# BCRP Corpus: An AI-Assisted Research and Information Retrieval System for Central Banking

Gonzalo Bueno Bustíos

21 November 2026

**Abstract**

Central banks accumulate economic knowledge across working papers, policy reports and other publications. Accessing this evidence is difficult when documents span decades, languages and formats. I document BCRP Corpus, an independent prototype for searching public publications of the Banco Central de Reserva del Perú (BCRP) and retrieving passages linked to original PDF pages and report editions. The 11 November 2026 snapshot contains 1,441 indexed document records and 96,695 chunks, including 474 working papers, 788 articles from *Revista Moneda*, 89 Spanish-language Inflation Reports and 90 separately indexed report boxes. A further 465 boxes have searchable catalogue metadata and can be accessed through their parent reports. In a traced bilingual search on exchange-rate pass-through, I recover studies showing attenuation from import to consumer prices and use the original pages to distinguish a rolling long-run estimate from a fixed-horizon response to an externally instrumented shock. The exploratory evaluations document numerical transcription errors, ambiguous evidence and limits of historical reconstruction.

**Keywords:** central banking, institutional knowledge, information retrieval, economic research, retrieval-augmented generation.
**JEL:** C88, D83, E58.

## Resumen

Los bancos centrales acumulan conocimiento económico en documentos de trabajo, reportes de política y otras publicaciones. Consultar esta evidencia puede resultar difícil cuando abarca varias décadas, idiomas y formatos. Presento BCRP Corpus, un prototipo independiente para buscar publicaciones públicas del Banco Central de Reserva del Perú y recuperar pasajes vinculados a las páginas de los PDF originales y a ediciones específicas de los reportes. La versión del 11 de noviembre de 2026 contiene 1 441 registros documentales indexados y 96 695 fragmentos de texto, entre ellos 474 documentos de trabajo, 788 artículos de *Revista Moneda*, 89 reportes de inflación en español y 90 recuadros indexados por separado. Otros 465 recuadros cuentan con metadatos consultables y pueden encontrarse en sus reportes originales. En una búsqueda bilingüe sobre el traspaso cambiario, recupero estudios que muestran una atenuación de la transmisión desde los precios de importación hasta los precios al consumidor. La revisión de los PDF originales permite distinguir una estimación de largo plazo basada en ventanas móviles de una respuesta a horizonte fijo ante un choque identificado mediante instrumentos externos. Las evaluaciones exploratorias documentan errores de transcripción numérica, evidencia ambigua y limitaciones para reconstruir la información disponible en fechas pasadas.

# 1 Introduction

Central banks publish economic research to deepen understanding of their economies and explain policy decisions. The value of these publications extends beyond their release dates and context: an earlier estimate of exchange-rate pass-through, an analysis of a financial crisis or a technical explanation of a forecast may be relevant to a new question many years afterwards. Yet locating the appropriate evidence can require familiarity with the terminology, publication series and date of the original work. As the stock of institutional knowledge grows and staff change over time, relying on researchers' recollection of past publications becomes increasingly difficult. General-purpose web search can help locate these materials, but it ranks institutional publications alongside other sources. Its results need not prioritize the relevance, quality or historical context of economic evidence.

Generative artificial intelligence (AI) can help researchers navigate this information, but fluent answers alone are insufficient for economic research. Language models may generate unsupported or incorrect claims, and retrieval does not by itself ensure that the cited evidence supports an answer. Moreover, the interpretation of economic evidence depends on its context. A forecast should not be confused with a realized outcome. It reflects the information and assumptions available when it was produced. A simulation does not by itself establish a causal effect, and a coefficient must be understood in relation to its shock, sample and units. An answer that omits these distinctions can misrepresent a finding or obscure its economic significance, even when it cites the correct paper. This motivates a workflow in which the primary output of search is inspectable evidence rather than an authoritative conclusion.

I document BCRP Corpus, an independent research prototype built using public BCRP publications. Its objective is to make published institutional knowledge accessible through a common search interface, while preserving the original PDF and the edition or information scope underlying each claim. The paper documents the inventory, the workflow that links retrieved passages to source pages and editions, and the feasibility of using that workflow for economic research. The evaluations describe performance on selected tasks and identify failure modes.

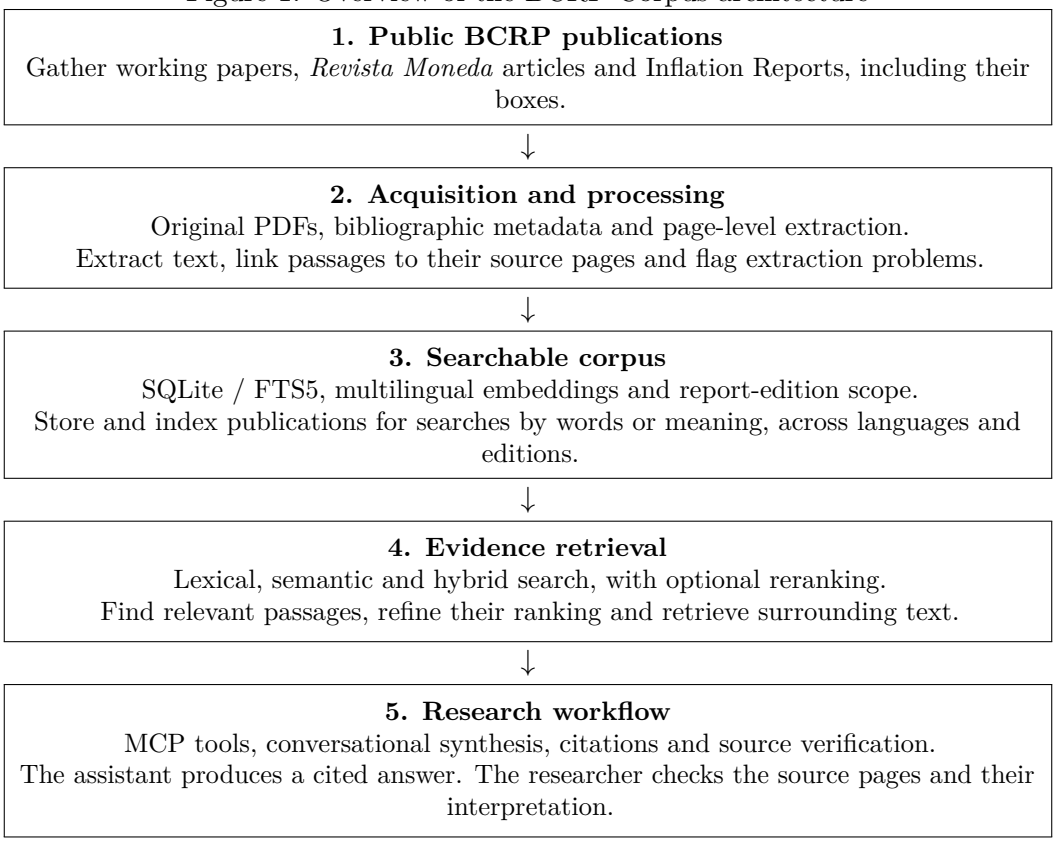
The design separates the preparation of the documentary collection from its use in answering research questions. As Figure 1 illustrates, I first web-scrape the public BCRP publications and download their original PDFs. I extract text from each document, record its bibliographic information and divide the text into short passages linked to the pages from which they were obtained. The resulting collection, or corpus, can be searched without opening every document individually. The original files remain accessible because the extracted text may contain errors, particularly in tables and equations.

