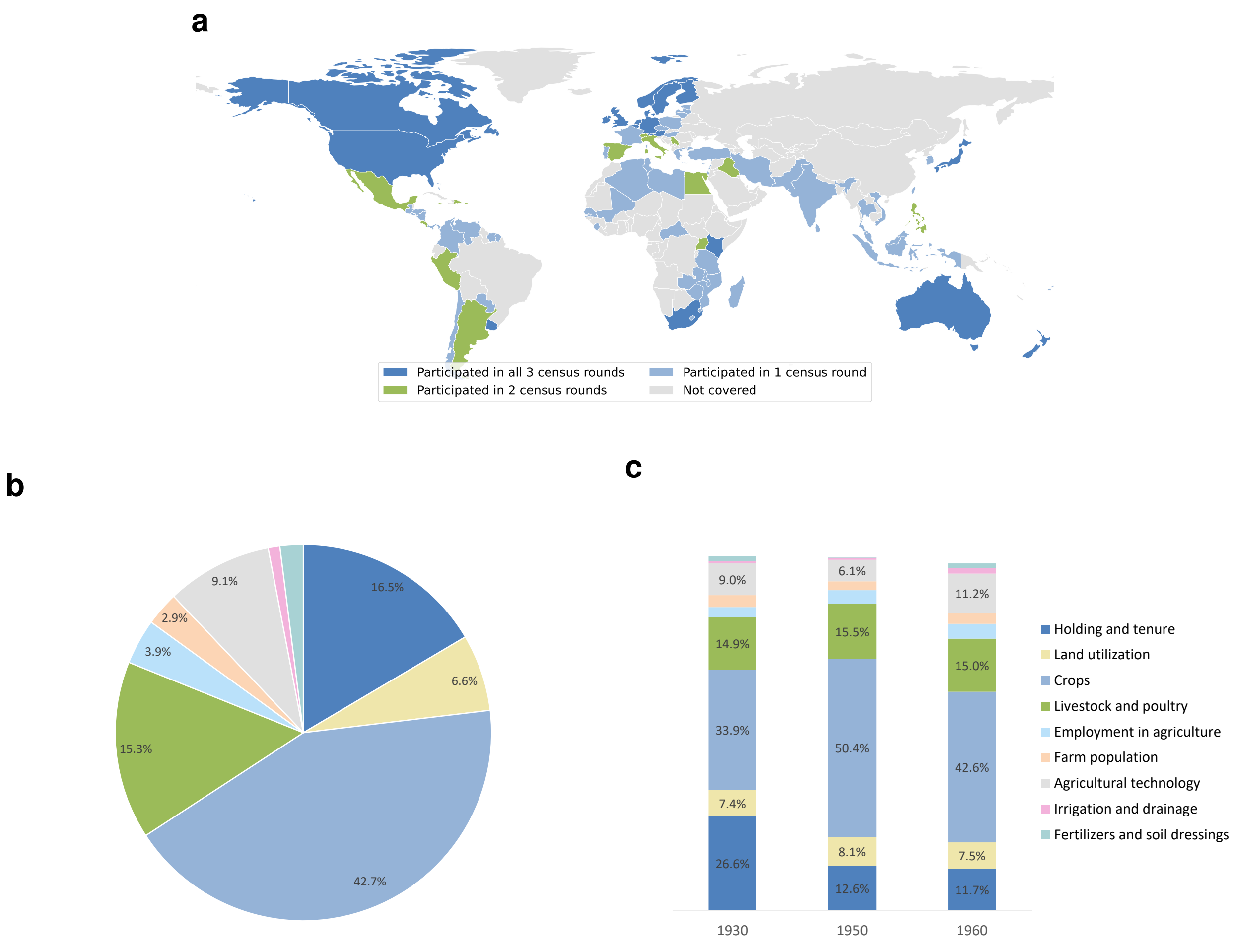

Figure 2: **Geographic and thematic structure of extracted agricultural census data (1930-1960).** Shown are: **a**, Global coverage of country participation across three census rounds. **b**, Distribution of extracted indicators by thematic categories. **c**, Temporal evolution of indicator distribution by category across three census rounds.

Country	Year	Indicator	Value	Unit
Austria	1930	Number of holdings	412 283	number
Austria	1930	Area of holdings	18 850 627	acres
Austria	1951	Number of holdings	432 848	number
Austria	1951	Area of holdings	7 726 228	hectares
Austria	1960	Number of holdings	396 530	number
Austria	1960	Area of holdings	7 683 888	hectares

Table 1: **Austria, number and area of holdings (1930–1960).**

Country	Year	Indicator	Value	Unit
Japan	1929	Rice - Area	7 868 124	acres
Japan	1929	Rice - Production	107 426 256	1000 lb
Japan	1949	Rice - Area	5 434 719	hectares
Japan	1949	Rice - Production	2 768 672	metric tons
Japan	1959-60	Rice - Number of holdings reporting	5 363 668	number
Japan	1959-60	Rice - Area	3 25 7 722	hectares
Japan	1959-60	Rice - Production	12 487 123	metric tons

Table 2: **Japan, harmonization of rice production data (1929–1960).**

Country	Year	Indicator	Value	Unit
Kenya European Holdings	1930	Cattle	540 445	head
Kenya Indian Holdings	1930	Cattle	1 974	head
Kenya European and Asia holdings	1954	Cattle	706 500	head
Kenya European and Asia holdings	1960	Cattle	979 600	head
Kenya African holdings	1960	Cattle	1 597 400	head

Table 3: **Kenya, cattle inventory (1930–1960).**

Country	Year	Indicator	Value	Unit
USA	1929	Tractors - Number of holdings reporting	851 457	number
USA	1929	Tractors - Number of tractors	920 021	number
USA	1950	Tractors - Number of holdings reporting	2 525 206	number
USA	1950	Tractors - Number of tractors	3 609 281	number
USA	1959	Tractors - Number of holdings reporting	2 679 561	number
USA	1959	Tractors - Number of tractors	5 138 921	number

Table 4: **United States of America: tractors and mechanization (1929–1959).**

Trends in agricultural transformation

To analyze agricultural transformation and structural differences in land use, we grouped countries by region: Europe, North America, Latin America, Oceania, Africa, and Asia.

