ParancU v1.1*

A Local-First Evidence Retrieval Prototype with Semantic-Anchor Scoring

Maurizio Scibilia

2026-08-19

Abstract

Dense retrieval is not inherently wrong, but its usual similarity objective is only an imperfect proxy for evidence containment. Its limitation is not that it represents meaning poorly, but that semantic proximity does not guarantee access to the specific evidence required by a question.

ParancU v1.1 is a local-first evidence retrieval prototype designed for small, user-controlled corpora. In the current implementation, local-first refers to on-device corpus persistence, embedding inference, scoring, ranking, and evidence inspection. Question-oriented retrieval metadata is still generated through a constrained external API step during corpus preparation.

The system compares user questions with guiding questions associated with source passages, while bounded lexical anchors allow exact or rare terms to influence ranking when relevant. Across three bounded corpora, semantic-anchor scoring improves Top-5 evidence containment relative to semantic-only retrieval within each benchmark. These results support the practical value of combining question-oriented semantic representations with lexical anchors, but they do not isolate representation quality from ranking quality.

ParancU v1.1 should therefore be read as an intermediate research stage. Its demonstrated contribution is a local multilingual retrieval pipeline with semantic-anchor scoring and inspectable source evidence. Its broader research hypothesis is that retrieval quality may also depend on how access to answer-bearing evidence is represented before ranking begins.

Guiding questions provide the current implementation of this idea, while future work will investigate automatically derived common representations between questions and source content that do not require generative interpretation.

1 Introduction

ParancU v1.1 evolves from EcoSearch v1.0, which began with a simple observation: semantic similarity does not entail evidence containment. Put plainly, users need passages that contain the evidence required by their questions, not merely passages that resemble those questions in topic or vocabulary.

In a common dense-retrieval workflow for bounded document collections, source documents are divided into chunks and each chunk is embedded directly into a vector representation. A user query is embedded in the same space, and the most similar chunks are passed downstream, often to a generative model that produces the final answer. This downstream generation can make imperfect retrieval less visible to the user: a fluent response may still be produced from evidence that is incomplete or merely related to the question. Generation may compensate for some retrieval imperfections, but it does not establish that the retrieval stage actually selected the evidence required by the query. ParancU therefore treats evidence access as a problem to be inspected before answer generation rather than inferred from the fluency of a downstream response.

The project retains that retrieval-first approach for small, user-controlled corpora while extending it into a local-first mobile application. Instead of relying only on direct similarity between a question and raw document chunks, ParancU associates each chunk with a guiding question intended to express one answerable intent supported by the source passage. Retrieval compares the user question with these question-shaped representations and returns the corresponding source evidence for inspection.

ParancU v1.1 develops this approach in a local-first mobile workflow. Corpus storage, embedding inference, retrieval, ranking, and evidence inspection are performed on the device. Corpus preparation still includes a constrained external step that generates summaries and guiding questions from each source chunk. Dense similarity is complemented by bounded lexical evidence when exact, rare, or corpus-distinctive terms matter.

The current prototype tests whether this combination improves access to answer-bearing passages beyond semantic

*ParancU v1.1 is available on the Apple App Store.

similarity alone. Across the reported local benchmarks, the best observed semantic-anchor settings improve Top-5 evidence containment relative to semantic-only retrieval. These experiments evaluate the combined behaviour of guiding-question representation, local E5 matching, lexical anchors, chunking, and ranking weights; they do not yet isolate representation, matching, or ranking quality.

This distinction motivates the broader research direction of the project. A passage may contain the evidence required by a question while its guiding question exposes a different or more general intent. Retrieval may therefore be limited not only by ranking, but also by how access to the evidence is represented before ranking begins. ParancU v1.1 uses one guiding question per chunk as a practical first implementation. Future work will investigate automatically derived, local, and reproducible common representations between questions and source content that do not require generative interpretation.

This paper makes four contributions. First, it operationalises guiding questions as question-oriented representations linked directly to inspectable source passages. Second, it implements local multilingual E5 embedding and exhaustive semantic-anchor ranking in a mobile application. Third, it evaluates this retrieval behaviour across three bounded corpora using evidence-containment metrics. Fourth, it identifies evidence-access representation as a research problem that should be evaluated separately from semantic matching and ranking. Earlier OpenAI-based experiments are retained as historical context, while the main validation focuses on the local E5 pipeline.

1.1 From EcoSearch to ParancU

The transition from **EcoSearch v1.0** to **ParancU v1.1** reflects more than a change of name. EcoSearch was originally conceived as a research prototype focused on evidence-oriented retrieval over user-provided documents. As the project evolved into a mobile, local-first application for capturing, importing, storing, and querying personal document collections, the original name no longer fully represented its scope or direction.

"U parancu" is the Sicilian expression for a block and tackle, a mechanical system of pulleys and ropes or chains used to lift heavy loads with reduced effort. The name was chosen in reference to the author's Sicilian origins and to the project's own roots. It also provides an appropriate metaphor for the system: ParancU is intended to reduce the effort required to extract useful evidence from complex or extensive documents.

ParancU identifies this next stage of the project. It preserves the retrieval principles developed in EcoSearch v1.0 while extending them into a more complete user-facing system centred on local document processing, multilingual embeddings, OCR-based capture, persistent corpora, and inspectable evidence. The change of name therefore marks the transition from the original retrieval prototype to a broader mobile application in which retrieval is one component of an end-to-end document interaction workflow.

Throughout this paper, **EcoSearch v1.0** refers to the original research prototype and to results obtained with that system, while **ParancU v1.1** refers to its current mobile, local-first evolution. This distinction preserves the continuity of the research lineage while making the architectural and product transition explicit.

2 EcoSearch v1.0: Guiding-Question Retrieval

EcoSearch v1.0 introduced the core idea that still defines the project: compare a user question with a question-shaped representation of corpus evidence rather than relying only on similarity with the raw passage text.

The first prototype was desktop-based. Users could upload a PDF, prepare it as a searchable corpus, and ask questions about its contents. During corpus preparation, the document was divided into chunks, and each chunk was associated with a short summary and a guiding question intended to express one answerable intent supported by the passage.

