Demo Walkthroughs

Dr Patrick Chu • Independent AI Consultant — Academic Research Pipelines • patrickchu.net • +852 5137 3793

One page per demo: what it proves, what data it runs on, what you will see during a 30-minute walkthrough, what you can try yourself, and the honest limits. Every demo is a working, reproducible pipeline — script + data + report — adaptable to your data in days.

Academic Admin Agent: Inbox Triage, Reference Letters & Meeting Minutes

WHAT THIS PROVES

One AI assistant handles three chores academics dislike most — and shows you exactly where it gets things right or wrong before it touches your real work.

THE DATA

The demo uses ten hand-written realistic academic emails, one fictional student CV (for Chan Hoi Yan, an MPhil Linguistics candidate), and a transcript of one fictional departmental meeting — all synthetic and modelled on a Hong Kong university context. Because the data is fixed and self-contained, results are stable across runs; a real deployment using your own emails would need two to four weeks of tuning and results would vary.

WHAT YOU'LL SEE

1. Ten academic emails load on screen, each with a sender, subject, and body.
2. The agent labels every email — urgent, reply, delegate, or ignore — and suggests a one-line action.
3. Results appear beside gold labels; all 10 categories matched, but priority levels were only half right.
4. A reference letter for the MPhil student appears; a coverage check confirms every CV fact was included, and three soft character claims are flagged as unverified.
5. Paste-in meeting notes are processed; three of four decisions and four of five action items surface, each with owner and deadline.
6. A summary panel shows where the agent succeeded and where you should review before acting.

TRY IT YOURSELF

Paste one of your own real emails (with any names changed) into the triage box and watch it classify and suggest an action — then judge whether you agree.

LIMITATIONS

All three tasks run on ten emails, one CV, and one meeting — a proof of concept, not a tested system. Gold labels reflect one person's judgment, so accuracy figures should be read as illustrative ranges rather than firm benchmarks. The agent drafts and flags; it does not send emails or submit forms — integration with Outlook or university portals is a separate engineering project not covered here. Priority scoring was only 50% accurate on this small set, so treat priority suggestions as a prompt to check, not a decision.

AI Speech Screening for Cantonese Children with Speech Sound Disorders

WHAT THIS PROVES

Can a computer hear a child's mispronunciation the way a speech therapist does?

THE DATA

Ten target words were recorded using synthetic Cantonese child and adult voices, each word embedded in a standard picture-naming phrase used in real clinical assessments. The voices are computer-generated, not real children — this is a mechanism test, not a clinical dataset. Because the errors and targets are built in by design, we know the exact right answer for every item, which makes the accuracy figures fully auditable.

WHAT YOU'LL SEE

1. The demo loads two voices — one adult, one child — reading the same Cantonese passage.
2. Both recordings are fed to a speech-recognition engine; transcripts appear side by side on screen.
3. Ten child-voice clips play, each containing a deliberate mispronunciation typical of speech sound disorder.
4. The transcripts appear; you see whether the mispronunciation survived into the text or was silently corrected.
5. An AI language model reads each transcript and names the disorder pattern — fronting, stopping, etc.
6. A results summary appears: 7 out of 10 disorder patterns correctly identified, 8 out of 10 errors confirmed present.

TRY IT YOURSELF

Point to any word in the results table and ask 'why did the system mark this one unclear?' — the per-item evidence is on screen and the consultant can walk through the reasoning live.

LIMITATIONS

All voices are synthetic, so the demo proves the pipeline works in principle — it does not yet measure how well it handles real children's voices, which are less precise and more variable. Ten items is a pilot; a clinically meaningful evaluation would need 100 or more items checked against human speech therapist judgements. The speech-recognition engine occasionally 'corrects' a mispronunciation back to the standard word before the disorder detector even sees it, which is a known technical hurdle this demo is designed to expose, not hide.

Back-Translation: Building a Bilingual Questionnaire

WHAT THIS PROVES

See how a questionnaire is translated, checked, and refined to match the published Chinese version.