In Europe, we can see the changes based on data from eight countries: Austria, Belgium, Denmark, Finland, Germany, the Netherlands, Norway, and Sweden. This group of countries provided the most complete and consistent statistical coverage, submitting reports for all three rounds of censuses. These eight countries experienced a wave of mechanization in the postwar period. The number of tractors in Norway increased 63-fold from 889 to 55,786 and, in Germany, 30-fold from 76,699 to 2,264,113. This technological revolution triggered a series of changes: the area under oats, the traditional feed for draft horses, plummeted by 77% in Austria, while the area under wheat and sugar beets increased by 21%. At the same time, farms were consolidated. In Austria, their number decreased by 8.5%, while the total area remained stable. This meant the absorption of small farms by large ones, which led to the enlargement of the average production unit.

Reports from North America (The United States, Canada, and Guam) show a high degree of mechanization and structural consolidation. By 1961, the USA tractor fleet had 5.1 million tractors, which exceeded the total for the whole of Europe. This led to a sharp decline in the number of horses in the US, which fell by 78% during the period under review. Structural consolidation can also be seen, with the number of farms falling by 40% and the average farm size increasing to 123 hectares. Data for Canada also confirm this trend: between 1931 and 1961, the farm population declined by 35.29%, the number of farms declined by 34%, and the size of farms increased by 60%.

In the reports on Latin America submitted by Uruguay, Puerto Rico, and the Virgin Islands, we see a picture that is the opposite of that in industrialized North America. In Uruguay, despite the fact that up to 82% of agricultural land was allocated to pastures, there was a catastrophic reduction in the cattle population of 57% between 1929 and 1961. In small territories such as Puerto Rico, the number of people employed in agriculture fell by 55.8%. This outflow was not caused by the successful replacement of labor with capital, as the level of mechanization remained extremely low only 3,338 tractors by 1959.

The Oceania reports are interesting because they combine two opposing types of agricultural systems: the industrial system (Australia and New Zealand) and traditional farming (American Samoa). The number of tracts in Australia increased by 133% between 1960 and 1970. The area of irrigated land in Australia increased almost ninefold, while in New Zealand, irrigation remained stable at around 64,000 hectares. Moreover, by 1960, chemicalization had become widespread in Australia, where inorganic fertilizers were used on an area of 17.3 million hectares. Livestock specialization increased. Sheep farming demonstrated growth (Australia: +48%, New Zealand: +79%), and the cattle population in New Zealand quadrupled. In contrast to Australia and New Zealand, in small areas such as American Samoa, the pattern was the opposite, where the number of farms increased by 161.96% between 1930 and 1960, and the average farm size was only 2.18 hectares. Production was concentrated on traditional permanent crops (coconut, cocoa, yams).

We can analyze the development of African territories using Kenya as an example. It demonstrates a classic case of colonial dualism, in which two agricultural sectors existed in the same territory but operated in fundamentally different economic and technological realities. The European sector consisted of only 3,609 farms, which controlled 3.13 million hectares of land with an average size of 867 hectares. This sector was fully mechanized, with up to 1,770 tractors per thousand farms. It is also possible to note that the European sector specialized in export crops (wheat, coffee, tea) and demonstrated growth in cattle numbers. The African sector, on the other hand, had 734,300 farms with an average size of only 4 hectares and focused on food crops (maize, millet), with a complete lack of access to mechanization.

Agricultural development in Asia is analyzed using Japan, where we see a shortage of arable land. By 1960, the average farm size was only 1.18 hectares, making Japanese farms among the smallest in our report. Pressure on land resources was extreme, as the population density on agricultural land reached 459 people per square kilometer. Nevertheless, the unique combination of land scarcity and labor surplus contributed to high yields. The main focus was on rice production, which accounted for 46% of arable land and yielded a record 4.45 tons per hectare. This was made possible by advanced irrigation (54% of arable land) and the widespread use of organic fertilizers (9.4 tons per hectare).

During the period under review, global agriculture entered an era of mechanization and structural transformation, as can be seen in the diagrams in Figure 3. Figure 3a shows changes in the average size of agricultural holdings in six regions between 1930 and 1960. Overall, figures for Europe showed little change during this period, but a detailed analysis of the data (see Supplementary Table S3) reveals two different trends. Central Europe largely experienced stability in terms of land area, with a reduction in the number of farms. The total area of agricultural land remained at the same level, but the number of farms decreased. In Scandinavia, the changes were more varied. For example, in Norway, there was a consolidation of farms from 5.7 to 17.6 hectares, +208%, while, in Sweden and Finland, land ownership was fragmented under the influence of social reforms and natural constraints.

The most notable changes are observed in Latin America and Oceania. In Oceania, the average farm size increased more than fivefold, from approximately 265 to 1,842 hectares (+540%). Latin America also recorded a large increase, from 31.8 to 132.6 hectares (4 times). This jump is related to accounting practices because, in the 1930s, a number of countries in the region did not include permanent meadows and pastures in their accounts, which understated the average figures. However, by 1950, the statistical coverage expanded, and permanent meadows and pastures accounted for about 80.6% of the total (see Figure 3c, which shows the distribution of agricultural land use by region).

A similar situation can be observed in Africa (Fig. 3c), where data for 1930 is based on a different accounting system, making it difficult to compare directly with later periods. The diagram for 1930 mainly shows forest land and other areas. In 1950, reporting became more detailed. Separate categories appear for arable land (15.4%) and permanent meadows and pastures, which dominate the structure of African land (accounting for about 65%). This redistribution reflects not radical changes in land use, but rather a transition to more accurate statistical accounting. In 1960, we see that the share of pastures and arable land visually decreases, while the category "other" increases. These changes may also be related to further refinement of reporting, with some areas being reclassified into new subcategories.

In Europe and North America, the quality of statistical land use records in 1930–1960 was higher than in other regions, which allows for more reliable comparisons of data over time. In Europe, the diagram shows a gradual change in land use structure. In 1930, 52.2% of the area was arable land, 16.6% was permanent meadows and pasture, and 31.2% was wood and forest land. By 1950, the share of arable land had decreased, while the share of forests had almost doubled to 54.9%. This was due to the post-war restoration of forest areas and the reclassification of underutilized land. In 1960, the growth of pastures and the reduction of arable land reflected a shift in agriculture towards livestock farming. Arable land and pastures each accounted for 20%. Finally, North America shows a different land use pattern compared to Europe. Here, there is a smaller proportion of forests and a significantly higher proportion of cropland. In 1930, arable land dominated, accounting for about 64%, reflecting the strong focus of agriculture in the US and Canada on grain and industrial crops. Pastures accounted for 22.4% during this period. By 1960, we see that pastures increased their share to 36.4%, which is associated with the growth of livestock farming, the expansion of pasture land, and the transfer of part of the land from the arable land category to the pastures category. Accordingly, the share of arable land decreased to 41.6%.

