

pdf2beamer: A Throttled, Resumable Pipeline from PDF Pages to Beamer

Shyamal Chandra

17 October 2026

Copyright (C) Shyamal Chandra, 2026. All rights reserved.

Abstract

pdf2beamer converts a PDF into a Beamer deck without handing the machine over to the conversion. A **zsh** orchestrator rasterizes or extracts each page, asks a local OCR model for reading-order text, drops running furniture, distills each remaining paragraph into bullets, polishes the wording on device, and emits formula-safe Beamer. The worker is an Objective-C++ program linked against Apple NaturalLanguage and libcurl. Scheduling stays at utility quality of service and reserves two hardware threads for the foreground. Finished pages and slides are checkpointed beside the output, so an interrupted run continues from the last completed unit. A single 343-page preprint is reported as an experience note, not as a benchmark: after furniture suppression the deck compiled to 974 pages, down from 1,333 pages in the unfiltered deck, and the author remained only on the title page.

1 Introduction

A long PDF is a poor slide deck. Pages mix prose, headings, formulas, running headers, author lines, dates, and bibliography entries. A presenter needs the claims, in order, as short bullets, with mathematics that **pdf_lat_ex** can still typeset. Doing that by hand does not scale past a few dozen pages. Doing it by sending every page through a large model, at full machine priority, freezes the workstation that the operator is trying to use.

pdf2beamer is a command-line pipeline for that job. It is aimed at a single macOS workstation with a local Ollama server, not at a cluster. The design has four obligations that are easy to break independently and hard to keep together:

1. Preserve reading order and the author’s claims.
2. Keep formulas as TeX, including formulas the summarizer drops or mangles.
3. Keep author, date, title repeats, correspondence, and citation stubs off the body of the talk.
4. Leave the foreground interface responsive, and survive a crash without redoing finished pages.

This paper records the engineering choices that meet those obligations. It is a system description. Section 9 reports one document processed on one machine. It is not a comparative evaluation.

2 Operating constraints

The implementation target is macOS with the Xcode SDK, Ghostscript, ImageMagick, Poppler or Xpdf text tools (`pdftotext`, `pdftotext`), and TeX Live’s `pdflatex`. OCR and distillation go to Ollama at <http://127.0.0.1:11434> unless the operator passes another host. The default OCR model is `glm-ocr`, which follows the fixed prompt `Text Recognition:` and does not reliably obey a free-form instruction to skip headers. Furniture is therefore removed after recognition, not inside the OCR prompt.

The practical summarizer on the development machine is `gemma3:12b`. The name `gemma4:13b` is not an Ollama tag. When `-model` is omitted, the script selects the first installed model in a fixed list: `gemma4:13b`, `gemma4:12b`, `gemma3:12b`, `gemma4:e4b`, `gemma4:26b`, `gemma4:31b`, `gemma3:1b`. It prints the substitution when the chosen tag is not `gemma4:13b`.

The scheduler reads `hw.logicalcpu` and `hw.physicalcpu`. If logical CPUs exceed physical CPUs, the pool may use hyper-threads. On the development machine they are equal: 28 hardware threads and no simultaneous multithreading. In every case the pool leaves two threads unused so WindowServer and the foreground session keep priority. The process asks `taskpolicy` for utility quality of service, or for background or maintenance if the operator says so, and renices itself by five. ImageMagick, OpenMP, vecLib, and OpenBLAS are pinned to one thread. Ghostscript renders with one rendering thread. Polish uses at most four processes of at most two C++ threads, and those threads set their own utility QoS.

3 Architecture

Two programs share the work. `pdf2beamer` is the orchestrator. It parses arguments, checks tools and models, creates the checkpoint directory, and runs a process pool. `d2s` (`bin/d2s`, built from `src/d2s.mm`) performs OCR calls, segmentation, distillation, polish, and Beamer assembly. The shell never parses model JSON beyond what the worker prints. The worker never decides how many jobs to run. The operator reference is the manual page `pdf2beamer(1)`. A Homebrew install places it on the default man path. From a source checkout, `man ./man/pdf2beamer.1` reads the same page.

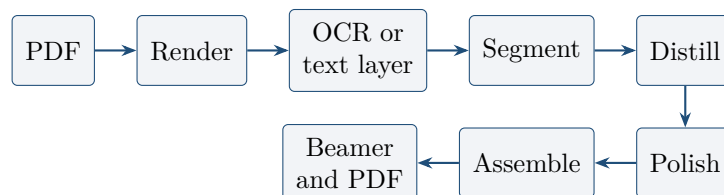


Figure 1: Stage order. OCR completes before distillation, so one model stays resident at a time. Assembly always runs again on resume. PDF compilation, when requested, always runs again.