THE DATA

The demo uses 11 public-domain items from the GAD-7 and PHQ-9 scales. It compares three translation drafts (plain, glossary-guided, and colloquial Cantonese) against the official Traditional Chinese versions. The data is synthetic in the sense that the drafts are generated by an AI, and results may vary slightly with each run.

WHAT YOU'LL SEE

1. You see the original English items and three Chinese drafts.
2. The pipeline runs forward translation and blind back-translation.
3. An AI judge scores each draft for meaning preservation.
4. A flagged item is automatically refined and re-checked.
5. A deliberately wrong translation is caught and corrected.
6. Final results show which draft best matches the published version.

TRY IT YOURSELF

None — the demo is a fixed pipeline.

LIMITATIONS

The AI acts as both translator and judge, so this validates the workflow, not a finished instrument. Results are ranges, not exact numbers, because the AI is stochastic. The comparison with published Chinese is indicative, as regional variants exist.

Classroom Discourse Analysis Pipeline

WHAT THIS PROVES

From a recorded lesson to a coded transcript — automatically, in minutes.

THE DATA

The demo uses a short synthetic Cantonese maths lesson (under a minute, 10 exchanges between one teacher and one student) recorded in a clean, controlled way so the pipeline always runs the same way live. Because it is synthetic, results are stable for this demo; a real classroom recording would introduce background noise, overlapping voices, and children's accents, so accuracy figures should be treated as a best-case range rather than a fixed number.

WHAT YOU'LL SEE

1. A short Cantonese maths lesson audio file loads — you see the waveform and duration.
2. The pipeline transcribes the audio; a turn-by-turn transcript appears on screen in Cantonese.
3. Each turn is automatically labelled: Teacher or Student, and a discourse code (Instruction, Question, Response, or Feedback).
4. A summary panel shows the teacher-talk ratio (roughly two-thirds) and a breakdown of question, response, and feedback turns.
5. A reliability table appears comparing the automatic codes against a hand-written gold standard — about two-thirds of discourse codes matched, which is moderate agreement for a fully automated first pass.

TRY IT YOURSELF

None — the demo is a fixed pipeline. However, you are welcome to bring a short audio clip or transcript from your own classroom data; the consultant can show how it would slot into the same pipeline after the live run.

LIMITATIONS

The audio is synthetic — two clean, adult voices with no overlap — so accuracy will be lower on real classroom recordings with children's voices, noise, and cross-talk. Discourse codes are illustrative; a real study would develop the codebook together with your research team. Treat the two-thirds code-agreement figure as an upper bound, not a guaranteed result.

EEG Brain-Signal Analysis: From Raw Data to Brain Maps

WHAT THIS PROVES

Watch a full brainwave analysis pipeline run in under two minutes — cleaning, averaging, and visualising neural responses to sound and vision.

THE DATA

The demo uses a standard open-access teaching dataset: a 60-channel EEG recording of a person responding to auditory and visual cues, roughly two and a half minutes long. It is a real (not synthetic) recording widely used in neuroscience training worldwide, so results are stable and repeatable across runs. Because this is a single-subject tutorial file rather than a full study dataset, the numbers shown are illustrative benchmarks, not publishable group statistics.

WHAT YOU'LL SEE

1. The raw brainwave recording loads and its 60 channels appear as scrolling coloured traces.
2. The pipeline filters out electrical noise and eye-blink artefacts automatically; a clean signal remains.
3. The recording is sliced into ~200 brief trial windows, one per stimulus event; 90% pass quality checks.
4. Averaged waveforms (ERPs) appear for auditory vs. visual conditions, showing distinct timing and scalp location differences.
5. A scalp topography map lights up showing where brain activity peaks for each condition at each moment in time.
6. A time-frequency colour map reveals how oscillatory brain power rises and falls around each visual stimulus.

TRY IT YOURSELF

None — the demo is a fixed pipeline. However, the researcher is welcome to point to any panel on screen and ask 'what does this mean for a reading/language study?' for a tailored explanation.