To analyze changes in the structure of agricultural crops, we selected the seven most common crops in our dataset: wheat, maize, oats, potatoes, barley, cotton, and rye (Fig. 3d). Their distribution allows us to observe long-term changes in agricultural specialization. In almost all countries, wheat accounts for a large part of the crops and has been one of the main crops throughout the whole period. In Europe, wheat has consistently accounted for around 40% of the area under cultivation in each census cycle. There has also been a decrease in the area under oats, while the share of potatoes and barley has increased. In 1960, barley accounted for 17.3% and potatoes for 11.9%. In North America, we also see a decline in the share of oats. At the same time, reports show data on cotton, whose share is falling and by 1960 is only 4.9%. Starting in 1950, maize appeared in the crop structure and quickly became an important element of North American agriculture. By 1960, its share reached 26.9%, which probably reflects the growth in demand for feed crops and the development of livestock farming. In Africa, data is only available for 1930 and 1960, as no reports were submitted for 1950. In 1930, maize dominated the crop structure, accounting for 72.3%, while wheat accounted for 25.6%. By 1960, the role of maize became even more pronounced, reaching 80.9%, while wheat declined to 7.9%. Asia, where the crop structure differs from other regions. The diagram does not include data on wheat. Barley dominates here, accounting for about 53–56% of the sown area throughout the period. It should be noted that rice, despite its central role in Asian agriculture, did not make it into the top 7 most common crops in the global dataset because countries outside Asia hardly grew rice during the period under review. Therefore, it is not included in the diagram (Fig. 3d).

Finally, we analyzed the process of agricultural mechanization around the world and compared how the average number of tractors and the average number of workhorses changed during this period (Fig. 3b). Between 1930 and 1960, the global agricultural system went through some fundamental changes. For example, the number of tractors grew steadily from 0.23 million to 1.14 million, while the number of workhorses declined sharply from 4.44 million to 2.39 million. This difference reflects the transition from traditional systems to mechanized farming. These changes influenced crop patterns, pasture use, and the formation of regional agricultural specialization.

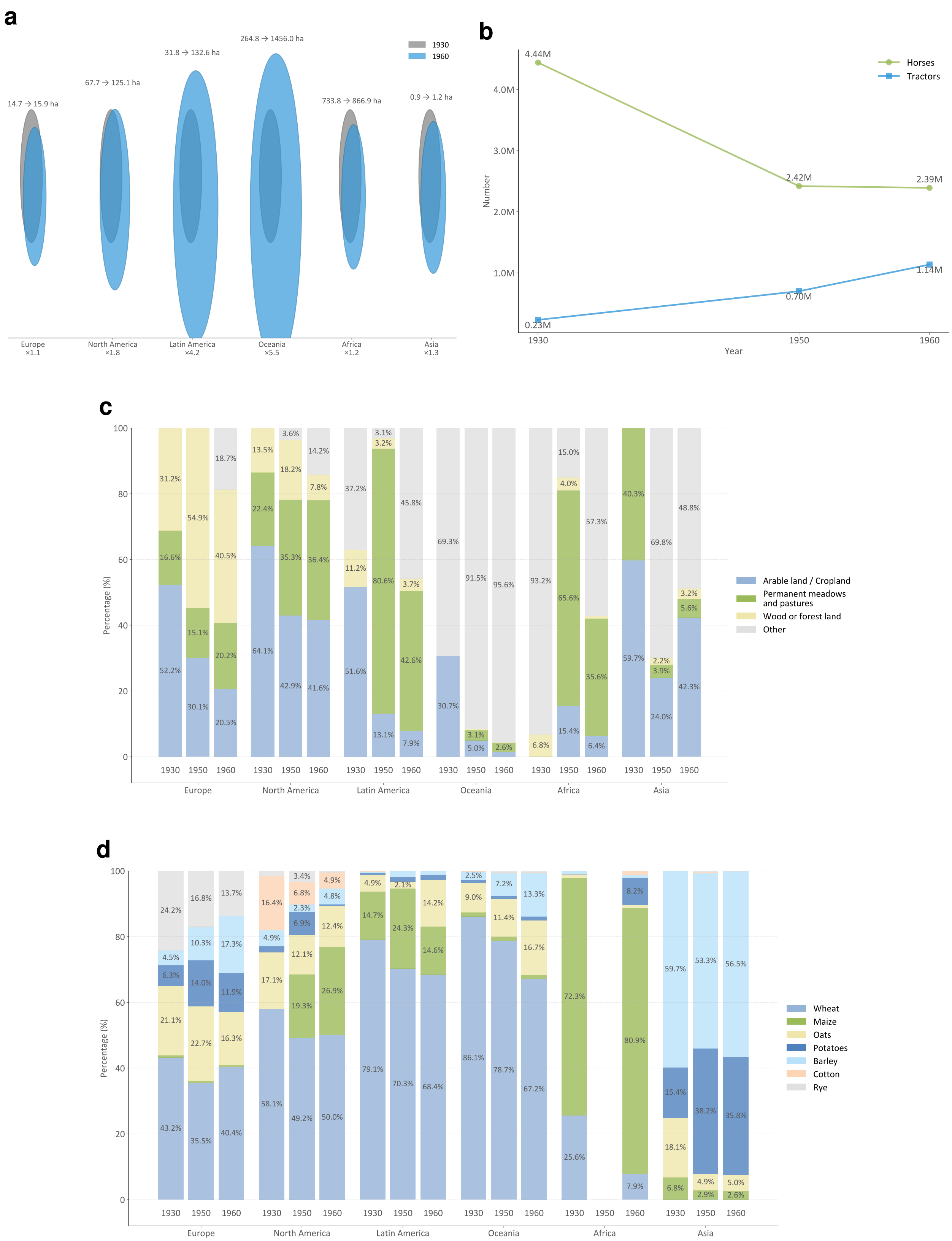

Figure 3: **Comparative regional trends in agricultural structure and production (1930-1960).** Shown are: **a**, Average holding size changes by region. Bubbles for 1930 are normalized to equal size, while bubbles for 1960 are scaled proportionally to the ratio of average holding sizes (1960/1930), illustrating how many times holding sizes increased in each region. **b**, Global dynamics of tractors and workhorses in agriculture. **c**, Land utilization by region. **d**, Crops by region and year.