At query time, the user's question was embedded and compared with the stored guiding-question embeddings. The retrieved result remained the original source chunk linked to the closest representation, allowing the user to inspect the evidence rather than receiving only a generated answer. When this route was unavailable or insufficient, the system could fall back to direct chunk retrieval.

The early EcoSearch v1.0 results were sufficiently promising to justify moving beyond the desktop prototype. The next step was not simply to reproduce the same workflow on a smaller screen, but to evolve the project into ParancU, a setting where its intended use became more natural: a personal, user-controlled retrieval tool operating directly on documents captured, stored, and queried from a mobile device.

This transition led to several practical extensions. OCR was added so that users could create corpora from scanned pages as well as uploaded documents. Corpus storage and retrieval were moved towards a local-first architecture. A Semantic/Lexical control was introduced to let users adjust the balance between broad semantic matching and

exact-term evidence when the corpus or question required it.

ParancU v1.1 therefore represents both a functional expansion and a refinement of the original guiding-question approach. It preserves retrieval-first access to source evidence while adapting the workflow to mobile document capture, local processing, multilingual embedding, and interactive ranking control.

In the current system, the guiding question remains the primary semantic representation associated with each chunk. It should nevertheless be understood as one practical approximation of what the passage can answer, rather than as a complete representation of every information need supported by that passage.

3 ParancU v1.1: System Architecture

ParancU v1.1 transforms source documents into a set of overlapping evidence units that can be queried locally. Source text is first segmented into overlapping chunks. Each chunk is then associated with a guiding question intended to express one answerable intent supported by the passage. A short summary and other metadata are stored alongside it. The guiding question is encoded locally with E5 and compared with the encoding of the user's question, while lexical evidence from the guiding question, summary, and original chunk provides a second ranking signal. The resulting semantic and lexical scores are combined to rank the original source chunks returned to the user.

The retrieval logic can be viewed as three distinct components. First, the guiding question provides a question-oriented access representation for each chunk. Second, E5 measures semantic similarity between the user question and that representation. Third, semantic similarity is combined with bounded lexical evidence to produce the final ranking score.

A defining feature of the architecture is therefore that the representation used for semantic access is not the same object ultimately returned to the user. The guiding question is used to access the evidence, while the original source chunk remains the inspectable retrieval result.

The experiments reported later evaluate the complete pipeline rather than the individual contribution of each component.

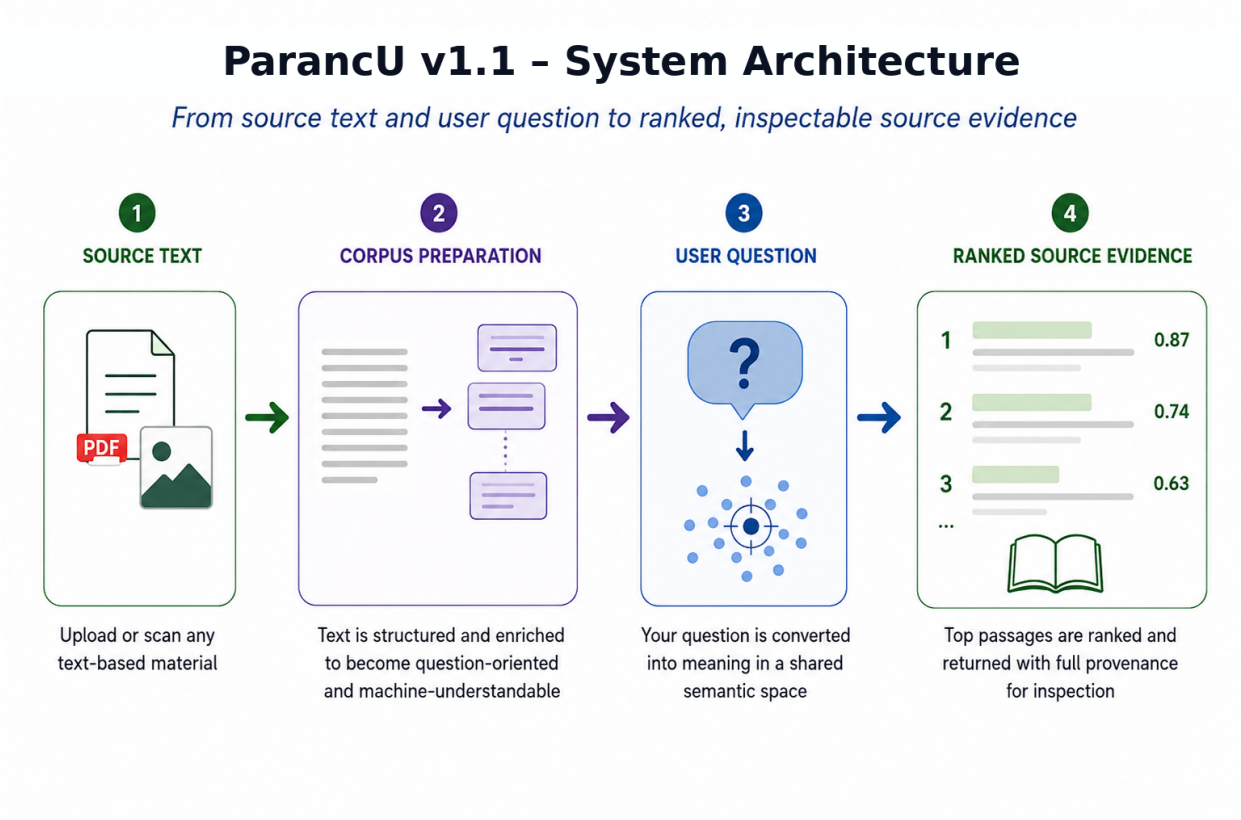

Figure 1: *ParancU v1.1 system architecture. Source material is prepared as a reusable corpus, queried through a shared semantic space, and returned as ranked source evidence.*

3.1 Source Text and Deterministic Sentence Splitting

ParancU can acquire source material as typed or pasted text, imported PDF content, or text extracted from images through OCR. Regardless of its origin, the material enters the retrieval pipeline as plain text.