LIMITATIONS

This is a single participant from a tutorial dataset, so no group statistics or individual differences are shown — a real study would add per-participant tuning and mixed-effects models. The pipeline settings (filters, epoch length, artefact thresholds) are reasonable defaults, not optimised for a specific research question; a live project would involve a pre-registered analysis plan. Results should be read as a proof-of-concept demonstration of the workflow, not as findings.

Automated Reaction-Time Analysis Pipeline

WHAT THIS PROVES

From raw experiment data to a draft Results section — in minutes, with every step auditable.

THE DATA

The demo uses a simulated word-response experiment: 32 participants each completing 80 trials, with a known 45 ms speed difference built in so the pipeline can be checked against a correct answer. The data are synthetic (not from a real study), designed to mimic standard cognitive-psychology experiments. Because the data are fixed for this demo, results are stable — a live pipeline on real collected data would produce ranges rather than single numbers.

WHAT YOU'LL SEE

1. Raw trial-by-trial response times load and display on screen.
2. The pipeline automatically removes invalid trials (errors, extreme outliers).
3. Group averages appear: related-prime responses average 45 ms faster than unrelated.
4. A significance test and effect-size score are calculated and shown instantly.
5. An AI-drafted Results paragraph appears, ready for the researcher to edit.

TRY IT YOURSELF

Paste in a short description of your own experiment design and ask the chatbot how it would handle your trial structure or exclusion rules.

LIMITATIONS

The data are simulated, so the pipeline has not yet faced the messier trial structures of a real study — though the steps carry over directly. Only a simple two-condition comparison is shown here; real multi-experiment papers typically need more complex models, which are an extension not yet in this demo. The AI-drafted paragraph is a writing aid and starting point, not a submission-ready text.

Eye-Tracking Reading Analysis: Automated Measures from Raw Gaze Data

WHAT THIS PROVES

Upload raw eye-tracking output, get publication-ready reading measures and a drafted results paragraph — automatically.

THE DATA

The demo uses simulated Chinese sentence-reading data (30 readers, 40 sentences each) with known word-frequency differences built in, so the pipeline can be checked against a correct answer. This is synthetic validation data, not a real corpus; real deployment would use your lab's own recordings or public datasets such as ZuCo or GECO. Because the demo data is fixed, the numbers you see today will be stable — but results on live lab data will naturally vary across runs and participant samples.

WHAT YOU'LL SEE

1. Raw fixation records load: one row per word per participant, nothing else.
2. Pipeline calculates five standard reading measures for every target word automatically.
3. Results appear: low-frequency words take longer to read on every single measure.
4. A summary paragraph — ready to paste into a manuscript — is drafted instantly.
5. You compare the pipeline's recovered differences to the known built-in effects as a quality check.

TRY IT YOURSELF

Bring a plain spreadsheet of your own word-level fixation exports and ask the consultant to swap it in — the pipeline will run on your data with no code changes needed.

LIMITATIONS

The data today is synthetic and small (30 simulated participants), so this validates that the measures are computed correctly, not that the norms match any real population. The pipeline does not yet include mixed-effects models, which are the current statistical standard for trial-level eye-tracking data — the paired t-tests shown are a useful first pass only. The AI-drafted paragraph is a starting point; authors must review it before submission.

Grant Proposal Alignment Check

WHAT THIS PROVES

See how a draft proposal measures up against official RGC criteria before you submit.

THE DATA

The demo uses a clearly labelled synthetic summary of a Hong Kong adolescent sleep and well-being study, written in the style of a typical RGC GRF application. It is not a real proposal. Because the underlying model and live data can vary, results are shown as ranges, not single fixed numbers.

WHAT YOU'LL SEE

1. Start the demo; the synthetic proposal summary loads.
2. The pipeline runs automatically, scoring the summary against official RGC criteria.
3. A colour-coded matrix appears, showing scores and evidence quotes per criterion.
4. Gaps and revision suggestions are listed below the matrix.
5. A coverage table shows which form sections the summary addresses.
6. You can hover over any score to see the exact wording from the RGC guidance.