A researcher interacts with BCRP Corpus through a conversational assistant connected to its local Model Context Protocol (MCP) server. The researcher asks a question in ordinary language and the assistant chooses a search tool and sends the query to the corpus rather than relying solely on its own learned information. The search layer returns passages with document identifiers, publication metadata and physical PDF page numbers. A lexical search ranks passages containing the query's words or expressions. A semantic search ranks passages by the similarity of their meaning to the question, including across Spanish and English. Hybrid search combines the two rankings. When enabled, a second-stage reranker reorders the leading passages according to their relevance to the specific question. These rankings help locate evidence, they do not establish the truth of the retrieved claims.

The assistant can follow an initial search with requests for neighboring passages or the original PDF page. It then uses the retrieved material to prepare a cited response. For example, a question in English about the effect of an exchange-rate depreciation on consumer prices can

retrieve a Spanish-language study of *traspaso cambiario* without the researcher knowing its title. A similar case can be made for a researcher looking for a specific concept, and the assistant seraches for many variations. Filters can restrict the search to a publication series or a particular Inflation Report edition. Because both extraction and interpretation can be imperfect, the researcher can inspect the original source before relying on a numerical estimate or causal conclusion.

Figure 1: Overview of the BCRP Corpus architecture




## 2 Related work and institutional motivation

Generative AI expands the set of tasks that economists can delegate to computational tools. Beyond writing and programming, large language models (LLMs) can assist in developing research ideas, reviewing literature, interpreting datasets and working through mathematical derivations. These applications can reduce the effort involved in individual research tasks, although their usefulness depends on the task and the researcher's ability to assess the output (Korinek, 2023). Background research is particularly relevant here: before estimating a model or assessing a policy, economists often need to establish what is already known about an economic mechanism and under what conditions previous findings were obtained.

For central banks, the practical issue is how to use AI without losing control over the evidence supporting policy analysis. This requires safeguards for sensitive information, checks on unreliable or biased outputs, and clear responsibility for the resulting decisions (de Araujo et al., 2024; Consultative Group on Risk Management, 2025).

One obstacle to using general-purpose LLMs for background research is that the knowledge acquired during training is stored implicitly in the model's parameters. The model may not

know a particular publication, its knowledge may be outdated, and an answer need not identify the original evidence. Retrieval-augmented generation (RAG) addresses this limitation by providing the model with relevant passages from an external collection at the time a question is asked (Lewis et al., 2020). In a typical workflow, the system first searches an index of the collection, selects passages related to the question and supplies them as context for the language model. The collection can be updated without retraining the language model, and the retrieved passages can be inspected independently of the generated answer.

However, retrieving evidence is not equivalent to interpreting it correctly. The search may omit the passage that establishes a result or retrieve a related but insufficient source. Even when the correct passage is available, the model may overlook a qualification, confuse two numbers or attach a citation to a claim that the source does not establish. Evidence from citation-generation benchmarks shows that answer correctness and citation support are distinct dimensions of performance: even apparently good answers can lack complete documentary support (Gao et al., 2023). Financial question answering presents a similar challenge. Evaluations based on company disclosures find that language models can produce incorrect answers or fail to answer even when retrieval is available (Islam et al., 2023). Although these exercises concern financial rather than macroeconomic documents, they highlight a common problem: answering questions that require locating a specific numerical result and interpreting it in its original context.

Central-bank applications already illustrate different ways of adapting these methods to institutional information. Domain-specific language models trained on central-bank publications can improve the interpretation of monetary policy terminology and the classification of policy stances (Gambacorta et al., 2024). Retrieval-based systems can also make official statistics more accessible by helping users identify relevant series in the ECB Data Portal and interact with economic datasets through natural-language queries (Petracca et al., 2026; Seah et al., 2026). More recent work extends AI assistance to the econometric workflow itself, including the selection of model specifications, data harmonization, estimation and diagnostics, while retaining economist control over consequential modelling decisions (Lombardi et al., 2026).

BCRP publications illustrate how computational methods support different parts of economic research. Text analysis has been used to study the readability of Peru's monetary policy releases (Lahura et al., 2023), and generative language models to classify the monetary policy tone of Federal Reserve communications in a *Revista Moneda* article (Woong et al., 2026). For Peruvian inflation forecasting, tree-structured search has been used to select model architectures and preprocessing choices, and institutional metadata to guide the selection of predictors (Garcia and Vega, 2024, 2026). I focus on an earlier step in the research process: finding and checking evidence in the publications themselves.

A particularly close institutional precedent is the development of offline assistants at the Central Bank of Bosnia and Herzegovina, where semantic document search and an HR chatbot were built on a common architecture under confidentiality and data-sovereignty constraints (Hajdarevic and Selimovic, 2026). I address a complementary task: discovering and verifying published economic evidence before a researcher develops a specification, performs an estimation or prepares a policy assessment. The emphasis here is on a multilingual public corpus, page-level provenance, edition-specific retrieval and the interpretation of historical economic evidence. The diagnostic evaluations distinguish the relevance of retrieved passages from the correctness and documentary support of the resulting answers.

# 3 Construction of the corpus

## 3.1 Coverage and source inventory

The prototype uses three official online collections: the BCRP's working papers, *Revista Moneda* and the Inflation Report archive. Table 1 reports the frozen 11 November 2026 snapshot. Coverage refers to the captured publisher inventories. The Moneda series starts at issue 131 because this was the earliest issue returned by the research portal capture. The legacy listing could not be reconciled, so earlier issues and any omissions from the captured listings remain coverage gaps.