The first architectural operation is deterministic sentence splitting over the normalised source text. The implementation identifies sentence boundaries using punctuation while applying rule-based protections for abbreviations, initials, and fragments that appear to continue the preceding sentence. It also recomposes some segments when the following text begins with a lowercase letter, a connective expression, or another pattern suggesting that the apparent boundary was not a complete semantic break.

This stage is entirely local and does not require a language model. Its purpose is not to interpret the document, but to produce stable sentence units while retaining their location in the normalised source text. Each sentence is assigned an identifier together with its start character offset and exclusive end character offset.

The resulting sentence representation provides the first provenance layer of ParancU. Later chunks retain the identifiers of the sentences they contain, allowing retrieved evidence to be traced back through those sentence records to its location in the source text.

3.2 Overlapping Chunk Construction

The sentence sequence is converted into overlapping chunks through a sliding-window procedure. In the current default configuration, each chunk contains three sentences and overlaps the next chunk by two sentences. The effective stride is therefore one sentence:

```
Chunk 1: sentences 0, 1, 2
Chunk 2: sentences 1, 2, 3
Chunk 3: sentences 2, 3, 4
```

Once at least one full three-sentence chunk has been created, a shorter trailing window is not emitted. Documents containing fewer than three sentences are retained as a single shorter chunk.

This high-overlap design reduces the risk that useful evidence will be divided at an arbitrary chunk boundary. A fact introduced at the end of one passage can remain associated with the context that follows it, while the same sentence may participate in several alternative evidence windows.

With a three-sentence window and a one-sentence stride, most interior source sentences occur in three consecutive chunks, each time with a different surrounding context. The resulting redundancy is intentional: the same piece of evidence can participate in several contextual windows, each of which can later produce a distinct retrieval representation, rather than depending on a single chunk boundary or contextual formulation. Because retrieval metadata is constructed independently for each chunk, this also creates multiple potential semantic access paths to evidence that appears in overlapping windows.

Each chunk retains:

- its own identifier;
- the identifiers of its component sentences;
- the position of its first sentence and the exclusive boundary after its last sentence;
- the original source text covered by the window.

At this stage, the summary, guiding-question, and answer-focus fields are already present in the chunk structure but remain empty. Chunk construction therefore remains a deterministic transformation of the source rather than an act of semantic interpretation.

3.3 GPT-Based Retrieval-Metadata Construction

Each chunk is enriched with a short summary and a guiding question. An `answer_focus` field is also generated and stored for compatibility, but it is not used by the active v1.1 ranking path.

In the current mobile implementation, these fields are generated inside ParancU through the OpenAI Responses API, using an API key supplied by the user and stored locally by the application. The currently configured default model is `gpt-5.4-mini`, although the model name can be changed through the application environment.

This stage uses a generative Transformer decoder, but its task is narrower than ordinary question answering or open-ended summarisation. The model is instructed to construct retrieval metadata grounded exclusively in the chunk text: it should not add outside knowledge, repair incomplete information, infer unstated relationships, or formulate questions that require facts absent from the passage.

The current constraints include:

- a summary of no more than twenty words;
- a guiding question of no more than ten words that must be directly answerable from the chunk.

The guiding question is the primary semantic representation used by ParancU v1.1. It attempts to express one explicit information need supported by the source passage. Instead of comparing the user question with the prose of the original chunk, ParancU compares it with a compact question-shaped representation deliberately constructed around answerability.

This representation is necessarily selective. A passage containing several facts, entities, relations, dates, or definitions may support multiple valid questions, while the current implementation normally stores only one guiding question. The guiding question should therefore be understood as a practical approximation of one access path into the passage, not as a complete account of everything the passage can answer.

The original chunk is not discarded. It remains linked to the guiding question as the inspectable source of evidence and also contributes to lexical-anchor scoring together with the summary and guiding question. This allows exact or rare terms in the source text to compensate partially when the semantic representation does not align closely with the user's wording.

This preparation stage is the main external-service boundary of the current architecture: each source chunk is sent to the configured generation service to construct its metadata. Once that metadata has been created, embedding, comparison, ranking, persistence, and evidence inspection proceed locally.

3.4 Local Multilingual E5 Encoding

The guiding question and answer-focus representation are converted into dense vectors using `intfloat/multilingual-e5-small`.

E5 is a family of encoder-only Transformer models designed for information retrieval. Unlike the GPT decoder used during corpus preparation, E5 does not generate summaries, questions, or answers. It maps text into a numerical representation whose position in a shared vector space reflects retrieval-oriented semantic relationships.

The model name identifies four relevant properties:

- `intfloat` identifies the checkpoint namespace used by ParancU;
- `multilingual` indicates support for multiple languages;
- `e5` identifies a retrieval-oriented embedding family;
- `small` identifies the compact model variant used by ParancU.

The compact multilingual variant provides multilingual retrieval within a model size compatible with the current local mobile implementation.

ParancU runs a quantised INT8 ONNX version of the model through `onnxruntime-react-native`. Tokenisation is performed locally using Hugging Face Tokenizers. Input sequences are limited to 512 tokens, and the model returns one 384-dimensional vector for each encoded text.

E5 distinguishes the two sides of retrieval through explicit prefixes:

```
query: <user question>
passage: <corpus representation>
```

The user's question is encoded in query mode. The guiding question associated with each corpus chunk is encoded in passage mode.

For every prepared chunk, ParancU stores:

- the guiding-question vector;
- the answer-focus vector;
- the original chunk text;
- the summary;
- the textual guiding question;
- the sentence identifiers linking the chunk back to the sentence-level provenance records.

The active retrieval path currently uses only the guiding-question vector. The answer-focus vector is generated and stored but is not used in semantic scoring or lexical matching.

The central geometric transformation can therefore be expressed as:

```
user question
    ↓ E5 query encoding
384-dimensional query vector
```

```
guiding question
    ↓ E5 passage encoding
384-dimensional corpus vector
```

Natural-language questions and corpus access representations are thus converted into comparable points in the same retrieval space.