Validation

We applied two validation strategies to assess the reliability of our ML extraction framework: (1) manual validation and (2) comparison against historical datasets. In the first approach, we manually verified 1,495 extracted records across all nine categories using a strict criterion: a record was considered valid only if the extracted value, year, and category were absolutely accurate. The results showed that 1,200 records met this standard and an accuracy rate of 80% (Table 5). These results demonstrate a high degree of internal consistency and reliability of the automated information extraction algorithms, confirming that the system is stable.

In manual inspection, most of the errors identified during the manual audit were caused by inaccurate reading of data from the documents themselves. These errors occurred at the OCR level. The most common types of errors were merging of values, where numbers from adjacent cells were combined into a single value. Replacement of a number, such as incorrect recognition of 1 instead of 7 or partial reading of a value, as well as shifting of rows or columns, when a number intended for one indicator was mistakenly read as a value for another, because the system shifted one row down in the table structure.

In the second approach, we evaluated the scientific plausibility of extracted trends by comparing them with independent historical datasets [9–11]. However, these datasets have limited coverage, with data available only for two categories: crops and agricultural technologies. After harmonizing measurement units (hectares and metric tons), this comparison, based on 200 comparisons, showed an accuracy of 69%. Correlation analysis using the Pearson correlation test revealed a strong positive association between the degree of temporal overlap and extraction accuracy (r = 0.726, $p < 0.001$). Independent samples t-test confirmed significant differences in accuracy based on temporal coverage (t = 17.554, $p < 0.001$), with accurate matches showing 96.5% temporal overlap compared to 28.0% for inaccurate matches. Importantly, this value should not be interpreted as a direct accuracy metric, as the benchmark datasets are not derived from original census records but from secondary compilations that often aggregate, round, or interpolate data across years. In contrast, our dataset is based on *raw* data extracted directly from FAO World Census of Agriculture reports, preserving the original granularity, definitions, and temporal specificity of the underlying sources. Thus, while existing databases provide valuable but simplified summaries, our dataset offers a faithful representation of historical agricultural conditions and enables analyses that were previously impossible due to missing or unstructured data.

Category	Manual validation accuracy (%)	Benchmarking accuracy (%)
Holding and tenure	83	—
Land utilization	87	—
Crops	75	71.3
Livestock and poultry	91	—
Employment in agriculture	87	—
Farm population	77	—
Agricultural technology	89	66.7
Irrigation and drainage	64	—
Fertilizers and soil dressings	70	—

Table 5: **Data extraction accuracy by category.**

Discussion

Large language models (LLMs) demonstrate great potential for generating, extracting, and analyzing information from unstructured text data in a variety of scientific fields [12–14]. LLM models have proven their effectiveness as a powerful tool for scaling analysis and reducing the manual labor required to synthesize huge and ever-growing volumes of data [13, 15]. In agricultural and environmental sciences, the application of LLM can help in the analysis of complex environmental and climate issues, such as biodiversity loss [16]. Yet, their potential for reconstructing and analyzing historical agricultural and environmental data remains largely untapped. This gap is critical because existing datasets on agricultural and land-use systems are often fragmented, inconsistently reported, or missing entirely for earlier periods, particularly for the historical agricultural landscapes that underpin long-term analyses of sustainability and land-use change [17].

It is important to note that unstructured and extremely heterogeneous archival materials pose a barrier to quantitative analysis, as reliable assessment of agrobiodiversity depends on understanding structural changes and land use dynamics at the farm level [18]. This problem is exacerbated by complex document structures, typographical defects such as non-standard fonts and low-quality fonts, as well as the presence of complex table structures and nested headings [19]. Our study aims to fill this gap: we demonstrate the effective use of LLM to extract long-term agricultural indicators from complex archival data, expanding the available data for land-use monitoring.

Our machine learning framework offers several key strengths. First, our framework addresses the challenge of mining massive data volumes with thousands of pages of archival material. Consistent with recent advances demonstrating the ability of LLMs to extract quantitative data from complex, unstructured documents [20–22], our pipeline makes it economically viable to produce 10,983 unique country–year observations across 130 countries. This highlights the scalability of LLMs in data extraction by circumventing the need for otherwise labor-intensive manual work. Second, our framework combines OCR with LLMs to extract data with contextual understanding rather than simple text recognition. This fusion allows the system to identify not only numerical values but also their meaning and placement within tables, captions, and paragraph structures. As a result, the pipeline can accurately associate figures with the correct variables, years, and categories, which is typically precluded in conventional OCR or rule-based methods (see the discussion in [23]). Third, as shown in our analysis, our framework is highly robust to variations in source quality, and it can reliably handle diverse layouts, typefaces. Fourth, our framework is built on state-of-the-art open-source LLMs, ensuring transparency, reproducibility, and accessibility for other researchers who wish to extend or adapt the approach to different historical datasets.

Our dataset and the underlying LLM pipeline open up new opportunities for monitoring and analyzing agricultural and environmental trends. Using previously unavailable historical agricultural data, we can track and analyze long-term changes in agricultural landscapes, which is important for assessing the effectiveness of environmental and agricultural policies.

Methods

Data

Our sample consists of official FAO World Census of Agriculture (WCA) reports, which compile country-level census results across Europe, the Americas, and countries concerned with statistical programs in underdeveloped areas [24]. They provide harmonizable statistics on holders, holdings and tenure, land utilization, crops (area and production), livestock and poultry inventories, farm population, agricultural power and machinery, irrigation and drainage, and fertilizers and soil dressings. Tables are issued in two formats: aggregate series not classified by size of holding and disaggregated series classified by size of holding. Based on this corpus, our dataset covers 130 countries drawn from 8 reports and includes around 275 agricultural indicators.

For all reporting countries, we collected PDF files between 1930 and 1960 directly from FAO's Open Knowledge and Statistics repositories [24]. Overall, PDF files have an average length of 300 pages, and the total length of the dataset is 2,400 pages. As part of our machine learning system, we standardize units of measurement and, where needed, normalize category labels. Country and census year are taken directly from the source reports and stored as metadata. Each scanned page is indexed as one independent chunk.

ML framework

We develop an ML framework to extract relevant agricultural indicators from archival FAO reports using retrieval-augmented generation (RAG) [25]. The pipeline has five steps (Fig. 1b): In Step 1, we collect FAO census PDFs, run image-to-text OCR on scanned pages, and segment pages into chunks (*preprocessing*). In Step 2, the data are converted into vector form using an embedding model and stored in a vector database (*indexing*). In Step 3, for a given parameter query, we fetch and re-rank the top-k most relevant chunks (*retrieval*). In Step 4, the target agricultural parameter and the retrieved report chunks are placed into a structured prompt and sent to the LLM, which outputs the generated responses (*generation*). Finally, in Step 5, data output is formatted, units of measurement and numerical formats are standardized (*postprocessing*).