Stages are barriers, not a free pipeline. All pages are read before any paragraph is distilled. That keeps `glm-ocr` hot during recognition and the summarizer hot during distillation, and it gives segmentation a whole-document view of repeated headers. The shell records a stage name (`read`, `segment`, `distill`, `polish`, `assemble`, `pdf`) and a stamp per finished unit.

Slots are directories created with `mkdir`. A worker takes a slot by creating it and releases the slot by removing it. The shell does not enable monitor mode: under `emulate -R zsh` that option cannot be changed, and an `EXIT` trap inside a worker would also hide the worker’s real status. Each background command’s status is collected with `wait`. Local variables never use the names `path` or

`status`, which are special in `zsh`. `typesetsilent` is set so that `local` declarations do not leak into the Beamer file.

4 Page acquisition

With the default path, Ghostscript renders one page to a 24-bit PNG at the requested resolution. The resolution flag must be a single token, `-r120` rather than `-r` and `120` as separate arguments. Ghostscript treats a separate argument as an input file. ImageMagick then shrinks the long edge to at most 1,536 pixels by default, between 512 and 4,096 if overridden. The PNG is base64-encoded and posted to Ollama `/api/generate` with temperature 0, a 32,768-token context, and `keep_alive` of two minutes. If `generate` fails, the same image is sent to `/api/chat`. A reply that is only the recognition prompt is discarded. An empty OCR result falls back to `pdftotext` for that page, and a note is written for the operator.

`-text-layer` skips rasterization and OCR and reads the embedded text directly. `-dpi` is accepted from 36 to 300. These choices trade recognition quality against memory and time. They are stored in the checkpoint identity so a resume cannot silently mix two renderings of the same PDF.

5 Segmentation and non-core text

OCR text is not a sequence of slides. The segmenter rebuilds paragraphs and then refuses several classes of text that would otherwise become slides in the middle of the talk.

A blank line ends a paragraph. A following line continues the paragraph only when it completes a hyphenated word, or when the previous line does not end a sentence and the next line starts in lowercase. A newline after `.`, `!`, or `?` followed by a capital therefore starts a new paragraph. That matches both OCR and the text layer, which usually break thoughts that way. A short line is a heading when it is a Markdown heading, begins with “chapter”, “section”, or “part”, ends with a colon, or is shouted in capitals.

Lines that are only page numbers, “page N ”, “ N/M ”, short roman numerals, copyright boilerplate, or a URL with no other words are dropped immediately.

Running matter is counted after case-folding and tokenization. A line of at most 100 characters that appears on at least 40 percent of pages is furniture. A short line, at most 14 words, is also furniture when it appears on at least two pages of a document of three pages or fewer, and otherwise on at least $\max(3, \lceil n/20 \rceil)$ pages. Every copy is dropped, including the first. The title page of the deck already carries the title, the author, and the date.

A further classifier rejects text that is non-core even when it appears once:

- a date, or a year, with no other claim;
- an email, a correspondence line, or a URL, arXiv identifier, or DOI with no other claim;
- a page range or a volume line;
- a close restatement of the document title;
- a line whose content words are only the PDF author;
- a citation stub (“authored”, “published in”, “appeared in the proceedings”, and similar) that contains no separate claim verb;

- a bare personal name or affiliation of two to eight capitalized words, unless a word is a topical heading such as “method” or “results”;
- an echo of the distillation instructions, including the placeholder `<bullet>` and a title that is only “presentation copy”.

Mathematics is never classified as non-core. A bullet that states a claim and then cites a paper keeps the claim. The citation bullet is removed. The same classifier runs again at assembly, because the summarizer will paraphrase “Aditya Cowsik” into “Authors of the Work” and a date into “Publication Year”. Exact duplicate slides are emitted once.

The title and author passed into segmentation are the deck title and the deck author, which default to PDF metadata. That is why `-title` and `-author` change both the title page and the filter.

6 Distillation and polish

Each remaining paragraph is one request, unless it does not fit the context budget. Token count is estimated as one token per three bytes, plus one. The context window sent to Ollama is the estimate of the prompt, the paragraph, and 768 generated tokens, plus a small margin, rounded up to a multiple of 256, and at least 2,048. It is capped by the minimum of the model’s advertised context and `-max-ctx` (default 8,192). The request sets `num_ctx` explicitly. Leaving it unset would reserve the model’s full window.