3.5 Exhaustive Vector Comparison

At query time, ParancU generates a local E5 embedding for the user's question and compares it with the stored guiding-question embedding of every chunk.

An earlier implementation explored FAISS-based vector indexing. For the small, user-controlled corpora targeted by the current mobile prototype, however, retrieval-time comparison was already fast in practical testing. FAISS was therefore removed from the active architecture because it introduced additional native integration, index persistence, and synchronisation complexity without providing a meaningful user-visible latency benefit at the present scale.

The current implementation consequently performs an exhaustive comparison across the prepared corpus. For each chunk i, the semantic score is calculated through cosine similarity:

$$s_i = \frac{q \cdot g_i}{\|q\|\|g_i\|}$$

where:

- q is the vector representing the user's question;
- g_i is the stored guiding-question vector for chunk i;
- s_i is their cosine similarity.

Cosine similarity measures the angle between the two vectors rather than their absolute magnitude. Questions and guiding questions that point in similar directions in the E5 space receive higher semantic scores.

The implementation bounds the resulting score to the interval from zero to one. Negative cosine values therefore contribute zero semantic evidence.

Exhaustive comparison also preserves an important property of the present semantic-anchor architecture: every chunk remains available for both semantic and lexical scoring. A FAISS-first pipeline would normally retrieve a semantic shortlist before applying anchor evidence, potentially excluding chunks whose semantic score is weaker but whose exact lexical evidence is strong.

FAISS or another approximate nearest-neighbour index remains a possible scaling strategy for substantially larger corpora. Its reintroduction should be driven by measured mobile retrieval latency rather than by architectural convention. For the file sizes currently handled by the prototype, exhaustive comparison is simpler, fully deterministic, and already sufficiently fast.

3.6 Lexical-Anchor Construction

Dense similarity captures broad semantic relationships, but it can underweight the exact textual features that distinguish an answer-bearing passage from a merely related one. ParancU therefore constructs a second, bounded lexical signal for every query-chunk pair.

Lexical evidence is calculated over the concatenation of:

- the guiding question;
- the summary;
- the original source chunk.

The current ranking path does not use the answer focus for lexical matching.

Two lexical components are calculated: exact anchors and soft keywords.

3.6.1 Exact Anchors

Exact anchors are query tokens that contain features commonly associated with identifiers or high-information expressions:

- uppercase form;
- digits;
- underscores;

- hyphens.

Examples include:

```
MS
LUNAR
BASEBALL
GPT-4
1130
```

For each candidate chunk, ParancU measures both how many query anchors are present and whether those anchors recur in the candidate text.

Coverage contributes eighty per cent of the exact-anchor score. Repetition contributes the remaining twenty per cent and is capped after three occurrences. This prevents repeated terms from dominating the ranking without limit.

3.6.2 Soft Keywords

Not all useful terms appear in uppercase or contain digits and symbols. ParancU therefore adds a weaker signal based on ordinary query words.

A soft keyword must:

- meet a minimum length, normally four characters;
- not belong to the stopword set supplied to the retrieval function, where one is configured;
- occur in the corpus;
- remain relatively rare across the collection.

The current document-frequency threshold retains terms appearing in no more than approximately twenty per cent of the corpus chunks, subject to a minimum threshold of one chunk.

This allows corpus-distinctive expressions such as names, medical terms, or specialised concepts to contribute even when they do not satisfy the exact-anchor rule.

The two lexical components are combined as:

$$a_i = \frac{2}{3}e_i + \frac{1}{3}k_i$$

where:

- e_i is the exact-anchor score;
- k_i is the soft-keyword score;
- a_i is the final bounded anchor score for chunk i.

Exact-anchor evidence therefore has twice the internal influence of softer keyword support.

3.7 Semantic-Anchor Score Fusion

ParancU combines semantic similarity and lexical anchor evidence into one final score for every chunk:

$$f_i = \alpha s_i + (1 - \alpha)a_i$$

where:

- s_i is the semantic cosine score;
- a_i is the combined anchor score;
- α is the selected semantic weight;
- $1 - \alpha$ is the corresponding lexical-anchor weight.

The default configuration uses:

$$\alpha = 0.65$$

which corresponds to a 65/35 Semantic/Lexical balance.

The mobile retrieval engine constrains the semantic weight to the interval from 0.4 to 1.0. The Semantic/Lexical control currently allows the interface to display values below 0.4; however, any semantic weight below this threshold is clamped to 0.4 before scoring. Consequently, settings from 0/100 through 40/60 all produce an effective 40/60

Semantic/Lexical balance. The lower bound preserves semantic retrieval as the primary signal and prevents the bounded anchor component from becoming a standalone sparse retriever. Above this threshold, the selected Semantic/Lexical ratio is applied directly.

The score is calculated directly for every chunk, and the complete set is then sorted by final score.

This distinction is architecturally important. ParancU v1.1 does not currently perform reranking in the strict sense. It does not first retrieve a dense shortlist and then pass those candidates to a second ranking model.

Instead, the active pipeline:

```
scores every chunk
    ↓
combines semantic and lexical evidence
    ↓
sorts all chunks by final score
```

The appropriate description is therefore semantic-anchor score fusion or semantic-anchor ranking, not reranking.

3.8 Ranked Evidence Output

After all chunks have been scored, ParancU sorts the complete corpus by final semantic-anchor score and returns the highest-ranking candidates.

The current mobile implementation preserves these candidates as separate evidence units. It does not automatically merge neighbouring chunks or reconstruct a larger passage around the Top-1 result. This keeps the displayed evidence aligned with the actual ranking and avoids adding lower-scoring text to the strongest retrieved chunk.

For a standard Top-k request, each returned result contains:

- the final semantic-anchor score;
- the semantic cosine score;
- the guiding question;
- the summary;
- the original source chunk;
- the chunk index;
- the identifiers of the source sentences contained in that chunk.

The default retrieval depth is five candidates. These Top-5 results form the inspectable evidence window used by the current prototype and correspond directly to the primary Top-5 FULL+PARTIAL evaluation metric. Because chunks overlap, neighbouring candidates can contain some of the same source sentences.