Step 1: Preprocessing

At the first preprocessing stage, the main goal was to extract text from scanned PDF files. Each page was first converted into a high-resolution image using the pdf2image library, which interfaces with Poppler [26]. The resulting images were then processed with optical character recognition (OCR) via the Mistral OCR to obtain selectable text [27]. We selected Mistral OCR after a comparative evaluation against PaddleOCR [28]. Specifically, Mistral achieved a higher accuracy rate: in our benchmark of 45 test cases, it correctly processed 37 (82%), while PaddleOCR reached only 53% accuracy on the same set. Although both tools are widely used, Mistral proved to be more reliable when working with noisy historical documents and complex tabular structures, which dominate our dataset. After text extraction, the content was split into chunks. In this project, we adopted a “one page–one chunk” strategy. This approach was chosen to preserve contextual integrity: in these reports, tables and statistical summaries often occupy an entire page and constitute a single semantic unit, so finer splitting would risk fragmenting the data and breaking relationships. This strategy is consistent with common guidance on chunking for retrieval-augmented systems [29].

Step 2: Indexing

Text chunks from Step 1 are encoded into 384-dimensional vectors using the Sentence-Transformers model all-MiniLM-L6-v2 [30]. The resulting embeddings, together with the original chunk text and a compact metadata dictionary (country, year, page), are stored in a persistent Chroma vector database collection [31]. Adding metadata to the embeddings enables efficient vector search with metadata filtering, allowing candidates to be restricted by country or by census year.

Step 3: Retrieval

Data retrieval is initiated by a natural language query for an indicator, which can be a simple question "What was the wheat production in Germany in 1930, 1950, and 1960?". The query specifies an indicator, a geographic location (country), and one or more years.

For each parameter query, the query text is embedded with all-MiniLM-L6-v2, and cosine-similarity nearest-neighbor search is performed over the Chroma index. The top-5 most similar chunks are returned and are concatenated in source order with clear delimiters to form the context window that is passed to the LLM in Step 4. In the baseline setup, retrieval relies on dense vector similarity, with optional metadata filters applied at query time to refine the candidate pool.

Step 4: Generation

At the generation stage, the top-k chunks retrieved in Step 3 are merged into a single context and, together with the original query, inserted into a structured prompt. The prompt was designed following best practices in prompt engineering [32–34] and consists of two main parts. The detailed prompt is available in Supplementary Materials S1. Moreover, the prompt includes recommendations for the expected output format, directing the model to structure the data as a markdown table, and the model is prohibited from using external knowledge, which is important for reducing the risk of hallucinations.

Generation is performed with deepseek-ai/DeepSeek-R1-Distill-Llama-70B served via Together AI [35]. We use this model because it is an open-source model, follows strict instructions, and keeps the fixed table schema reliably, which is crucial for extracting numbers with units from historical, table-heavy documents. In practice, the model adheres well to the context-only rule, interprets complex headers and footnotes, and returns stable, schema-compliant outputs under low-temperature decoding, which makes it well suited to our RAG pipeline.

Step 5: Postprocessing

Finally, we perform the necessary postprocessing steps to ensure that the data is structured and ready for quantitative analysis. The results are formatted according to a specific standardized scheme with mandatory columns: country, year, indicator, value, and unit of measurement. We apply coverage rules by grouping by country, returning all available values, and including multi-year and multi-parameter cases. The final dataset is stored in a standardized CSV file.

Validation

The validation of our ML framework consists of two methods: (1) a human-annotated dataset evaluates the accuracy of the framework. This check compared automatically extracted values with manually entered reference data. The sample size for this technical check was 1,495 comparisons, as the amount of manual reference data covered all seven categories. The comparison was performed line by line, with a match recorded only under the strictest condition: if the entire extracted line was completely identical to the reference, including the value, year, and category. (2) Benchmarking against existing databases. This verification focused on 200 comparisons of extracted values with reference series from historical literature [9–11]. For the categories crops, agricultural power, and machinery, linear trends were constructed based on reference data, against which actual values were evaluated. It should be noted that, in these categories, crop areas were recorded in hectares and production volumes in metric tons, which we transformed to SI units.

Both validation methods have limitations. (1) The human-annotated dataset, although extremely rigorous and covering all seven categories, is limited in size due to the high labor costs associated with manual annotation. (2) Benchmarking against existing databases validation was limited to only two categories (crops, agricultural power, and machinery) as well as a lack of historical literature, resulting in a relatively small sample size. This means that it confirms the overall reliability of the data only for these specific areas. Consequently, neither method alone is capable of covering the entire volume of data we have extracted, but their combined use provides a reliable and multifaceted assessment of our framework.

Data availability

The complete dataset of extracted agricultural indicators is available at https://github.com/daryateplykh/agricultural_indicators_extractor/tree/main/rag_outputs

Code availability

All code to replicate our analyses is available at https://github.com/daryateplykh/agricultural_indicators_extractor. `deepseek-ai/DeepSeek-R1-Distill-Llama-70B` is available at https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Llama-70B.

References

[1] Nunn, N. The historical roots of economic development. In *Science, 367.6485*, eaaz9986 (2020).

[2] Nunn, N. The importance of history for economic development. *Annu. Rev. Econ.* **1**, 65–92 (2009).

[3] Secretariat of the Convention on Biological Diversity. Kunming-montreal global biodiversity framework (2022). URL https://sdgs.un.org/un-system-sdg-implementation/secretariat-convention-biological-diversity-cbd-54146.

[4] Programme, U. N. E. Un decade on ecosystem restoration: Action plan booklet (2023). URL https://wedocs.unep.org/20.500.11822/42535.

[5] Mirzabaev, A. & Wuepper, D. Economics of ecosystem restoration. In *Annual Review of Resource Economics, 15.1*, 329–350 (2023).

[6] Wuepper, D., Bukchin-Peles, S., Just, D. & Zilberman, D. Behavioral agricultural economics. *Applied Economic Perspectives and Policy* **45**, 2094–2105 (2023).

[7] Kuemmerle, T. *et al.* Challenges and opportunities in mapping land use intensity globally. In *Current Opinion in Environmental Sustainability, 5(5)*, 484–493 (2013).

[8] Philips, J. P. & Tabrizi, N. Historical document processing: Historical document processing: A survey of techniques, tools, and trends (2020). URL https://arxiv.org/abs/2002.06300.