TRY IT YOURSELF

None — the demo is a fixed pipeline.

LIMITATIONS

The demo uses a synthetic summary, not a real application, so scores reflect only what that summary evidences. The scoring model is a proxy reviewer and can vary slightly between runs; the keyword baseline is simple. Always check the current RGC guidance, as wording may change.

Literature Matrix: ChatGPT & Generative AI in Higher Education

WHAT THIS PROVES

One automated pass turns 11 real papers into a structured literature matrix with themes, gaps, and a reliability check.

THE DATA

The demo uses 11 real, open-access, top-cited papers on ChatGPT in higher education (2023–2026), fetched from OpenAlex and pinned for reproducibility. The sample is small and selective (a top-cited sample, not an exhaustive search), and because live data can drift, results are reported as ranges rather than single numbers.

WHAT YOU'LL SEE

1. Start: the pipeline loads the pinned corpus of 11 papers.
2. Watch as each abstract is read and structured into a matrix row.
3. See the design column cross-checked against a keyword baseline.
4. Themes and gaps appear automatically from the matrix.
5. A one-page summary shows key numbers and limitations.
6. You can click any cell to verify it against the abstract.

TRY IT YOURSELF

None — the demo is a fixed pipeline.

LIMITATIONS

Extraction is from abstracts only, not full texts, so some details are thinner. The LLM is stochastic but pinned to a cache for reproducibility; observed agreement with the baseline ranges from 0.76 to 0.88. The corpus is a small, top-cited sample, not an exhaustive review.

Translation with Terminology Memory: Does Giving the AI an Official Glossary Actually Help?

WHAT THIS PROVES

A working demo of a simple idea: feed a list of official Hong Kong government terms into an AI translator and measure whether it actually uses them correctly.

THE DATA

14 Chinese-language government press-release sentences, hand-authored to include a mix of everyday official terms (e.g. 教育局, 立法會) and rarer policy-specific terms (e.g. 簡約公屋 → Light Public Housing, 明日大嶼 → Lantau Tomorrow Vision). The data is small and synthetic — it is a mechanism demo, not a full study. Because the underlying AI model can be updated at any time, exact numbers may shift slightly between runs; the demo quotes ranges from repeated runs, not single best scores.

WHAT YOU'LL SEE

1. The demo loads 14 press-release sentences and a 22-term official glossary on screen.
2. Each sentence is translated twice: once with no glossary, once with the glossary injected.
3. Results appear side-by-side; correct and incorrect term choices are highlighted per sentence.
4. A summary panel shows overall terminology accuracy: 80–90% without the glossary, 100% with it.
5. You can scroll to sentences 13–14 to see where the glossary makes the clearest difference — rare policy terms.

TRY IT YOURSELF

Paste in one of your own Chinese government sentences and watch both translations appear alongside the term-hit highlights.

LIMITATIONS

Only 14 sentences — enough to demonstrate the mechanism, not to publish. A real study would need 100+ sentences with professional translators setting the gold standard. The quality scores (fluency, adequacy) are generated by an AI judge, not a human, and both translation versions are already fluent, so those scores are close and should not be over-interpreted. Results are ranges across runs, not guaranteed single figures.

Mental-Health Screening Chatbot with Safety Rails

WHAT THIS PROVES

A Cantonese-language chatbot that spots student distress, always refers a crisis to a real hotline, and never plays therapist.

THE DATA

Fifteen scripted Cantonese student messages, split evenly across everyday chat, moderate distress, and crisis statements (including self-harm and suicidal ideation). The messages are purpose-written for this demo, not drawn from real student records — this is a proof-of-concept mechanism test, not a clinical study. Because the scripts are fixed, results are stable across runs; a live deployment with real users would behave differently and would need proper ethics-approved evaluation before any conclusions could be drawn.