The prompt requires a title of at most eight words, a kind of **steps**, **process**, or **items**, and one bullet per sentence. It forbids new facts. It requires every formula to be copied in inline or display math, or in its original equation environment, without spelling symbols out or rewriting scripts. Temperature is 0.3 on the first attempt. If the server rejects the **think** field, the request is retried without it. A failed request falls back to on-device sentence splitting. `-no-llm` uses that fallback for every paragraph.

Kind selects the Beamer list. Items become **itemize** under the kicker “Key points”. Steps and processes become **enumerate** under “Steps” or “Process”. A heading becomes a centered section frame. `allowframebreaks` is set only when a frame has more than six bullets. Shorter frames would otherwise display a roman “I”.

Before polish, any formula fingerprint of length at least three that was present in the source and absent from the model output is appended. Polish then runs `NLTagger`, `NLTokenizer`, and `NLLanguageRecognizer`. Contractions expand. A small lexicon upgrades informal verbs (“get” to “obtain”, “start” to “commence”, and others) only when the following content word is a noun, a name, or a determiner, so “get ready” stays. Filler adverbs drop. Non-English text skips the lexicon. Math spans are split out, tagged as the probe noun “value” when a verb upgrade needs a following noun, and restored. All-caps titles become title case. When more than one bullet is present, polish runs on a small `std::thread` pool at utility QoS.

7 Formula-faithful Beamer

The emitted preamble loads `amsmath`, `amssymb`, and `mathtools`, provides `\ldots` as `\ldots`, and selects `professionalfonts` so the theme does not replace math glyphs. The theme name must be alphanumeric, which blocks injection through `\usetheme`. The default theme is Madrid with the seahorse color theme, 16:9, 11 point, and hidden navigation symbols. The subtitle is “Converted from” the PDF file name. The date is `\today`.

Prose is escaped. Math is not, once it has been checked. The splitter recognizes $\$ \dots \$$, $\$ \$ \dots \$ \$$, $\backslash(\dots\backslash)$, $\backslash[\dots\backslash]$, fenced math blocks, and the usual math environments, including their starred forms. A dollar amount whose interior is only digits, commas, dots, and spaces remains text. An identifier such as `keep_the_name` is escaped.

Inside a math span the sanitizer joins a command that OCR split ($\backslash \Sigma$ becomes $\backslash \text{Sigma}$), converts Unicode Greek, operators, superscripts, and subscripts to TeX, and braces a doubled script of the same kind so $10^{2^{-6}}$ is not parsed as two superscripts. A mixed script such as x_0^1 is left alone. $\backslash \text{Ldots}$ becomes $\backslash \text{ldots}$. An unknown letter-command becomes $\backslash \text{mathrm}$. That is how a model invention such as `\g` compiles instead of aborting `pdflatex`. Percent and hash are escaped unless they are already escaped.

Two spans are refused and escaped as ordinary text. The first has unbalanced braces, which would otherwise swallow the rest of the frame. The second contains a command that can read files, write files, change category codes, or redefine macros: `\input`, `\write`, `\def`, `\includegraphics`, verbatim environments, and the rest of a fixed denylist. Moving arguments (frame titles, title, subtitle, author, section) use `\texorpdfstring`. Display math is omitted from the short bookmark text. A slide that is only a formula is titled “Formula” rather than “Notes”.

8 Control plane

The working directory is `<output-basename>.d2s`, next to the TeX file. Its identity file records the absolute PDF path, byte size, modification time, page count, and the options that change bytes on disk: text-layer mode, whether the summarizer is disabled, both model names, resolution, maximum edge, and context cap. A second invocation with the same identity resumes. A mismatch stops the run and tells the operator to pass `-fresh`, which deletes the directory first.

Resume skips stamped OCR pages, existing slide JSON for a paragraph, and polished files that are not older than their source JSON. Segmentation and assembly always run again. They are cheap relative to OCR, and assembly must see the furniture filter even when distillation was resumed from an older worker. If `-pdf` is set, `pdflatex` always runs twice. Auxiliary files (`.aux`, `.log`, `.nav`, `.out`, `.toc`, `.snm`, `.vrb`) are removed after success. A compile failure copies the log to `<output-basename>.build.log` and leaves the TeX in place.

A directory lock stores the orchestrator pid. If that pid is alive, a second run refuses to start. If it is dead, the lock is taken. On success the working directory is deleted unless `-keep-tmp` was passed. On failure after the identity file exists, the directory is kept and the script prints the resume command. Progress is a six-bar redraw on a terminal and plain `progress` lines otherwise. `NO_COLOR` disables color. Block glyphs are used when the locale is UTF-8.