Table 1: Composition of the corpus, 11 November 2026

| Collection | Indexed | Metadata only | Excluded |
|---|---|---|---|
| Working papers (1997–2026) | 474 | 0 | 0 |
| *Revista Moneda* (issues 131–207) | 788 | 0 | 1 |
| Inflation Reports (Spanish) | 89 | 0 | 0 |
| Inflation Report boxes | 90 | 465 | 0 |
| Total | 1,441 | 465 | 1 |

*Source: BCRP Corpus inventory, 11 November 2026. Report and box records may refer to the same physical PDF pages.*

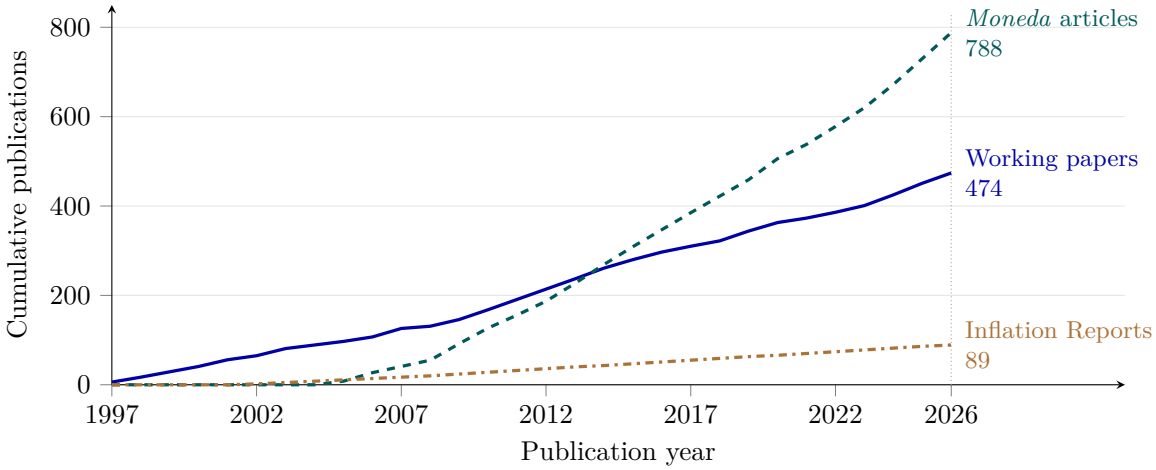
One record from *Revista Moneda* (issue 144) was excluded because it contains a museum donation acknowledgement rather than economic research or policy analysis. The original PDF is retained.

All 89 indexed Spanish Inflation Reports match the editions in the captured archive, from June 2002 through September 2026. The archive lists two in 2002, three per year in 2003–2008, four per year in 2009–2013, three in 2014 and four in 2015. The 2014 published schedule placed the fourth release in January 2015. There are four per full year from 2016 through 2025 except in 2020, when the BCRP annual report records three releases during the pandemic. The partial 2026 inventory contains three. English editions are outside the indexed scope, and the completeness of earlier archival holdings remains unverified.

The database has 30,389 page records and 96,695 searchable chunks. Of 555 catalogued boxes, 90 own separately searchable passages and 465 retain metadata with body text under the parent report. The importer assigns reviewed source elements to one owner and leaves uncertain boundaries with the parent. Standalone PDFs serve as alternative sources. All 348 page records for indexed boxes also have parent-page records. Several distinct passages can come from one physical page, and parent context can display box text again. Search therefore avoids duplicate ownership of mapped source elements, while ranking and optional reranking do not establish independence across passages, documents or alternative editions.

Figure 2 shows how the indexed publications accumulated by publication year. Each article or working paper counts once, and each Spanish Inflation Report counts once.

Figure 2: Cumulative indexed publications by year, 1997–2026




*Source: Frozen inventory, 11 November 2026. The 2026 year is partial. Report boxes and the excluded* Moneda *record are outside the counts.*

### 3.2 Research topic prevalence

To examine whether the subjects covered by BCRP research change over time, I measure the prevalence of ten economic topics among newly published Working Papers. This provides a descriptive view of changing publication emphasis and illustrates how the corpus allows for older work to remain accessible even when few recent papers address the same topic. The exercise is confined to the 474 indexed Working Papers dated 1997–2026. Each paper is assigned to its publication year using the catalogue record. I retain *Revista Moneda* and Inflation Reports outside this figure because their editorial purpose, publication frequency and archival coverage differ from those of the working-paper series.

I classify each paper by searching its title and abstract for predefined English and Spanish expressions associated with each topic. The dictionary includes specific variants of relevant economic terms rather than treating every occurrence of a broad word as a match. For example, exchange-rate pass-through is distinguished from the transmission of policy interest rates, and the financial-crisis/liquidity category requires expressions that identify the intended financial context. These rules were refined after inspecting sampled matches and plausible omissions. Searching titles and abstracts limits incidental matches in bibliographies, although it can miss papers that examine a topic only in the body. The COVID-19 series is defined separately from unconventional monetary policy. A paper may match more than one topic. Appendix C documents the matching rules and reproducibility files.