Overlapping chunks may therefore appear separately in the ranked results. For example, chunks covering sentences 1-3 and 2-4 remain distinct candidates rather than being fused into a synthetic passage covering sentences 1-4. This preserves the provenance, score, and retrieval metadata of each candidate.

The final output of ParancU is not an automatically generated answer. It is a ranked set of source-grounded evidence candidates whose ordering can be inspected together with the guiding questions and scores that produced it.

The next chapter describes how this retrieval architecture is exposed through the ParancU application.

4 The ParancU Mobile Application

ParancU v1.1 exposes the retrieval architecture described in Chapter 3 through a workflow organised around source acquisition, corpus preparation, evidence retrieval, and local corpus management. Rather than presenting retrieval as an opaque conversational interface, the application keeps these stages explicit so that users can inspect how source material becomes a searchable corpus and how ranked evidence is returned.

The interface is organised around three primary sections: **Scan**, **Upload**, and **Ask**. A separate **Manage corpus** panel provides corpus persistence, import/export, clearing, and API-key controls. The language selector in the upper-right corner changes the interface language without modifying the active corpus or its retrieval configuration.

4.1 Source Acquisition

ParancU supports camera capture, stored images, pasted text, and imported PDFs. The **Scan** section exposes camera capture and image-based OCR routes. The **Upload** section supports corpus creation from pasted text and imported PDF files. At this stage, no corpus is active. These routes differ only in how source text enters the

Figure 2: *The ParancU mobile application. Main interface states from source acquisition and corpus preparation to evidence retrieval and local corpus management.*

system; once text has been obtained, all sources proceed through the same corpus-preparation workflow. Interface language and corpus language remain separate aspects of the application.

4.2 Corpus Preparation

After a PDF has been selected, the application displays it as draft content and makes clear that it is not yet searchable. OCR follows the same preparation model: the selected image and extracted text are shown before the corpus is built. Selecting **Prepare Corpus** starts the pipeline described in Chapter 3. The source is segmented deterministically, retrieval metadata is generated through the constrained external preparation step, and the resulting guiding questions are embedded locally with E5.

The OCR preview allows the user to inspect and edit the text that will enter the retrieval pipeline. Camera images, stored images, pasted text, and PDFs therefore converge on the same chunk, metadata, embedding, and provenance structure. Corpus preparation does not translate or rewrite the original source passage, so retrieved evidence remains in the language of the source material.

4.3 Evidence Retrieval

Once a corpus has been prepared and activated, users query it through the **Ask** interface. The **Semantic/Lexical** control exposes the ranking balance used for the current retrieval request. After a question is submitted, ParancU compares it with the guiding-question representation associated with every chunk, combines the resulting semantic score with lexical-anchor evidence, and returns the five highest-ranking source candidates.

The candidate selector identifies the currently displayed result within the Top-5 list. Each candidate presents its final score, guiding question, summary, and original source passage. Because candidates are not merged after ranking, the displayed passage remains directly aligned with the score and metadata assigned to that chunk.

The interface deliberately exposes evidence rather than hiding retrieval behind a generated answer. The example shown in Figure 2 also illustrates the multilingual design: an English question is retrieving evidence from an Italian corpus. Qualitative testing also showed that the same Italian corpus could be queried with semantically equivalent questions in other scripts, including Simplified Chinese, while preserving the same relevant evidence near the top of the semantic ranking. This observation is illustrative rather than a multilingual benchmark result. Cross-language retrieval is supported by the embedding model architecture, although it is not evaluated separately in the experiments reported here.

4.4 Local Corpus Management

Prepared corpora can be saved locally, reloaded across sessions, imported from an existing snapshot, or exported for reuse and inspection. The management panel separates the active corpus from the draft acquisition flow. It also exposes the OpenAI API key used only during retrieval-metadata construction.

Once preparation has completed, a corpus can be reused without repeating PDF extraction, OCR, retrieval-metadata generation, or local embedding computation. Corpus persistence, E5 query embedding, exhaustive semantic-anchor scoring, ranking, and evidence inspection then operate directly on the device.

The application therefore mirrors the architecture introduced in Chapter 3: acquisition and constrained metadata construction occur during corpus preparation, while corpus persistence, retrieval, ranking, and evidence inspection remain local operations under the user's control.

5 Experiments

ParancU is evaluated as an evidence retrieval system. The central question is not whether a retrieved passage appears topically related to the query, but whether it contains the evidence needed to answer it.

The experiments reported here evaluate the complete retrieval pipeline, including guiding-question construction, local or hosted embeddings, lexical-anchor scoring, chunking, and ranking. They therefore measure end-to-end evidence containment rather than isolating the contribution of each component.

The evidence comes from two experimental phases: an earlier OpenAI-based benchmark series and a later local E5 validation series. These phases use different embedding models and partially different benchmark provenance, so their results are reported separately rather than treated as a direct model comparison.

5.1 Corpora

Three corpora were used to test different retrieval conditions.

Table 1: *Experimental corpora used in the ParancU v1.1 evaluation.*

Corpus	Description	Why it matters
Sicilia Normanna	Historical Italian corpus with names, places, dates, battles, rulers, migrations and administrative events.	Tests entity-heavy historical retrieval in Italian.
Question Answering	Technical/conceptual corpus about question-answering systems, including terms such as LUNAR, BASEBALL, BART and MathQA.	Tests conceptual and technical retrieval with many system names and domain-specific terms.
Multiple Sclerosis	Medical corpus with biomedical vocabulary, risk factors, epidemiology, symptoms, diagnostic criteria and study findings.	Tests a semantically dense domain in which many passages share overlapping biomedical vocabulary, while only some contain the exact evidence required by a question.

5.2 Two Experimental Phases

The historical benchmark used hosted OpenAI embeddings and LLM-assisted oracle construction. It established the first quantitative evidence that semantic-anchor calibration could improve evidence containment. These figures remain useful as a historical baseline, but they should not be interpreted as measurements of the current local pipeline.