WHAT YOU'LL SEE

1. Fifteen Cantonese student messages load on screen, labelled everyday, distress, or crisis.
2. The chatbot replies to each message in Cantonese with an empathetic response.
3. A separate automated judge reads each reply and scores distress level, empathy, hotline presence, and whether handoff was recommended.
4. A summary panel appears: 15/15 distress levels correctly identified, all 5 crisis messages triggered a real hotline number and a human-handoff recommendation.
5. You can scroll the full conversation log and see how the tone shifts — warm reassurance for everyday chat, gentle concern for distress, immediate safety referral for crisis.
6. A limitations card is shown last, reminding you that scripted demos are not clinical trials and a real service needs further validation.

TRY IT YOURSELF

Type your own Cantonese sentence into the chatbot input box during the demo — try something everyday ('今日天氣好好') and then something that sounds distressed, and watch how the response and safety flags change.

LIMITATIONS

All 15 messages were written by the research team, so the chatbot has never been tested on real students; strong results here confirm the design works in principle, not in practice. The automated judge is itself an AI, so empathy scores are indicative, not clinically validated — the one hard check is a simple text-search confirming the real hotline number appears in every crisis reply. This is a triage-and-handoff tool only: it does not diagnose, treat, or counsel, and any live deployment would need ethics approval, clinician review of the scripts, and verification of local crisis resources.

Can a Published Hong Kong Study Be Reproduced? A Live Replication Demo

WHAT THIS PROVES

The pipeline re-runs the core statistical analyses from a 2022 PLOS ONE study on child-care worker burnout — and checks how closely the numbers match what was published.

THE DATA

The dataset is the publicly released SPSS file attached to the original paper (open licence, CC-BY 4.0): 381 children living in Hong Kong residential care homes, assessed repeatedly over two years by 76 care workers. Because this is a fixed, archived dataset — not a live feed — the numbers are stable and will look the same every time the demo runs.

WHAT YOU'LL SEE

1. The published SPSS data file loads automatically; row and column counts appear on screen.
2. The pipeline reproduces the study's sample descriptives — 76 workers, 381 residents, stay-length split.
3. Growth-curve models run; the screen shows behavioural-problem trajectories matching the paper's two-phase decline.
4. Burnout regressions appear; standardised effect sizes land close to the published values.
5. A summary panel highlights which findings reproduced, which are slightly attenuated, and the one flagged discrepancy.

TRY IT YOURSELF

None — the demo is a fixed pipeline.

LIMITATIONS

The dataset is small (76 workers) and covers only one sector in Hong Kong, so results should not be generalised broadly. The replication used a standard multilevel approach rather than the original software (Mplus), which means slope values look numerically different until a simple scaling adjustment is applied — the direction and significance still match. One phase-2 effect appeared marginally significant in the replication but not in the paper; this is flagged honestly rather than explained away. Effect-size ranges, not single 'best' numbers, are the appropriate way to read the burnout results (roughly 25–33% of variance explained).

Automated Questionnaire Validation & Draft Write-Up

WHAT THIS PROVES

Run your survey data through one pipeline and get reviewer-ready statistics plus a first draft of your Methods and Results — in under two minutes.

THE DATA

The demo uses a fully synthetic dataset: 240 simulated respondents, 12 questions, answered on a 1–5 scale, with a known three-group structure built in from the start. Because the correct answer is known in advance, we can prove the pipeline finds it before anyone trusts it with real data. A real deployment would use your own survey responses instead.

WHAT YOU'LL SEE

1. The pipeline loads 240 simulated survey responses across 12 items.
2. Reliability scores appear for each of the three subscales — all comfortably above the accepted 0.70 threshold.
3. Item-level statistics show how well each question pulls its weight in its subscale.
4. A factor analysis runs and assigns every item to a group — 100% match to the known structure.
5. The screen displays a ready-to-edit Methods paragraph and Results paragraph, written in plain academic English.

TRY IT YOURSELF

Paste in a short list of your own item labels and watch the AI redraft the Methods paragraph using your actual scale name and subscale names.

LIMITATIONS

The data here is synthetic and tidy — real surveys bring missing responses, scale bunching at one end, and cross-cultural wording effects that can lower these numbers. The factor analysis shown is exploratory only; a journal submission would also need a confirmatory stage. The AI-generated paragraphs are a first draft: the interpretation, emphasis, and final wording must come from you as the author.