[9] Binswanger, H. Agricultural mechanization: a comparative historical perspective. In *The World Bank Research Observer*, 27–56 (1986).

[10] Gollin, D., Hansen, C. W. & Wingender, A. M. Two blades of grass: The impact of the green revolution. In *Journal of Political Economy, 129(8)*, 2344–2384 (2021).

[11] Trends in U.S. Agriculture (2018). URL https://www.nass.usda.gov/Publications/Trends_in_U.S._Agriculture/Mechanization/index.php.

[12] Feuerriegel, S., Zschech, P., Hartmann, J. & Janiesch, C. Generative AI. In *Business & Information Systems Engineering, 66(1)*, 111–126 (2024).

[13] Raeissi, M. M. & Knapen, R. Applications of generative large language models in environmental science: A systematic review. In *Advances in Environmental and Engineering Research, 6(3)*, 1–15 (2025).

[14] Toetzke, M., Probst, B. & Feuerriegel, S. Leveraging large language models to monitor climate technology innovation. In *Environmental Research Letters, 18(9)*, 091004 (2023).

[15] Moorthy, S. M. K., Qi, M., Rosen, A., Malhi, Y. & Salguero-Gomez, R. Harnessing large language models for ecological literature reviews: A practical pipeline (2025). URL https://ecoevorxiv.org/repository/view/8516/.

[16] Hald-Mortensen, C. The main drivers of biodiversity loss: a brief overview. In *Journal of Ecology and Natural Resources, 7(3)*, 000346 (2023).

[17] Foster, D. *et al.* The importance of land-use legacies to ecology and conservation. In *BioScience, 53(1)*, 77–88 (2003).

[18] Reidsma, P., Tekelenburg, T., van den Berg, M. & Alkemade, R. Impacts of land-use change on biodiversity: An assessment of agricultural biodiversity in the european union. In *Agriculture, Ecosystems & Environment, 114(1)*, 86–102 (2006).

[19] Nguyen, T. T. H., Jatowt, A., Coustaty, M. & Doucet, A. Survey of post-ocr processing approaches. In *ACM Computing Surveys (CSUR), 54(6)*, 2 (2021).

[20] Forster, K. *et al.* Assessing corporate sustainability with large language models: Evidence from europe. In *Available at SSRN 5361703* (2025).

[21] Stürenburg, F. *et al.* Tracking funding disparities in global health aid with machine learning. In *medRxiv*, 2025.06 (2025).

[22] Toetzke, M., Banholzer, N. & Feuerriegel, S. Monitoring global development aid with machine learning. In *Nature Sustainability, 5(6)*, 533–541 (2022).

[23] Vilkomir, K. & Herndon, N. Challenges of automatic document processing with historical data. In *Proceedings of the 2024 ACM Southeast Conference*, 50–59 (2024).

[24] FAO Statistics Resource Repository (2025). URL https://www.fao.org/statistics/resources/3/en.

[25] Lewis, P. *et al.* Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems, 33*, 9459–9474 (2020).

[26] pdf2image's documentation (2022). URL https://pdf2image.readthedocs.io/en/latest/index.html.

[27] Mistral OCR (2025). URL https://mistral.ai/news/mistral-ocr.

[28] PaddleOCR Documentation (2025). URL https://www.paddleocr.ai/latest/en/index.html.

[29] LangChain. Text splitters (2025). URL https://python.langchain.com/docs/concepts/text_splitters.

[30] Hugging Face. all-MiniLM-L12-v2 (2021). URL https://huggingface.co/sentence-transformers/all-MiniLM-L12-v2.

[31] Chroma (2025). URL https://docs.trychroma.com/docs/overview/introduction.

[32] Lin, Z. How to write effective prompts for large language models. In *Nature Human Behaviour*, 611–615 (2024).

[33] Giray, L. Prompt engineering with ChatGPT: a guide for academic writers. In *Annals of Biomedical Engineering , 51(12)*, 2629–2633 (2023).

[34] Feuerriegel, S. *et al.* Using natural language processing to analyse text data in behavioural science. In *Nature Reviews Psychology, 4(2)*, 96–111 (2025).

[35] Together AI (2025). URL https://www.together.ai/.

Acknowledgments

SF acknowledges funding via the Swiss National Science Foundation (SNSF), Grant 186932.

Author contributions

DT, KF, and SF designed the LLM pipeline. DT implemented the LLM pipeline and performed the analysis. DT and SF wrote the first draft. All authors contributed to conceptualization, manuscript writing, and approved the manuscript.

Competing interests

The authors declare no competing interests.

Supplements

Supplementary Tables

Supplementary Figures

Supplementary Materials

Supplementary Tables

Table S1: List of countries grouped by continent

Africa	
Aden Protectorate	Northern Rhodesia
Algeria	Nyasaland
Bechuanaland Protectorate	Portuguese Guinea
British Somaliland	Saint Helena
Central African Republic	Senegal
Egypt	Seychelles
French West Africa	Sierra Leone
Gambia	South Africa
Gold Coast	South West Africa
Kenya	Southern Rhodesia
Lesotho	Swaziland
Libya	Tanganyika
Madagascar	Tunisia
Mali	Uganda
Mauritius	Union of South Africa
Mozambique	Zanzibar and Pemba
Nigeria and British Cameroons	
North America	
Alaska	Hawaii
Bermuda	Mexico
Canada	United States of America
Guam	

Table S1: List of countries grouped by continent (continued)

Latin America	
Argentina	Jamaica
Bahamas	Leeward Islands
Barbados	Nicaragua
British Guiana	Panama
British Honduras	Paraguay
Chile	Peru
Colombia	Puerto Rico
Costa Rica	Surinam
Dominican Republic	Trinidad and Tobago
El Salvador	Uruguay
Falkland Islands	Venezuela
Guatemala	Virgin Islands (U.S.)
Honduras	Windward Islands

Asia	
Brunei	Lebanon
Ceylon (Sri Lanka)	Malaysia
China (Taiwan)	North Borneo
India	Pakistan
Indonesia	Philippines
Iran	Ryukyu Islands
Iraq	Sarawak
Israel (Arabs, Druzes and other Minority Groups Sector)	Singapore Island
Israel (Jewish Sector)	Thailand
Japan	Turkey

Table S1: List of countries grouped by continent (continued)