The later validation used `intfloat/multilingual-e5-small` locally. The Question Answering benchmark was available as a frozen local grid. The Sicilia Normanna benchmark used a revised 50-question oracle. The Multiple Sclerosis benchmark had to be reconstructed from the recovered source text and corpus: 50 short questions were written directly from the text, with a ten-word design limit and a maximum observed length of six words, then linked to source-sentence evidence before being mapped to corpus chunks.

Because the local benchmarks do not share identical question sets and oracle provenance with every historical benchmark, the two phases are reported separately. Their purpose is complementary: the historical series shows the potential of the original architecture, while the local series verifies the behaviour of the present local embedding pipeline. The reconstructed Multiple Sclerosis test set is text-grounded and human-readable, but it is not an independently sampled distribution of real user queries.

The two phases should therefore be interpreted as complementary evidence about the behaviour of the ParancU architecture, not as a controlled comparison between embedding models.

5.3 Evaluation Methodology

5.3.1 Oracle Overlap

Oracle overlap measures whether the retrieved candidate window intersects the source evidence associated with each question. The primary metric is Top-5 FULL+PARTIAL evidence containment: at least one of the first five retrieved chunks contains all or part of the oracle evidence. A PARTIAL match shows access to some oracle-linked evidence, but not necessarily enough evidence to answer the question completely. The metric does not evaluate generated-answer correctness or completeness.

Metric	Meaning
Top-1 FULL	The first retrieved chunk contains all oracle evidence sentences.
Top-1 FULL+PARTIAL	The first retrieved chunk contains all or part of the oracle evidence.
Top-5 FULL	At least one of the first five chunks contains all oracle evidence.
Top-5 FULL+PARTIAL	At least one of the first five chunks contains all or part of the oracle evidence.

The historical phase cached LLM-selected oracle spans so that changes in retrieval weights did not regenerate the evidence reference. In the reconstructed Multiple Sclerosis validation, questions were derived from the source text first, associated with source-sentence IDs, and only then mapped to chunk IDs. This avoids generating test questions from the corpus chunk structure itself.

5.3.2 Expected-Focus Sanity Check

A secondary expected-focus diagnostic was retained for the historical Multiple Sclerosis experiment. It checked whether retrieved chunks contained an expected answer cue such as a name, acronym, technical term, or short phrase. This check is weaker than oracle overlap and is not treated as the primary metric.

5.4 Historical OpenAI-Based Results

The earlier OpenAI-based experiments produced the highest reported scores within their original benchmark settings. They remain useful as historical evidence of the performance observed under the original hosted embedding configuration.

Table 2: *Historical OpenAI-based oracle-overlap results.*

Corpus	Baseline	Best setting	Best Top-5	Gain
Multiple Sclerosis	67.16%	50/50	95.52%	+28.36
Question Answering	77.97%	65/35	84.75%	+6.78
Sicilia Normanna	75.51%	45/55	79.59%	+4.08

The baseline column reports the semantic-heavy 90/10 setting used in the historical experiments; gains are percentage-point differences within each original benchmark.

These experiments are retained as a baseline rather than as the current product configuration. They used hosted embeddings and different benchmark provenance, and therefore cannot be treated as a like-for-like comparison with the present local pipeline.

5.5 Local E5 Validation Results

The local validation used 384-dimensional multilingual E5 embeddings. Once the corpus had been prepared, query embedding, semantic-anchor scoring, ranking, and evidence inspection were performed locally and did not require an OpenAI key.

Table 3: *Local E5 validation results. Best Top-5 is the highest observed result on the tested weight grid; gains are measured against the 100/0 semantic-only setting within each benchmark.*

Corpus	Semantic-only	Best Top-5	Gain	Balance
Question Answering	78.33%	85%	+6.67	65/35
Sicilia Normanna	54%	82%	+28	55/45
Multiple Sclerosis	42%	78%	+36	70/30

For Multiple Sclerosis, 55/45 and 50/50 produced the highest mean oracle-overlap score, while 70/30 produced the highest observed Top-5 hit rate. The gains above are internal to each benchmark and should not be compared across corpora as if the test sets had identical provenance or difficulty.

The local values should not be compared directly with the historical figures because benchmark provenance differs. They are used here to characterise the current E5 pipeline within each local benchmark.

The Multiple Sclerosis grid also illustrates why a single number is insufficient. The highest observed Top-5 hit rate occurs at 70/30, while the highest mean oracle-overlap score occurs at 55/45 and 50/50. With 50 questions, a two-point difference corresponds to one query, so neighbouring settings should be interpreted as an observed plateau rather than as statistically distinct optima.

5.6 Interpreting the Results

The results are best understood as evidence of potential rather than as a finished product claim. The historical experiments show that strong evidence-containment scores were observed under the original hosted embed-

ding configuration. The local experiments show that the same general retrieval workflow remains viable with `intfloat/multilingual-e5-small`.

Within each local benchmark, the best observed semantic-anchor setting improves Top-5 evidence containment relative to the corresponding semantic-only setting. This is the principal empirical finding of the present study. The gain varies across corpora, suggesting that lexical-anchor usefulness depends on corpus characteristics, question formulation, benchmark construction, and the availability of exact or corpus-distinctive terms.

These experiments evaluate guiding-question construction, chunking, multilingual embeddings, lexical anchors, and ranking weights together. They therefore establish an end-to-end ranking result, not an independent measure of the quality of the guiding-question representation or E5 matching. The possible causes of a miss are separated conceptually in the Discussion rather than treated here as experimentally established error classes.

The reported percentages should also be interpreted within the limits of each benchmark. The question sets are relatively small, and in the 50-question benchmarks a two-percentage-point difference corresponds to one query. Neighbouring weight configurations should consequently be read as observed plateaus rather than as statistically distinct optima.

Finally, the Python benchmark results do not establish exact numerical parity with the mobile ONNX runtime. The local mobile pipeline implements the same retrieval design, but runtime-level equivalence should be tested separately through a reproducibility benchmark executed directly on the application build.

6 Discussion

Dense semantic similarity remains useful. In the three local benchmarks studied here, however, at least one non-zero lexical-anchor setting produced a higher observed Top-5 result than the semantic-only guiding-question baseline. This does not show that lexical retrieval should replace dense retrieval, or that hybrid retrieval is universally superior; it shows that ParancU's current semantic path can benefit from bounded exact-term evidence.