For topic $k$ and publication year $t$, the annual prevalence in Figure 3 is

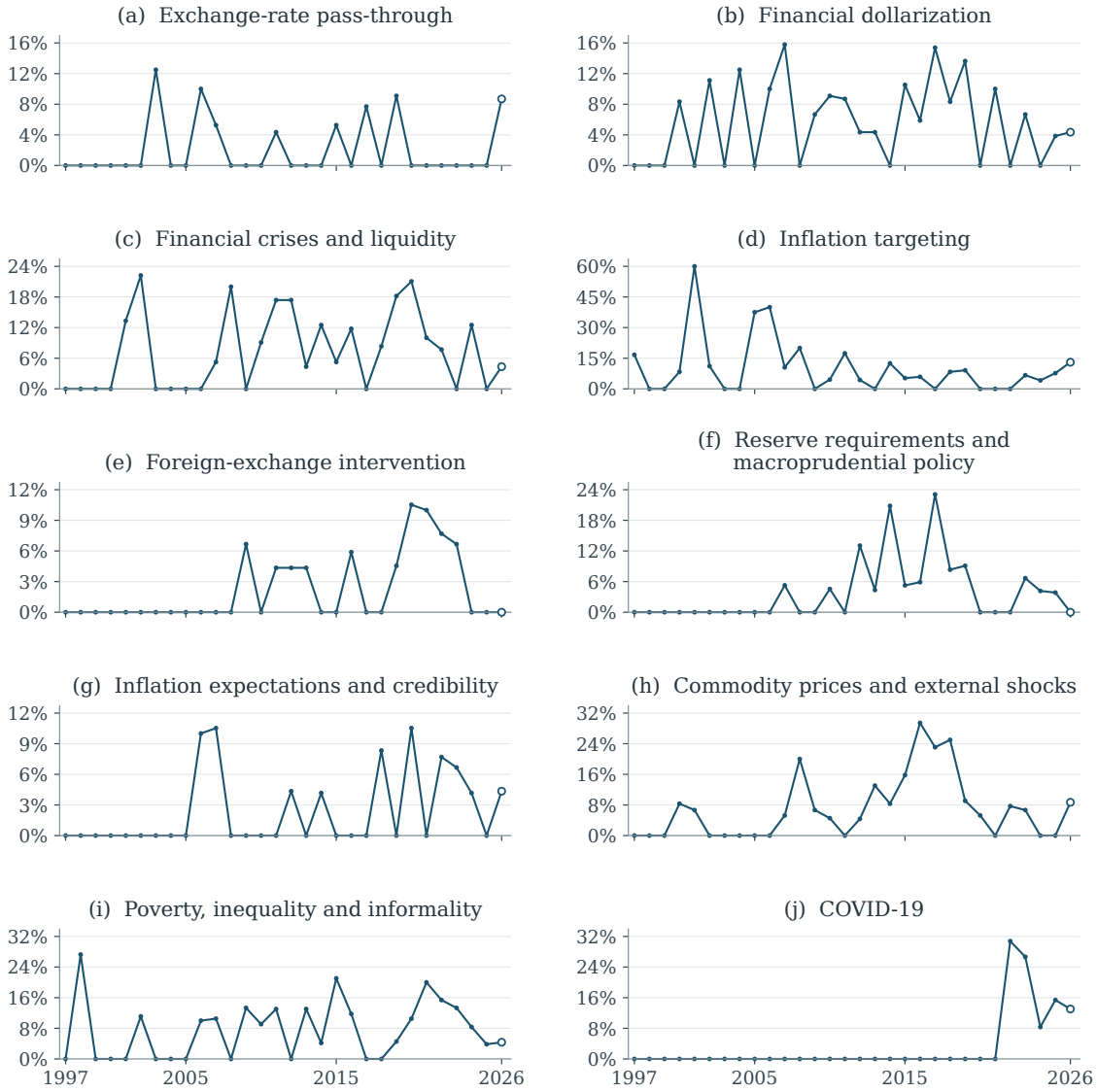
$$P_{k,t} = 100\frac{N_{k,t}}{N_t},$$

where $N_{k,t}$ counts distinct indexed Working Paper IDs from year $t$ that match topic $k$, and $N_t$ is the total number of indexed Working Papers from that year. Multiple mentions within a paper count only once.

The clearest long-run change concerns inflation targeting. It appears in 19 of 107 Working Papers in 1997–2006 (17.8%), 14 of 190 in 2007–2016 (7.4%) and seven of 154 in 2017–2025 (4.5%). Its pronounced early peak occurs in 2001, before the adoption of inflation targeting in 2002. Among the 2017–2025 papers, the most frequent matches in this ten-topic classification

are financial crises and liquidity (14), COVID-19 (14), poverty, inequality and informality (12), and commodity prices and external shocks (11). Foreign-exchange intervention and inflation expectations each match six papers over the same period. Financial-crisis and liquidity matches increase around 2019–2020, while COVID-related Working Paper matches appear from 2022 onward. A decline in new publications on a subject does not diminish the availability of earlier research. For example, the 1997 inflation-targeting study by Perea and Soto (`dt-1997-001`) remains identifiable in the indexed collection. The corpus preserves a route to such documents and their original pages.

Figure 3: Research topic prevalence in BCRP Working Papers




*Notes:* Each panel shows the percentage of indexed Working Papers published in that year whose title or abstract matches the corresponding bilingual keyword rules. Panel-specific vertical scales apply. Comparisons of plotted heights across panels are not meaningful. Open circles identify the incomplete 2026 observation (23 indexed papers). *Source:* BCRP Corpus, frozen 11 November 2026 snapshot.

### 3.3 Acquisition, conversion and quality control

The acquisition stage collects the publications listed in the BCRP's online catalogues and downloads the corresponding PDFs. It records each document's title, source URL, page count and SHA-256 hash, a digital fingerprint used to check whether the PDF file has changed. Then conversion is done by extracting text and page information from each PDF so that it can be searched. The PDF is kept because an extracted table or number may need to be checked against the original.

Different collections need different conversion methods. For Working Papers, Poppler, a set of PDF-reading utilities, extracts text already stored in the file. Tesseract performs optical character recognition (OCR) on selected pages where text must instead be read from an image. It was used on 1,801 Working Paper pages across 296 papers. *Revista Moneda* uses Docling, which also tries to recognize columns, tables and reading order. OCR was selectively enabled for one indexed article, accounting for three page records. Another article was excluded after OCR still failed to produce usable text. Inflation Reports and boxes were processed with Docling, with separate Poppler text records for comparison. OCR was switched off for the 636 report and box conversion receipts. That setting does not establish that every page was extracted correctly, particularly where text appears in images.

Every extracted passage carries a document ID and a physical PDF page number, counting the first page of the PDF as page 1 even if it is a cover. This may differ from the page number printed within the document. Conversion receipts also retain warnings, including pages with unusually little text, possible scan-like content and replacement characters that indicate unresolved encoding. The broad `needs_review` status is a request for verification, not a measured transcription error. Manually checking every page in a corpus of this size would be costly. A more useful quality-control strategy combines automated diagnostics with targeted visual checks of pages most likely to affect retrieval or numerical interpretation.

Table 2: Selected metadata and searchable-text limitations

| Issue | Recorded evidence | Interpretation and possible follow-up |
|---|---|---|
| Missing JEL metadata | 204 Working Papers lack JEL codes in the captured portal metadata | Topic classification by JEL is incomplete. Candidate codes require source verification. Missing codes should not be inferred from silence. |
| Sparse or scan-like text | 77 *Moneda* page records and 315 Inflation Report page records | Flags identify candidates for inspection, not confirmed missing text. Check source-page images and apply OCR selectively where needed. |
| Replacement characters | 42 Inflation Report page records. Seven box views | Encoding or extraction may have damaged some characters. Compare affected passages with the original pages before citing them. |
| No directly indexed passages | Nine report-page records. Three box-page views | Search may not return text under that particular page record. Check whether its content is indexed under the parent report or a separate box before classifying it as a gap. |

*Source: Frozen corpus diagnostics. Counts refer to recorded warnings or metadata gaps, not independently measured transcription error rates. Categories can overlap.*

Automated checks help decide which pages merit inspection. A page is flagged as *sparse* when very little text was extracted. This may be harmless, as on a cover or a largely graphical page, or it may indicate a scanned page whose words were missed. A *replacement character* is a placeholder inserted when text could not be decoded properly. Table 2 also records missing

JEL classifications and page records without searchable passages. These are possible problems to investigate, not confirmed errors on every flagged page.

A missing search passage does not always mean that the information is missing from the corpus. Inflation Report boxes can be indexed separately from their parent report. For example, the June 2002 box about exchange-rate regimes is searchable as `ri-2002-06-es-box-02`, on physical PDF page 8, although the parent report has no separate search passage for that page. The other pages without passages need similar checks before they can be called retrieval gaps. Tables, figures and equations also remain possible sources of extraction errors, which is why the original PDF is retained.

# 4 Retrieval architecture and evidence workflow

The system stores document information and extracted text in SQLite, a database held in a local file. A text index makes word searches fast. A separate index represents each passage as a vector of numbers, called an *embedding*, so that passages can also be found by meaning rather than exact wording. Python functions expose these searches, neighboring passages, document sources and PDF pages to a conversational assistant through the Model Context Protocol (MCP). In the documented demonstration, the researcher interacts with Claude, which calls the corpus tools and receives the retrieved evidence. There is no separate BCRP Corpus chat interface in that workflow.

## 4.1 Search methods

A search returns a ranked list of passages, much like the results of a web search. *Lexical search* favors passages containing the words in the query. It uses SQLite's full-text search (FTS5) and the BM25 ranking method, which weighs term matches by how informative they are and gives more weight to titles than to body text (Robertson and Zaragoza, 2009). *Semantic search* can find related passages even when they use different words or a different language. It uses the multilingual E5-small model to convert questions and passages into embeddings, then compares their direction using cosine similarity (Wang et al., 2024).