Korea, Republic of	United Arab Republic
Malaya, Federation of	Viet-Nam, Republic of
Europe	
Austria	Latvia
Belgium	Lithuania
Cyprus	Luxembourg
Czechoslovakia	Malta and Gozo
Denmark	Netherlands
England and Wales	Norway
Estonia	Poland
Finland	Portugal
France	Saar, the
Germany	Scotland
Greece	Spain
Hungary	Sweden
Ireland	Switzerland
Irish Free State	United Kingdom
Italy	Yugoslavia
Oceania	
American Samoa	New Hebrides
Australia	New Zealand
British Solomon Islands	Cook and Niue Islands
Fiji	Tonga
Gilbert and Ellice Islands	Western Samoa

Table S2: List of agricultural indicators grouped by thematic category

1. Holding and tenure	
Number of holdings	
Area of holdings	
Holdings fully owned (area of holdings, number of holdings)	
Holdings rented from others (area of holdings, number of holdings)	
Holdings operated under mixed forms of tenure (area of holdings, number of holdings)	
Holdings not owned (area of holdings, number of holdings)	
Number of farms and their distribution according to size	
Area of farms and their distribution according to size	
Number of agricultural and forest holdings	
2. Land utilization (area)	
Total area	
Cropland	
Arable land	
Land for growing trees, vines and shrubs	
Permanent meadow and pasture	
Wood or forest land	
All other land	
3. Crops (area, production, number of holdings reporting)	
Wheat	Sweet potatoes
Winter Wheat	Yams
Spring Wheat	Sugar Cane
Rye	Sugar Beets
Rice	Cotton
Millet and Sorghum	Flax
Millet	Hemp

Table S2: List of agricultural indicators grouped by thematic category (continued)

Sorghum	Groundnuts
Maize	Linseed
Barley	Hempseed
Oats	Castor beans
Spelt	Rapeseed
Maslin	Colza
Other mixed grains	Sesame
Soybean	Sunflower
All dry beans and peas	Tobacco
Edible dry beans	Coffee
Lentils	Tea
Chickpeas	Cacao
Edible dry peas	Coconut
Potatoes	Oil Palms
Manioc	Rubber
Arrowroot	

4. Livestock and poultry (number of heads, number of holdings reporting)

Horses	Cattle
Sheep	Goats
Pigs	Poultry
Buffaloes	Camels

Table S2: List of agricultural indicators grouped by thematic category (continued)

5. Employment in agriculture (all persons, male, female)	
Holders and members of their families	
Holders operating their own holding	
Holders not operating their own holding	
Family members permanently employed	
Family members not permanently employed	
Persons working for pay on the holding	
Employed temporarily	
6. Farm population (all persons, male, female))	
All persons	
Holders and members of their households	
Other persons living on the holding	
Farm population by main occupation	
7. Agricultural technology (number of machines)	
Tractors	Mowers
Plows	Rakes
Iron plows	Reapers
Disk plows	Binders
Wood plows	Combines (harvest-threshers)
Ridging plows	Corn pickers
Tine harrows	Potato-harvesting machinery
Rotary tillers	Sugar-beet harvesting machinery
Disk harrows	Threshers
Cultivators	Hay balers
Hoes	Sugarcane crushers
Seed drills	Carts

Table S2: List of agricultural indicators grouped by thematic category (continued)

Sprayers	Jeeps
Dusters	Station wagons
Rollers	Trucks
Fertilizer distributors	Automobiles
Grain harvesters	Ploughs
Potato lifters	Tedders
Cleaners and sorters	Maize shredders
Hay and forage presses	Chaffcutters
Rootcutters	Grinders
Shedders	

8. Irrigation and drainage (area, number of holdings reporting)

Land irrigated

Area of land drained

9. Fertilizers and soil dressings (area, number of holdings)

Artificial fertilizers

Nitrogenous fertilizers

Phosphate fertilizers

Potash fertilizers

Natural fertilizers

Other fertilizers and fertilizer compounds

Improvements

Organic fertilizers

Inorganic fertilizers

Mixed fertilizers

Table S3: Holdings, total area, and average size per holding

Country	**Year**	**Holdings**	**Area (ha)**	**Avg area (ha/holding)**	**Δ 1930–1960 (%)**
Austria	1930	433 560	7 630 000	18.5	
Austria	1950	432 848	7 726 228	17.8	
Austria	1960	396 530	7 683 888	19.4	+5
Sweden	1932	669 751	20 593 000	30.8	
Sweden	1951	674 624	16 609 642	24.6	
Sweden	1961	264 580	3 866 484	14.6	-52
Denmark	1929	205 991	3 247 000	15.8	
Denmark	1949–50	206 635	3 597 922	17.4	
Denmark	1959–60	196 506	3 108 267	15.8	0
Germany	1933	3 046 226	42 121 000	13.8	
Germany	1949–50	2 011 992	21 979 025	10.9	
Germany	1960	1 761 114	21 369 649	12.1	-12
Norway	1932	299 360	1 713 000	5.7	
Norway	1948-49	349 528	7 052 895	20.2	
Norway	1959	433 920	7 622 744	17.6	+208
Netherlands	1930	372 081	2 151 000	5.8	
Netherlands	1950	282 119,	2 314 424	8.2	
Netherlands	1959–60	300 702	2 658 297	8.8	+53
Belgium	1929	1 131 146	1 998 000	1.7	
Belgium	1950	280 015	1 726 865	6.2	
Belgium	1959	269 069	1 660 831	6.2	+254
Finland	1929	287 171	15 365 000	53.5	
Finland	1950	465 655	15 534 357	33.4	
Finland	1959–60	387 962	15 959 621	41.1	-23
United States	1929	6 208 648	403 680 000	65.1	
United States	1950–51	5 382 162	468 848 429	87.1	
United States	1961	3 707 973	454 608 633	122.6	+88
Canada	1931	728 623	66 027 000	90.7	
Canada	1950–51	623 091	70 433 200	113.0	
Canada	1959	480 903	69 827 959	145.2	+60
Guam	1930	2 104	2 464	1.2	
Guam	1949–50	2 262	10 025	4.43	

Table S3: Holdings, total area, and average size per holding (continued)