Detecting AI-Generated Text vs. Human Writing

WHAT THIS PROVES

A live, auditable research pipeline that tests two AI-text detection methods side-by-side — matching the exact questions the detection literature is asking.

THE DATA

The demo uses 20 short passages: one genuine human excerpt (from Wikipedia) and one AI-generated passage on the same topic, repeated across 10 Hong Kong-relevant topics (e.g. Cantonese, MTR, Hong Kong cuisine). Because the labels are known by construction — we made the AI passages ourselves — no human annotation is needed to check whether the detectors are right. The texts are a small proof-of-concept set; a real study would require hundreds of passages across many writing styles and AI models.

WHAT YOU'LL SEE

1. The 20 passages load on screen, each tagged with its true human-or-AI label.
2. The pipeline extracts writing-style statistics (word length, punctuation patterns, vocabulary variety) for every passage.
3. A simple classifier trains on 19 passages and predicts the 20th — repeated for all 20; final score appears: 16 out of 20 correct.
4. An AI judge reads each passage blindly and gives a verdict with a short reason; per-text results populate a results table.
5. A side-by-side summary appears: the style classifier got 16/20 correct; the AI judge got all 10 topic-pairs right when comparing the two passages directly, but only 11/20 when judging each passage in isolation.
6. The researcher is invited to inspect any row — clicking a topic shows the original passage, the judge's stated reason, and whether the verdict was right or wrong.

TRY IT YOURSELF

Point to any topic in the results table and ask 'why did the judge get this one wrong?' — the demo shows the judge's exact reasoning for that passage. If time allows, suggest a Hong Kong topic you care about and we can generate a new paired example on the spot.

LIMITATIONS

This is a 20-text mechanism demo, not a benchmark — results would shift with a larger or more varied dataset. The style classifier (16/20) is a weak-but-real signal; the AI judge is stronger at paired comparison (10/10) but unreliable on single passages (11/20), and it tends to flag encyclopedic human writing as AI. Detection accuracy is also a moving target: newer AI models write less predictably, so these numbers reflect today's AI output style and should be treated as indicative ranges, not fixed figures.

LLM-Assisted Qualitative Coding with Reliability Check

WHAT THIS PROVES

An AI reads open-ended text, applies your codebook, and reports how closely it agrees with a human coder — using the same reliability standard academic reviewers expect.

THE DATA

Thirty real consumer-complaint narratives from a US financial regulator's public database, standing in for open-ended survey or interview responses. The regulator's own category labels serve as the human-coded ground truth. Because this is a fixed public dataset the results are stable, but if you substituted live or updated text the model's agreement scores would shift — treat any figures as a range, not a fixed number.

WHAT YOU'LL SEE

1. A spreadsheet of 30 short complaint texts loads on screen alongside a five-category codebook.
2. The AI reads each text and assigns one category; a progress indicator shows coding in real time.
3. A summary appears: 27 of 30 texts coded identically to the human — 90% agreement overall.
4. A reliability score (kappa ≈ 0.73) is displayed; a plain-English label reads 'substantial agreement'.
5. Three disagreements are highlighted; clicking one shows the text, both codes, and a note on why definitions may overlap.
6. A clean CSV of all coded items is ready to download for further analysis.

TRY IT YOURSELF

Paste in one or two open-ended survey responses of your own and watch the AI assign a category from the codebook in real time.

LIMITATIONS

The dataset is small (30 items) and comes from one specific domain, so the kappa figure (roughly 0.70–0.80 across runs) is illustrative rather than definitive. Categories with only one or two examples — such as 'credit card' or 'debt collection' here — can show wide swings in agreement by chance. A real study would need a larger, domain-matched sample and a codebook refined through the disagreement-review step shown in the demo.

Automatic Transcription & Prosody Analysis for Cantonese Speech

WHAT THIS PROVES

Turn a Cantonese audio recording into a word-by-word transcript — with pitch and loudness measurements — in under a minute, no manual annotation required.