*Hybrid search* combines the lexical and semantic result lists using reciprocal rank fusion (Cormack et al., 2009):

$$s(d) = \sum_{m:\, d \in L_m} \frac{1}{60 + r_m(d)}.$$

Here $d$ denotes a retrieved passage, $L_m$ is the ordered list produced by search method $m$, and $r_m(d)$ is that passage's position in the list. The score $s(d)$ increases when a passage ranks highly in one or both lists. The constant 60 moderates the influence of the highest-ranked results. As an optional final step, the BGE reranker compares the question with the leading candidate passages and reorders them (Beijing Academy of Artificial Intelligence, 2024). This can improve their order, but cannot find a passage that the first search did not retrieve. Appendix A gives the exact model and parameter settings.

## 4.2 From a question to inspectable evidence

In the documented workflow, a researcher asks Claude a question and Claude uses the corpus tools to locate publications, inspect passages and retrieve original PDF pages. When the question names an Inflation Report, the system first identifies the relevant edition. For a question about what was known by an earlier date, the strict historical search requires evidence

that the edition had been released and that the stored PDF was already available by that date. Because many historical PDFs were downloaded only in 2026, this restriction may prevent an answer even when the report's original publication date is known.

The researcher can specify a report edition or search across the available collections. Important numerical or causal claims should then be checked against the PDF, including their sample, units, time horizon and identification assumptions. If the search does not find a requested estimate, the answer should state what was searched rather than conclude that the estimate does not exist.

# 5 Illustrative applications

These examples combine a documented corpus search with cases drawn from earlier project audits. They illustrate research tasks the prototype can support. Page references are physical PDF pages. I checked six selected factual reference items against the original PDFs, including the pass-through, GDP, liquidity and El Niño examples below. The checks cover those specific claims.

## 5.1 Empirical estimates: Exchange-rate pass-through

I asked how Peruvian exchange-rate pass-through has evolved across prices, horizons and directions of exchange-rate movement. The Full Corpus returned passages from four BCRP studies using queries in English and Spanish. Figure 4 shows selected results and two estimates checked against their original PDF pages. The English semantic query returned five passages from papers in each language. A longer Spanish hybrid query returned ten passages from Spanish-language papers, all through semantic search. A shorter Spanish query returned results through both lexical and semantic search. Follow-up queries recovered the 2011 study and the 2026 analysis of asymmetry between depreciation and appreciation. This example demonstrates cross-language retrieval for one question, not a general performance advantage.

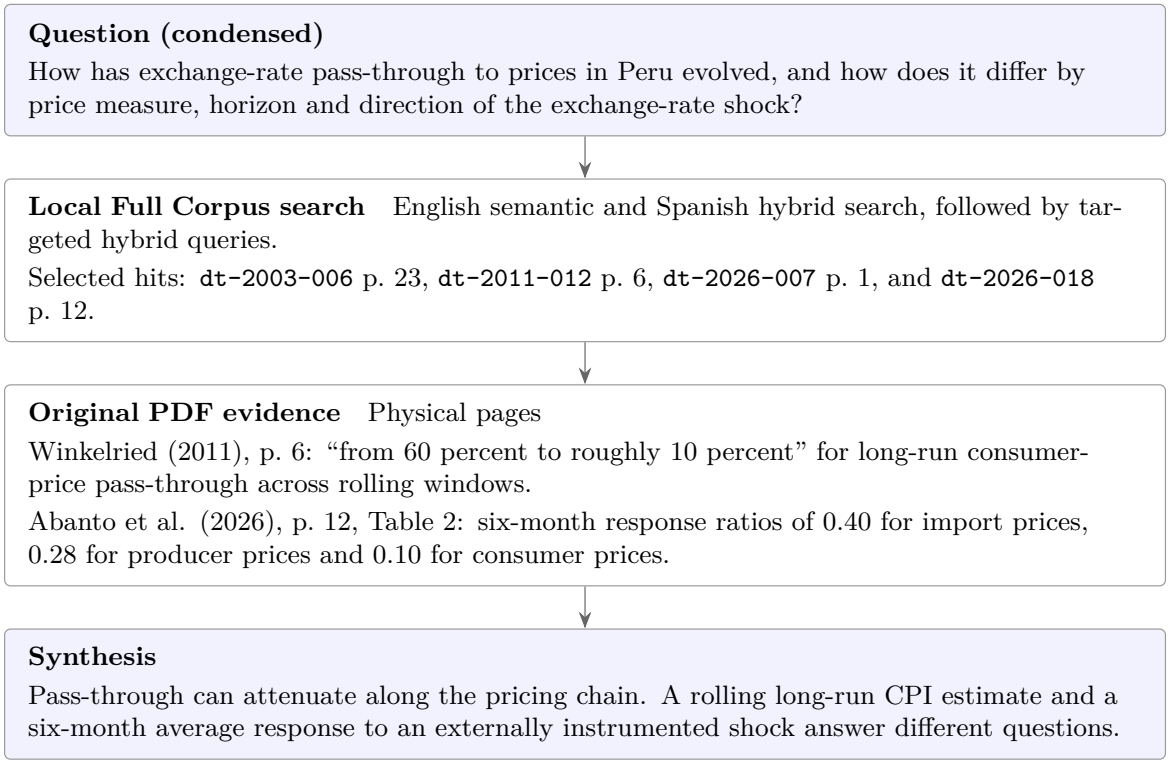
The early literature distinguished state dependence from average pass-through. Winkelried (2003) evaluates price responses to depreciation in a smooth-transition VAR. Pass-through varies with the strength of the depreciation and the economic state. A threshold at zero intended to separate depreciations from appreciations fit poorly, so the exchange-rate regimes in the reported model describe high and low depreciation (physical PDF pp. 9, 12 and 20). Winkelried (2011) instead estimates a recursively identified structural VAR on rolling windows of 72 monthly observations. The long-run consumer-price response relative to the exchange-rate response fell from about 60 percent to about 10 percent by the final window ending in April 2011 (physical PDF pp. 4–6). This is evidence of a decline in that specification. Its timing is consistent with the transition to low inflation, and the design does not isolate the effect of adopting inflation targeting.

The 2026 studies separate two sources of variation. Pérez Forero (2026) uses a common Cholesky ordering across linear and nonlinear Bayesian VAR specifications. In the censored model, the one-year CPI response ratio is about 0.23 after a depreciation shock, while the corresponding appreciation effect cannot be distinguished from zero (physical PDF pp. 10–12). Abanto et al. (2026) uses DXY and VIX as external instruments in a time-varying Bayesian VAR with stochastic volatility. Its historical average six-month ratios of cumulative price and exchange-rate responses are 0.40, 0.28 and 0.10 for import, producer and consumer prices, respectively (1996–2024, physical PDF pp. 11–12, Table 2). The interpretation depends on the instruments primarily affecting domestic variables through the exchange rate, an assumption

that can be challenged if global shocks also have direct domestic effects.