Country	Year	Holdings	Area (ha)	Avg area (ha/holding)	Δ 1930–1960 (%)
Guam	1961	2 028	12 994	6.4	+448
Uruguay	1929	37 306	2 050 000	54.9	
Uruguay	1950–51	85 258	16 973 632	199.1	
Uruguay	1961	86 928	16 988 408	195.4	+256
Puerto Rico	1930	52 965	802 000	15.1	
Puerto Rico	1950	53 515	725 086	13.5	
Puerto Rico	1959	45 792	661 245	14.4	-5
Virgin Islands	1930	329	27 700	84.0	
Virgin Islands	1950	755	25 800	34.2	
Virgin Islands	1959–60	501	17 831	35.6	-58
Australia	1928-29	201 225	58 500 000	290.7	
Australia	1950	245 267	375 788 373	1532.0	
Australia	1960	252 243	464 575 646	1842.0	+540
New Zealand	1929–30	85 167	17 563 000	206.3	
New Zealand	1949–50	90 290	17 465 309	193.4	
New Zealand	1960	76 928	17 813 567	231	+12
American Samoa	1929	815	392	0.5	
American Samoa	1949-50	1 490	1 490	1.0	
American Samoa	1959-60	2 135	4 662	2.2	+340
Kenya (Eur+Asian)	1930	2 836	2 081 000	734.0	
Kenya (Eur+Asian)	1954	3 163	2 838 300	898.0	
Kenya (Eur+Asian)	1960	3 609	312 8547	866.0	+18
Kenya (African)	1960	734300	2979659	4.1	–
Japan	1929	5 575 583	5 037 000	0.9	
Japan	1949	6 189 700	10502618	1.7	
Japan	1959–60	6 056 534	7 141 941	1.2	+33

Supplementary Figures

Agricultural indicators coverage by region and category (1930-1960)

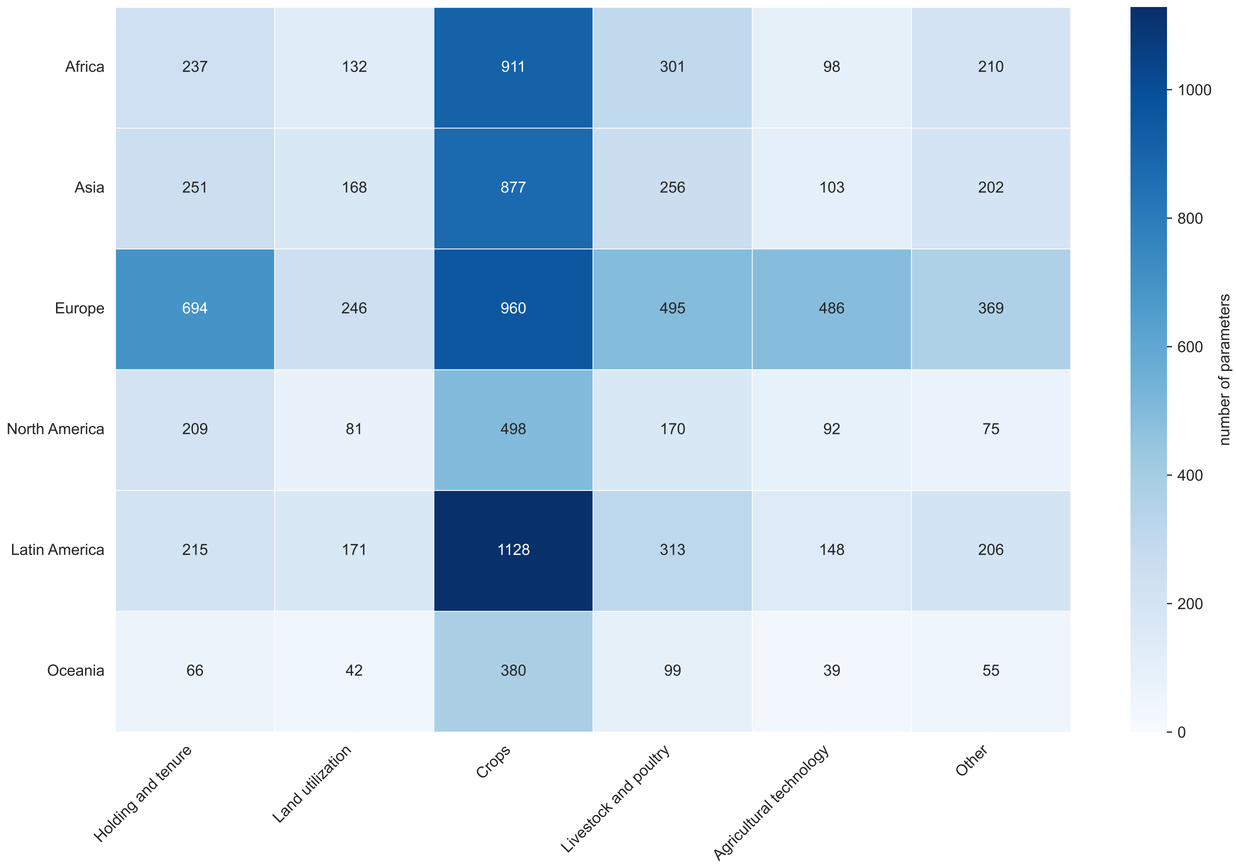

Figure S1: **Agricultural indicators coverage by region and category (1930-1960).** Shown is the number of available agricultural indicators for each region across six key categories: holdings and tenure, land utilization, crops, livestock and poultry, agricultural machinery, other indicators. The color intensity reflects the degree of coverage: the darker the cell, the more indicators were reported.

Supplementary Materials

S1 Prompt design

The prompt was designed following best practices in prompt engineering [32–34] and consists of two main parts: (1) the system prompt sets a role and instructs to use only the provided context. Moreover, it provides comprehensive guidelines on the expected output format, specifying a markdown table. This combination is important: the role aligns the model's behaviour with a narrow data-assistant task, and the context-only rule reduces hallucinations by prohibiting the use of outside knowledge. (2) The user prompt specifies the target agricultural parameter, country, and relevant year(s). It also contains the top-k most relevant report chunks retrieved by the hybrid search method. The full prompt is provided in Supplementary Fig. S2.

System prompt

You are a data assistant for agricultural census reports. Use only the context below to answer the question.

If the question includes multiple countries or multiple indicators, return all available values for each country separately.

If the question asks about data across multiple years, note that different years may have different parameters or indicators available. Include all available data for each year, even if the parameters differ between years.

Please format your answer as a markdown table with the following columns:

| Country | Year | Indicator | Value | Unit |

Group all indicators by country, do not mix countries and do not create extra combinations.

User prompt

{context}

Requested data point: "What are the {indicator} data for {country} across {years}?"

Answer:

Figure S2: **Prompt design.** (1) The system prompt section describes the role, rules for extracting and formatting data. (2) The user prompt provides the top-k retrieved report chunks (context), and an example of a user query with an indicator, country, and year.