The empirical result is therefore primarily a ranking result. The architecture raises a separate question upstream of ranking: whether the chunk exposes the information need through an adequate access representation. Even when it does, E5 must still align that representation with the user query before score fusion can rank the candidate. Representation, semantic matching, and ranking are therefore distinct possible failure points, although the present benchmarks do not measure them independently.

The original chunk remains part of the retrieval process. Although semantic similarity is calculated primarily through the guiding-question embedding, lexical-anchor scoring also examines the guiding question, summary, and source chunk. Exact or rare terms in the source can therefore compensate partially when the guiding question does not align closely with the user query.

Qualitative inspection nevertheless suggests several ways in which a single guiding question may provide incomplete semantic access to a passage. It may compress several possible information needs into one broad question, omit a valid but less salient intent, or emphasise one interpretation while obscuring another. These patterns are working hypotheses and require a documented error analysis before they can be treated as established failure categories.

The best semantic-anchor balance also varies by corpus and evaluation metric. Historical Multiple Sclerosis experiments favoured stronger lexical pressure, while the reconstructed local benchmark produced its highest Top-5 hit rate at 70/30 and its best mean overlap at 55/45 and 50/50. Question Answering and Sicilia Normanna show different plateaus. There is therefore no evidence for one universal semantic-anchor ratio.

Top-5 remains important because ParancU is designed to expose an inspectable evidence window rather than to force a single generated answer. Bringing the correct passage into the candidate set may be more valuable than presenting an overconfident Top-1 result. This also provides a practical foundation for future reranking, fallback strategies, and optional answer generation.

ParancU v1.1 should therefore be read as an intermediate retrieval prototype. Its current contribution lies in the combination of local multilingual embeddings, bounded lexical evidence, inspectable source passages, and a mobile workflow. Its broader research value lies in exposing questions about how source evidence should be represented before ranking, without claiming that those questions have already been resolved.

6.1 Product Implications

The Semantic/Lexical control should remain available in the research and prototype interface because it makes retrieval behaviour inspectable and allows different corpora to be explored without hiding the scoring balance.

For broader product use, simplified presets may be more appropriate than requiring users to understand embeddings

or lexical weighting. Presets such as Balanced, Technical, and Anchor-heavy could provide useful starting points, while the slider remains available as an advanced override.

These presets should be presented as interface hypotheses rather than validated universal defaults. A future version may also estimate an initial balance automatically from corpus characteristics such as acronym density, rare-term density, entity frequency, vocabulary overlap, and semantic concentration.

The longer-term product direction remains a standalone local-first application. Corpus storage, import/export, embedding inference, retrieval, ranking, and evidence inspection should remain under the user's control. External generation should be limited to the current preparation stage and progressively removed where practical.

7 Limitations and Future Work

ParancU v1.1 is an evidence-retrieval study, not a complete evaluation of generated answer quality. Several limitations remain:

- only three corpora were tested;
- the question sets are curated and relatively small;
- the reported best Semantic/Lexical balances are selected and evaluated on the same benchmark grids rather than on held-out questions;
- Top-5 FULL+PARTIAL can credit incomplete oracle evidence, and overlapping chunks can make neighbouring candidates redundant;
- historical and local experiments use different embedding models and partially different benchmark provenance;
- the reconstructed Multiple Sclerosis benchmark is grounded in the recovered text but does not reproduce the original historical benchmark;
- results depend on chunk size, overlap, guiding-question quality, lexical-anchor availability, and ranking weights;
- the current evaluation measures the full retrieval pipeline rather than isolating individual components;
- a single guiding question may not represent every information need supported by a multi-fact chunk;
- local embedding inference has been implemented, but retrieval-metadata construction still depends on a constrained external generative step;
- direct numerical parity between the Python benchmarks and the mobile ONNX runtime has not yet been established;
- cross-language retrieval is supported by the multilingual embedding model but has not been evaluated separately.

7.1 Separating Representation, Matching, and Ranking

Future experiments should distinguish three questions. First, does the representation associated with a chunk expose the information need contained in the benchmark question? Second, if it does, does the embedding model align the query and representation strongly enough? Third, do semantic and lexical scores place the relevant candidate within the desired window?

This separation would make it possible to distinguish chunking or representation failures from semantic-matching and score-fusion failures. A documented analysis should use explicit criteria, representative examples, and counts rather than relying only on informal inspection.

7.2 Beyond Generated Guiding Questions

The current implementation normally associates one generated guiding question with each chunk. This is a practical first approximation, but the larger research problem is not simply to generate more questions. It is to find a common representation that can be derived automatically from source content and compared reliably with user questions without requiring generative interpretation.

Multiple guiding questions remain one useful baseline. Other candidates include micro-facts, claims, definitions, relation statements, event descriptions, structured entity and date cues, and learned prototypes. These forms should not be assumed equivalent: some represent answer-oriented content, while others function more naturally as lexical or structured access cues. The goal is to improve coverage of the information needs supported by a passage while preserving provenance and controlling redundancy, storage cost, and retrieval complexity.

A separate qualitative observation concerns multilingual access. Figure 3 shows a Simplified Chinese query applied to the Italian Sicilia Normanna corpus using semantic-only retrieval. The example is included as an implementation-level illustration of cross-language semantic matching rather than as part of the formal multilingual evaluation.