Taken together, the studies document changes over time, state dependence and attenuation along the pricing chain. The rolling long-run CPI estimate, direction-specific one-year CPI response and six-month historical averages under external instrumentation reflect distinct shocks, samples and horizons. Differences between their reported estimates therefore cannot be assigned to structural change alone.

Figure 4: From a pass-through question to original PDF evidence




*Source: Selected results from the recorded Full Corpus calls. The question is condensed and the synthesis is interpretive.*

## 5.2 El Niño and cross-publication research

The El Niño example shows why economists may need to search beyond Working Papers. Inflation Report boxes contain shorter applied studies that are relevant to research questions, including the effects of weather shocks on household outcomes. Locating such a study is only the first step. Its data, comparison group and interpretation must also be checked.

Box 3 of the December 2023 Inflation Report (Banco Central de Reserva del Perú, 2023) classifies 68 of the 194 provinces in its geographic tabulation as highly exposed to the 2017 coastal El Niño. The tabulation excludes Datem del Marañón and Putumayo, where the source lacked sufficient information. The criterion is a three-month standardized precipitation index of at least two in any month of the first half of 2017. The separate poverty analysis uses 16,162 household-year observations from the 2016–2018 ENAHO panel. The exposure indicator compares households in highly exposed provinces in 2017–2018 with households in less-exposed provinces in those years and with the pre-event year 2016. The Probit model, which controls for prior poverty and other covariates, reports a conditional marginal association of about 1.4 percentage points in the probability of poverty. See physical pp. 1–4 of the standalone box

(printed pp. 71–74). The study does not identify a causal effect through this comparison.

### 5.3 When the system should not supply a number

Two errors in the November 2 answer audit concerned foreign-exchange intervention. One answer assigned a coefficient to intervention when the equation and table labels in Arena and Tuesta (1999) (`dt-1999-003`, physical PDF p. 24) were inconsistent. Another repeated a numerical transcription error in a table on physical p. 21, which inspection of the original page resolved. The coefficient on p. 24 remains ambiguous and is omitted here.

The examples motivate a conservative rule: withhold disputed coefficient magnitudes and causal interpretations until the estimand, units and row identity can be established. Page-grounded retrieval makes verification possible but does not replace it.

## 6 Preliminary evaluation

I report two early checks of the search results, followed by a pilot that compares answers produced with different tools. They used different questions and assessment rules, so their results are presented separately. None is a representative test of the system's overall accuracy.

### 6.1 Retrieval tests and earlier answer checks

The search test on 28 October used the version of the corpus that then contained Working Papers and *Revista Moneda*. Inflation Reports and their boxes were added later, so these results do not describe the final collection. Twenty fixed questions or controls were submitted to each of three search methods: keyword, semantic and hybrid search. Eighteen asked about genuine economic topics and two were negative controls with no supporting answer expected. Eight selected cases were also run with an additional reranking step, giving 68 search calls in total. For six ordinary research questions, an unblinded AI reviewer examined the first returned passage and classified it as directly useful, partly useful or not useful as evidence.

Table 3 evaluates only the first passage returned by each search. *Direct* means it provides evidence answering the question, *Partial* means it is relevant but incomplete, and *None* means it provides no useful evidence. *Calls* counts the searches performed. *Median (s)* is the middle search time, measured in seconds. The first three columns refer to six questions, while search times use all 20 calls per basic method and eight with reranking.

Table 3: First-passage usefulness and local retrieval time

| Mode | Direct | Partial | None | Calls | Median (s) |
|---|---|---|---|---|---|
| Keyword | 0 | 0 | 6 | 20 | 0.003 |
| Semantic | 0 | 3 | 3 | 20 | 2.413 |
| Hybrid | 0 | 3 | 3 | 20 | 2.393 |
| Hybrid + reranking | 3 | 2 | 1 | 8 | 31.609 |

*Source: October 2026 experiment. First-passage judgments use six common queries. Timing uses all calls per mode.*

Reranking returned directly useful evidence for three of the six questions, but increased the median search time for those questions from about 2.4 to 30.2 seconds. Across the broader set of 18 substantive queries, keyword, semantic and hybrid search returned six, five and five directly useful first passages, respectively. The negative controls found no supporting evidence

through keyword search, although semantic methods returned related but unhelpful passages. These small, unblinded checks do not establish which method performs best.

A separate test on 2 November examined the answers produced by the assistant, rather than just the search results. Eight research questions were each answered using keyword-only and hybrid search, producing 16 answers. An unblinded AI reviewer checked whether each response addressed the question, supported its claims with the cited pages and acknowledged uncertainty. Four met all the requirements. The other 12 failed at least one requirement, but were not necessarily factually wrong. Two contained serious errors, and 29 selected PDF pages were visually inspected. On 11 November, eight answers to a different set of questions met a different assessment standard. Since the tests were not comparable, this is not evidence of an improvement in accuracy.

### 6.2 Feasibility pilot with recorded tool use

The next pilot, run on 15 November, explored whether giving an AI assistant access to the corpus changed the answers it produced. Five questions from system development covered tasks such as exchange-rate pass-through, finding literature in two languages and answering questions about Inflation Report boxes. Each question was asked of Claude Haiku 5.5 and Sonnet 5.5 with four levels of access: no tools, web search, basic corpus keyword search or the full corpus toolkit. This produced 40 answers, ten per condition (Table 4). The questions were known in advance, so the comparison is exploratory rather than a test on new questions.

Table 4: AI-scored outcomes on the matched development questions

| Tool access | Answers | Correct | Grounded | Critical |
|---|---|---|---|---|
| No tools | 10 | 1 | 0 | 1 |
| Native web | 10 | 7 | 0 | 1 |
| Basic corpus retrieval | 10 | 3 | 1 | 0 |
| Full corpus retrieval | 10 | 9 | 5 | 0 |

*Source: November 15 feasibility pilot. Each outcome is an answer count out of ten per arm. Original AI grades.*

An AI evaluator graded each answer for factual accuracy as incorrect (0), partly correct (1) or correct (2), by comparing it with a prepared reference answer. The *Correct* column counts only scores of 2, regardless of citations. The *Grounded* column applies a separate standard: the answer must be at least partly correct, cover all points required by the reference, support at least 80% of its factual claims with cited evidence, and provide valid citations for at least 90% of them. It must also avoid a critical error, such as inventing a source, reporting a materially wrong number or claiming to have inspected a page that was not inspected. For questions with no answer in the available evidence, an appropriate refusal can pass. These are AI judgments using the pilot's scoring rules.

Table 5: Factual correctness and source-grounded acceptability

| Factual score | Grounded | Did not pass grounding | Total |
|---|---|---|---|
| 2 (correct) | 6 | 14 | 20 |
| 0 or 1 | 0 | 20 | 20 |
| Total | 6 | 34 | 40 |

*Source: Archived AI grades for the same 40 development answers as Table 4.*

The full-corpus condition gave the strongest results in this small pilot: the AI evaluator judged nine of its ten answers factually correct, and five of the ten also supplied sufficiently complete

and verifiable citations. An answer can state the right economic result without showing where the evidence comes from. Across all four conditions, 20 of 40 answers were judged correct, but only six also passed the stricter evidence check (Table 5). Web search produced seven correct answers, yet none met that check, partly because the rules required PDF-page citations rather than ordinary web links. The comparison therefore does not establish that the corpus is better than web search. It shows why a correct answer and a well-supported research answer must be assessed separately.