The file creation mask is 077. Page artifacts are zero-padded so lexical order is page order.

9 One document

The development input was the 343-page PDF `2609.30063v1.pdf`. Its metadata title is “Self-Play Pretraining with Zero Data” and its metadata author is Aditya Cowsik. This section describes that file only. It is not a claim about other PDFs.

The first full generation, before furniture suppression was strong enough, produced on the order of 1,324 content frames. Running headers had been paraphrased into repeated slides titled with the author, the date 17 October 2026, and the paper title. Bibliography lines had become slides titled “Authors of the Work”, “Publication Year”, and “URL Identification”. The distillation prompt itself sometimes came back as a slide. Separately, `pdflatex` aborted. The fatal errors, in the order they

appeared, were an undefined `\g`, an unclosed math brace, Unicode subscripts, a doubled superscript $10^{2^{-6}}$, and commands split by a space after the backslash. Each of those is now handled in the emitter described in Section 7. After those repairs the unfiltered TeX compiled to a 1,333-page PDF.

The furniture filter was then applied to that TeX using the same rules as assembly. It removed 357 frames and 482 bullets. The author string remained in `\author` and in no frame. The filtered file compiled, in two `pdflatex` passes, to 974 pages. The original OCR checkpoint for that run no longer existed, so the 343-page document was not re-recognized. Later runs of the same command checkpoint automatically, and segmentation drops the same classes of text before the summarizer is called.

10 Verification

`make test` builds `d2s` and runs `pdf2beamer -selftest`. The worker checks JSON decoding, LaTeX escaping, inline and display math, neutralization of `\input`, rewriting of `\g` and `\Ldots`, unbalanced math, doubled superscripts, currency, polish of verbs next to formulas, Unicode superscripts, heading segmentation, and removal of author, date, correspondence, citation, URL, title-echo, and prompt-echo slides. A core bullet is required to survive beside a citation bullet that must not. The shell test builds a one-page PDF, runs the offline pipeline with `-pdf`, checks that a successful run deletes its checkpoint, then seeds a checkpoint with `-keep-tmp` and checks that a second run prints that it is resuming and deletes the checkpoint after success.

These tests pin the contracts above. They do not measure recognition accuracy, summarizer faithfulness on arbitrary scientific prose, or throughput.

11 Limitations

The furniture classifier is lexical. A short technical heading that happens to be two capitalized words outside the topic list can be treated as a name. A repeated short sentence that clears the page threshold is treated as a header. A citation that also contains a claim verb is kept, which is intentional, and a citation that hides inside a long paragraph may survive until assembly removes only the bullets that match.

OCR quality is the quality of `glm-ocr` at the chosen resolution and edge length. The text-layer path cannot invent text that was never embedded. Context capping can split a long paragraph, and the split is by a byte estimate rather than by the model’s tokenizer. Polish is an English lexicon plus Apple language identification. It is not a general rewriter.

Checkpoint identity does not include the theme, title override, or author override. Changing those and resuming will reuse OCR and distillation and then assemble with the new strings. Changing resolution, models, or the PDF bytes will not resume. Two `pdf2beamer` processes must not share an output path. The lock enforces that only while the first pid is alive.

The emitter’s denylist is not a full TeX sandbox. It is a filter on commands that showed up in model output or that are obvious file and definition primitives. A novel dangerous command outside that list would pass through a math span. Operators who do not trust the model should read the TeX before compiling, or compile with shell escape disabled, which is already how `pdf2beamer` invokes `pdflatex`.

12 Conclusion

`pdf2beamer` is a single-machine pipeline from PDF pages to a Beamer file. It separates orchestration from linguistic work, sizes each summarizer call to the paragraph, restores and sanitizes mathematics, and refuses to turn the author line, the date, and the bibliography into the talk. Utility QoS and a reserved pair of threads keep the conversion off the foreground. Checkpoints make a long document restartable. The 343-page preprint shows why the furniture filter and the math sanitizer are not optional: without them the deck repeats its front matter for hundreds of frames, and a single undefined command stops the compile.

References

- [1] Till Tantau, Joseph Wright, and Vedran Miletić. *The `beamer` class*. CTAN package `beamer`.
- [2] Leslie Lamport. *L^AT_EX: A Document Preparation System*. Addison-Wesley, 2nd edition, 1994.
- [3] Donald E. Knuth. *The T_EXbook*. Addison-Wesley, 1984.
- [4] Artifex Software. *Ghostscript*. <https://www.ghostscript.com>.
- [5] Ollama. *Ollama*. <https://ollama.com>.
- [6] Apple. *NaturalLanguage*. Apple developer documentation.