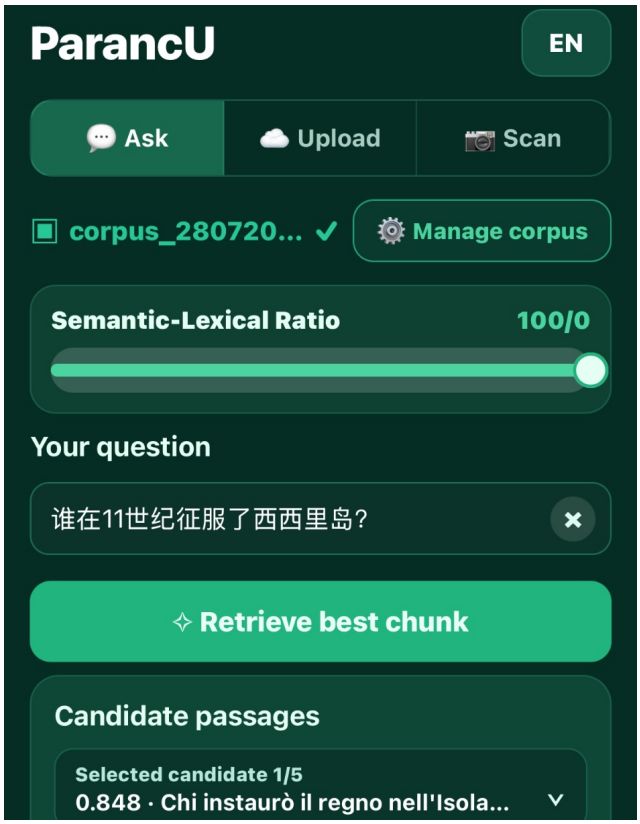

Figure 3: *Qualitative cross-language retrieval example in the ParancU mobile application. A Simplified Chinese question queries an Italian Sicilia Normanna corpus under semantic-only retrieval (100/0). The relevant Norman evidence is ranked first with a semantic score of 0.848. This example illustrates the shared multilingual embedding space of* `intfloat/multilingual-e5-small`*; it is not part of the formal benchmark evaluation.*

7.3 Local and Reproducible Metadata Construction

The long-term architecture should reduce or remove dependence on free-form autoregressive generation during corpus preparation. Any replacement should be locally derivable where practical, reproducible for a fixed source, source-faithful, and traceable to the original passage.

Possible directions include:

- extractive templates;
- micro-fact extraction;
- relation extraction;
- pseudo-query mining;
- contrastive representation learning;
- answer-aware representation learning;
- clustering-based prototypes;
- constrained autoencoding.

These are research directions rather than selected implementations. They should be evaluated according to retrieval coverage, reproducibility, latency, storage cost, and provenance preservation.

7.4 Stronger Baselines and Runtime Validation

Future experiments should compare ParancU against:

- dense-only retrieval;
- BM25;
- conventional dense-plus-BM25 hybrid retrieval;

- direct chunk embedding;
- summary embedding;
- guiding-question embedding without lexical anchors;
- multiple guiding questions per chunk;
- lightweight reranking over the Top-5 or Top-10 candidates.

Ranking weights should also be selected on calibration questions and evaluated on held-out questions. The benchmark should additionally be executed through the mobile ONNX implementation to verify whether ranking outputs and aggregate scores match the Python evaluation pipeline closely enough for reproducible reporting.

7.5 Multilingual Evaluation

Because `intfloat/multilingual-e5-small` maps queries and passages from multiple languages into a shared embedding space, ParancU can support scenarios such as asking a question in English about an Italian corpus. Figure 3 provides a qualitative example using a Simplified Chinese query over the Italian Sicilia Normanna corpus under semantic-only retrieval. The observation demonstrates that cross-language semantic matching is operational in the current mobile implementation, but it is not treated as a formal multilingual benchmark result.

This capability should therefore be evaluated explicitly rather than inferred from either model design or isolated qualitative examples. Future tests should include:

- same-language Italian retrieval;
- same-language English retrieval;
- English questions over Italian corpora;
- Italian questions over English corpora;
- mixed-language corpora;
- language-specific lexical-anchor behaviour.

7.6 Corpus-Level Balance Selection

Future versions may estimate an initial Semantic/Lexical balance from corpus statistics. Relevant signals may include:

- rare-term density;
- entity and acronym density;
- document-frequency distribution;
- vocabulary overlap;
- semantic concentration;
- observed retrieval confidence.

The slider could remain available as an inspection and override tool, while the application proposes a corpus-aware default.

7.7 Additional Future Directions

Future work should also:

- expand evaluation to legal, scientific, financial, software-documentation, and multilingual corpora;
- test real user questions rather than only curated benchmark questions;
- report latency, memory use, corpus size limits, and battery cost on representative devices;
- evaluate whether overlapping Top-5 candidates create useful evidence coverage or excessive redundancy;
- test lightweight candidate diversification;
- validate the interface language selector and retrieval presets with real users;
- evaluate optional answer generation separately from retrieval quality.

8 Conclusion

ParancU v1.1 demonstrates a local-first evidence-retrieval workflow for small, user-controlled corpora. The system combines guiding-question representations, local multilingual E5 embeddings, bounded lexical-anchor evidence, exhaustive score fusion, and inspectable Top-5 source passages.

The principal empirical result is that the best observed semantic-anchor setting improves Top-5 evidence containment relative to semantic-only retrieval within each local benchmark. The observed maxima were 85% for Question Answering, 82% for Sicilia Normanna, and 78% for the reconstructed Multiple Sclerosis benchmark. Because the

balance was selected on the same benchmark grid, these are descriptive maxima rather than held-out performance estimates, and they should not be compared directly across corpora or with the historical OpenAI-based experiments.

The study also highlights an unresolved issue upstream of ranking. The current implementation normally represents each chunk through one guiding question, while the source passage may support several distinct information needs. This may limit semantic access to some evidence, although the present experiments do not yet isolate or quantify representation failures independently from ranking failures.

ParancU v1.1 should therefore be understood as an intermediate prototype rather than a completed retrieval architecture. Its demonstrated contribution is a working mobile pipeline in which corpus persistence, embedding inference, semantic-anchor scoring, ranking, and evidence inspection operate locally once the corpus has been prepared.

Its broader research direction is to investigate automatically derived, local, and reproducible common representations between questions and source content that do not require generative interpretation. Multiple representations per chunk are one possible strategy rather than the objective itself. This direction remains to be tested through stronger baselines, component-level evaluation, multilingual experiments, mobile runtime validation, and documented failure analysis.

The product direction is equally clear: preserve local control and inspectable source evidence, reduce dependence on external preparation services, simplify retrieval settings for ordinary users, and move from a research prototype toward a scalable application without obscuring how evidence is selected.