The recorded model use also limits comparison. All five Sonnet responses made with web access show some Haiku usage in their execution records, so the web condition was not strictly produced by Sonnet alone. A second AI evaluator reviewed 12 selected answers from the wider pilot. The two evaluators agreed on whether the evidence standard was met in all 12 cases, on full factual correctness in nine, and on critical errors in 11. This is agreement between two AI assessments, not independent human validation. I separately checked six reference facts against the original PDFs and assessed four answers without seeing their configurations. Those ten checks are useful for finding disagreements, but too few to certify the full pilot. Appendix B explains them.

The historical tests are reported separately from the 40-answer comparison. Twelve answers were generated under restrictions intended to reflect knowledge available by June 2020, and eight more were generated using the current corpus. The strict historical search refused a report when it could not verify that the stored PDF existed in that form in 2020. The later-access questions also retained some historical-cutoff wording. Neither group provides a clean measure of how the tool would have performed in 2020.

## 7 Operational limitations and institutional implications

Keeping the search index on a local computer does not mean the language model also runs locally. When the assistant uses an external provider, retrieved passages may be sent to that provider. This prototype uses public publications. Institutional deployment would require explicit rules about what data can leave the institution, who can access or retain them, what source documents may be shared, and who reviews the results.

The timing figures measure local search, not the total time a researcher spends reaching a conclusion. The study also does not compare the same PDFs with a general-purpose document assistant or web search under matched conditions. Claims about relative performance or research productivity would require such tests with new questions and independently checked answers.

A possible institutional version would keep the corpus and search services on central-bank servers. The assistant could run on an internal language model or use an external provider under an approved confidentiality and data-processing agreement. Either arrangement would need access controls, limits on retention and model training, and an assessment of what information may be transmitted. Subject to those safeguards, the collection could extend beyond public papers to internal research notes and other authorized records, making more of the institution's accumulated knowledge available to its economists. This remains a potential application, not a deployment or benefit demonstrated in this study.

# 8 Conclusion

BCRP Corpus provides a structured approach to searching several decades of public central-bank publications. The design links retrieved evidence to its original PDF, preserves document and edition scope, and gives researchers a practical means of verifying generated claims. In the pass-through example, the workflow recovered evidence of attenuation from import to consumer prices and exposed why a rolling long-run CPI ratio and a fixed-horizon instrumented response should be interpreted separately. Other cases illustrate numerical transcription errors, source ambiguity and limits to the reconstruction of historical information sets.

The prototype is intended to help economists find and verify published evidence, not to replace their assessment of it. Central banks with comparable public archives could adopt the same approach of linking search results to original pages and editions, although this has not been tested. Further evaluation is needed before drawing conclusions about research productivity or operational performance.

# A Local configuration and evaluation provenance

Table 6 records the settings needed to interpret or reproduce the retrieval exercises. Technical names are retained here because they identify the software and models actually used.

Table 6: Retrieval and evaluation settings

| Component | Implemented configuration or archived run setting |
|---|---|
| Search units | Physical-page passages, keyed by document, page and ordinal. Default text windows contain 240 whitespace-delimited words with 40-word overlap. *Moneda* preserves extracted table blocks. Verified report-box pages use contiguous groups of assigned Docling elements. Lexical search also returns separate metadata hits. |
| Lexical ranking | SQLite FTS5 with `unicode61` and diacritic removal. Title, abstract and body weights 5, 2 and 1. FTS5 fixes $k_1 = 1.2$ and $b = 0.75$ in its BM25 implementation. |
| Dense and hybrid ranking | `multilingual-e5-small`, 384 dimensions, mean pooling and L2 normalization. Title and section accompany passage text. The local worker uses CPU float32 with four threads. The 512-token input limit uses a 40-token overlap when splitting. Each parent passage keeps its best-scoring segment. Hybrid fuses up to 100 candidates per leg with equal weights and $k = 60$. |
| Optional reranking | `bge-reranker-v2-m3` cross-encoder on up to 20 distinct document/page pairs, with 512 tokens per query–window pair, up to 16 windows per page and 64 per call in the current implementation. Keyword-only mode has no reranker. |
| October 28, 2026 | Earlier working-paper and *Moneda* snapshot. Retrieval timed on the local CPU, including initialization where it occurred. Top five requested. The eight reranked calls scored 39–57 windows each. No answer generator was used. |
| November 15, 2026 | Feasibility pilot on the 11 November corpus. Claude Code 2.1.325, using Claude Haiku 5.5 and Sonnet 5.5. Hosted generation hardware and per-run semantic-index hash were not recorded. |
| Documentary example | WP-3 used the 11 November database and index, with Python 3.12.12, local CPU retrieval, and Codex synthesis. No reranking was requested. |

The code, data-processing procedures, model configurations and evaluation records have been preserved to document how the results were obtained. The original BCRP publications are publicly available. However, the research repository, processed database and detailed evaluation records are not currently public and may be shared upon request and authorization.

## B Selected author checks and score reconciliation

I agreed with the six selected reference items. The documented searches in answer D21 did not identify a causal estimate of the effect of illegal mining on the policy rate. They returned only *Moneda* articles. The original PDFs nevertheless contain working-paper mentions of illegal or informal mining in `dt-2017-005` (physical pp. 9, 10 and 22) and `dt-2015-006` (physical p. 10). As I noted in the reference review, physical p. 26 of the latter paper is in its references section. These checks qualify the coverage claim. They leave my original grade intact.

Of the four blinded answers, I judged three correct with all their cited support and one incorrect with citation support inapplicable. The original AI scorer gave one answer its top factual score, passed two through the grounding gate, and flagged no critical errors. Using the top AI factual score as the fully correct label, my judgments matched two of the four. The critical-error flags matched in three cases. I marked D07 as critical because it left the financial-dollarization question unanswered. The archived configuration and startup trace confirm that this baseline deliberately provided no built-in tools or MCP servers. The run ended without a tool call or permission denial. The archived AI grade records a failed answer and no critical factual error.

C17 and C11 answered the same historical question. C17 used an allowlist with a cutoff of 19 June 2020. C11 had access to the full index without that allowlist, but its prompt retained the May 2020 cutoff wording. Both answers used a search filter through 2020. These runs do not provide a controlled comparison of historical retrieval performance.

C17 repeated a figure of 1% from the February 2017 working paper `dt-2017-002` (physical p. 3). The BCRP decision of 6 August 2009 lowered the reference rate from 2.0% to 1.25%. The citation identifies the source of the figure, while the contemporaneous decision establishes the historical rate. The disputed figure appears in the answer's prose, outside the structured claims used for the grounding assessment. The archived AI scorer assigned factual score 1 and passed the grounding gate. My original review marked the answer correct with all citations supported. Both grades remain as recorded. These selected reviews do not estimate scorer reliability, and the pilot aggregate retains the original AI grades.