THE DATA

The demo uses a short synthetic Cantonese passage (~39 seconds, about 110 words) generated by a text-to-speech voice, so results are perfectly reproducible for a live showing. Real interview or corpus recordings can be substituted directly — the pipeline handles them the same way. Because synthetic speech is cleaner than natural conversation, pitch and timing figures on real recordings will vary; treat any numbers as illustrative ranges, not fixed benchmarks.

WHAT YOU'LL SEE

1. A Cantonese audio clip loads and plays — you hear the sentence spoken aloud.
2. The pipeline transcribes the audio; a full Chinese-character transcript appears on screen.
3. Each word lights up with its precise start and end time in the recording.
4. A results table appears showing pitch (F0) and loudness for every word.
5. A summary line confirms 110 words processed, all with pitch measurements, at ~2.9 words per second.

TRY IT YOURSELF

Paste in a different Chinese sentence or short paragraph and watch the pipeline re-run on it — the new transcript and pitch table appear within seconds.

LIMITATIONS

The audio here is synthetic (computer-generated speech), so it is unusually clean; natural recordings — with background noise, overlapping speakers, or heavy dialectal variation — will produce messier pitch tracks and occasional transcription errors. Timing is at the word level, not the individual sound level, so fine-grained phoneme work would need an extra alignment step. Pitch figures are raw and would require speaker-specific tuning before appearing in a publication.

AI-Assisted Systematic Review Screening

WHAT THIS PROVES

An AI reads research abstracts, decides what to include or exclude, and shows its reasoning — letting you audit every decision before it enters your review.

THE DATA

Sixteen short research abstracts were written by hand for this demo: eight that genuinely fit the review question (technology-enhanced feedback on school students' writing) and eight that do not, covering common exclusion reasons such as wrong age group, wrong subject, or no real study data. The dataset is synthetic and intentionally clean; real-world records are messier, and a proper validation would need 100+ abstracts checked by two independent human raters. Because this is a fixed pilot set, the headline numbers are illustrative ranges rather than a certified benchmark.

WHAT YOU'LL SEE

1. The review question and inclusion criteria (PICOS) load on screen — no coding needed.
2. Sixteen abstracts appear one by one; the AI reads each and gives an Include or Exclude verdict.
3. Every verdict is accompanied by a plain-English reason you can read and challenge.
4. A PRISMA-style flow diagram builds live, showing how many records survived each stage.
5. A summary panel shows the AI agreed with the gold-standard labels on 15 of 16 abstracts.
6. Click any row to compare the AI's reason with the human gold label side by side.

TRY IT YOURSELF

Paste the abstract of one of your own papers into the input box and watch the AI decide whether it meets the stated criteria — then ask it why it excluded or included it.

LIMITATIONS

This is a 16-abstract pilot built on hand-written, unusually tidy records — not a real-world benchmark. Gold labels come from a single rater, not the two independent reviewers a real review requires. The one missed study (a meta-analysis mis-labelled by the demo) shows the AI is not error-free; recall figures in production will likely be lower and should be reported as a range. A live deployment would also need deduplication and full-text screening stages not shown here.

At-Scale Text & Document Intelligence

WHAT THIS PROVES

Turn thousands of reviews, reports, or policy documents into validated sentiment scores, structured themes, and a plain-English executive summary — automatically.

THE DATA

The demo uses 2,000 real IMDB consumer reviews drawn from a publicly available dataset, with a 120-review sub-sample that carries human-verified sentiment labels used to check the AI's accuracy. Because this is a fixed public dataset the results are stable for today's walkthrough; if the same pipeline were pointed at live social-media or regulatory feeds, scores would shift over time and should be treated as ranges rather than single fixed numbers.

WHAT YOU'LL SEE

1. The dataset loads: 2,000 real reviews appear on screen, raw and unprocessed.
2. The AI reads each review and labels it positive or negative; accuracy checks at 96% against human labels.
3. The topic model groups all 2,000 reviews into six themes automatically — no manual coding.
4. A sentiment score is attached to each theme, revealing which topics drive praise or complaint.
5. An AI-drafted executive summary appears, translating the theme table into plain business findings.