# C Topic-prevalence classification

The trend analysis uses the 474 indexed Working Papers in the 11 November 2026 snapshot. For each topic, the script searches the English and Spanish titles and abstracts using a fixed list of phrases. SQLite's full-text search treats accents as equivalent and matches the indexed words in each phrase. The examples in Table 7 illustrate the rules. The complete dictionary, rather than these examples, determines the reported counts.

Table 7: Examples from the bilingual topic dictionary

| Topic | Example English and Spanish search phrases |
|---|---|
| Exchange-rate pass-through | *exchange rate pass through*, *traspaso cambiario* |
| Financial dollarization | *financial dollarization*, *dolarizacion financiera* |
| Financial crises and liquidity | *liquidity crisis*, *crisis de liquidez* |
| Inflation targeting | *inflation targeting*, *metas de inflacion* |
| Foreign-exchange intervention | *foreign exchange intervention*, *intervencion cambiaria* |
| Reserve requirements and macroprudential policy | *reserve requirements*, *requerimientos de encaje* |
| Inflation expectations and credibility | *inflation expectations*, *expectativas de inflacion* |
| Commodity prices and external shocks | *commodity prices*, *precios de materias primas* |
| Poverty, inequality and informality | *poverty*, *pobreza* |
| COVID-19 | *covid*, *reactiva peru* |

*Source: Illustrative entries from the fixed dictionary.*

A paper is counted once for a topic if at least one listed phrase matches its title or abstract. A paper may be counted under several topics. Annual percentages divide topic matches by the number of indexed Working Papers published that year. The rules were checked against sampled matches and plausible omissions. An early pass-through expression also matched interest-rate transmission, so the final dictionary requires an explicit exchange-rate connection. Likewise, bare mentions of liquidity do not count unless they meet a more specific rule. The analysis does not use AI to classify publications.

The full dictionary, annual denominators, document-level matches and validation sample are registered. The separately tracked unconventional-policy series is not included in the COVID-19 panel. Results refer to the captured corpus and may miss studies whose title and abstract do not use the selected expressions.

# References


Flavio Abanto, Alan Ledesma, and Fernando Pérez. El efecto traspaso del tipo de cambio y su evolución en Perú: evidencia de un VAR bayesiano con parámetros cambiantes para Perú. Working Paper 2026-018, Banco Central de Reserva del Perú, May 2026.

Marco Arena and Pedro Tuesta. El objetivo de la intervención del banco central: ¿el nivel del tipo de cambio, la reducción de la volatilidad cambiaria o ambos?: Un análisis de la experiencia peruana 1991–1998. Working Paper 1999-003, Banco Central de Reserva del Perú, 1999.

Banco Central de Reserva del Perú. El fenómeno El Niño Costero 2017 y la vulnerabilidad de los hogares peruanos. Reporte de Inflación, diciembre 2023, recuadro 3, pp. 71–75 (printed pagination), December 2023.

Beijing Academy of Artificial Intelligence. BAAI/bge-reranker-v2-m3: Model card, 2024.

Consultative Group on Risk Management. Governance of AI adoption in central banks. Cca report, Bank for International Settlements, January 2025.

Gordon V. Cormack, Charles L. A. Clarke, and Stefan Büttcher. Reciprocal rank fusion outperforms Condorcet and individual rank learning methods. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 758–759, 2009. doi: 10.1145/1571941.1572114.

Douglas Kiarelly Godoy de Araujo, Sebastian Doerr, Leonardo Gambacorta, and Bruno Tissot. Artificial intelligence in central banking. BIS Bulletin 84, Bank for International Settlements, 2024.

Leonardo Gambacorta, Byeungchun Kwon, Taejin Park, Pietro Patelli, and Sonya Zhu. CB-LMs: Language models for central banking. BIS Working Paper 1215, Bank for International Settlements, 2024.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6465–6488. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.398.

Andrew R. Garcia and Marco Vega. Forecasting inflation with a framework for model and neural architecture search with tree-structured search spaces. Working Paper 2024-009, Banco Central de Reserva del Perú, May 2024.

Andrew R. Garcia and Marco Vega. Can metadata guide variable selection for macroeconomic forecasting? Working Paper 2026-012, Banco Central de Reserva del Perú, March 2026.

Kemal Hajdarevic and Jasmina Selimovic. Offline NLP assistants for internal knowledge access in a central bank: Proof-of-concept systems for semantic search and HR chatbot. *Journal of Central Banking Theory and Practice*, 15(2):99–123, 2026. doi: 10.2478/jcbtp-2026-0014.

Pranab Islam, Anand Kannappan, Douwe Kiela, Rebecca Qian, Nino Scherrer, and Bertie Vidgen. FinanceBench: A new benchmark for financial question answering. arXiv:2311.11944, 2023.

Anton Korinek. Generative AI for economic research: Use cases and implications for economists. *Journal of Economic Literature*, 61(4):1281–1317, 2023. doi: 10.1257/jel.20231736.

Erick Lahura, Luz Maria Pumacahua, and Rebeca Sosa. Análisis textual y legibilidad de las notas informativas de la política monetaria en el Perú: 2001–2021. Working Paper 2023-004, Banco Central de Reserva del Perú, August 2023.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474, 2020.

Marco Jacopo Lombardi, Taejin Park, and Fernando Perez-Cruz. MaRCo@BIS: AI-agentic econometric analysis for central bankers. BIS Working Paper 1395, Bank for International Settlements, 2026.

Fernando J. Pérez Forero. Asymmetries and non-linearities in the exchange rate pass-through to inflation: Evidence for Peru. Working Paper 2026-007, Banco Central de Reserva del Perú, 2026.

Luca Petracca, Simone De Benedictis, Thomas Gottron, and Zlatina Hofmeister. Supporting users in seeking data on the ECB data portal: A use case for retrieval augmented generation. IFC Bulletin 67, Bank for International Settlements, 2026.

Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389, 2009. doi: 10.1561/1500000019.

Boon Keong Seah, Eilyn Chong, and Max Chew. Using retrieval-augmented generation (RAG) and LangChain to drive chatbots for interactive economic data analysis and insight. IFC Bulletin 67, Bank for International Settlements, 2026.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. Multilingual E5 text embeddings: A technical report. arXiv:2402.05672, 2024.

Diego Winkelried. ¿es asimétrico el pass-through en el Perú?: Un análisis agregado. Working Paper 2003-006, Banco Central de Reserva del Perú, 2003.

Diego Winkelried. Exchange rate pass-through and inflation targeting in Peru. Working Paper 2011-012, Banco Central de Reserva del Perú, 2011.

Daniel Woong, Iván Ortiz, David Ricardo Mora, and Juan Jesús Ramírez. Midiendo el tono de la reserva federal: un índice de sentimientos de banca central basada en IA generativa. *Revista Moneda*, (207):86–92, September 2026.