TRY IT YOURSELF

Paste a short paragraph from one of your own student surveys, course evaluations, or policy documents into the summary box and watch the pipeline label its sentiment and assign it to the nearest theme.

LIMITATIONS

The accuracy figure (96%) comes from English-language reviews; Chinese or Cantonese text needs additional language-specific setup before the same reliability applies. Topic names are generated from the most frequent words in each cluster — a human researcher should review and rename them for any publication. The pipeline is validated on a 120-review sample; a full deployment on your own corpus would require a fresh calibration run on that material.

AI-Assisted Essay Scoring & Feedback

WHAT THIS PROVES

An AI reads student essays against your rubric, scores them, and writes actionable feedback — benchmarked against real human raters so you can judge whether to trust it.

THE DATA

The demo uses 25 real student essays drawn from a large public dataset of grade 7–8 persuasive writing, each independently scored by two trained human raters on a 1–6 scale. The data is real, not synthetic, but it comes from a US classroom context — results on Hong Kong student writing or different rubrics would need separate validation before any conclusions are drawn.

WHAT YOU'LL SEE

1. 25 anonymised essays load on screen, each showing two human rater scores.
2. The AI reads every essay against the official rubric and assigns its own score.
3. A results table appears: AI matched human scores exactly 68% of the time.
4. The key comparison appears: human raters agreed with each other at a similar rate, setting the honest benchmark.
5. Click any essay row to see the AI's full written feedback — three specific, rubric-linked comments per essay.
6. A summary panel highlights where AI and human scores diverged most, flagging essays for closer review.

TRY IT YOURSELF

Click on any essay row in the results table and read the AI feedback aloud — ask whether you would find that comment useful if you were the student's teacher.

LIMITATIONS

The 25-essay sample is small, so treat the agreement figures as indicative ranges rather than fixed numbers — a larger local sample would tighten them considerably. All essays here are from US middle-school students writing in English; the AI's scoring reliability on Hong Kong student writing, or against a different institutional rubric, is unknown and would need its own validation study (typically 50–100 essays with two local raters). The AI also produces only a single overall score in this demo; separate scores for content, organisation, and language conventions are possible but not shown here.

How Official Media Frames China's Inclusive Education: A Text-Analysis Walkthrough

WHAT THIS PROVES

Automatically reading 77 news articles to test whether one discourse frame dominates — and how closely a computer coder can match a human one.

THE DATA

The corpus is 77 official English-language news articles on inclusive education in China, drawn from a public archive deposited by the paper's authors on OSF. The texts are real published news items, not synthetic, and the deposit is stable — the same file is retrieved every run. Because the authors' original line-by-line coding sheets were not deposited, the pipeline builds its own independent coding for comparison rather than replaying the authors' exact work.

WHAT YOU'LL SEE

1. The corpus file loads: 77 articles, roughly 104,000 words appear on screen.
2. The pipeline scans every sentence and flags those discussing inclusive-education discourse — about 1,000 sentences highlighted.
3. Top words appearing near 'inclusive education' display: children, schools, quality, promote, concept.
4. Each flagged sentence is automatically labelled with one of four published categories: efforts, consensus, challenges, or other.
5. A comparison panel appears showing the published split alongside the pipeline's own split — 'efforts' leads in both, at roughly 65% (published) and 61% (pipeline).
6. A plain-language summary card appears noting where the two codings agree and where they diverge.

TRY IT YOURSELF

Paste in a short news sentence about disability or special-needs education and ask the demo to classify it — the category label and reasoning appear instantly.

LIMITATIONS

The authors' original sentence-by-sentence coding was never made public, so this demo cannot reproduce their exact numbers — it produces an independent re-coding for comparison only. The two codings use different counting units (the paper counts 520 concordance instances; the pipeline counts roughly 1,000 sentences), which partly explains differences in the minority categories. The automated coder agrees only moderately with a simple keyword check, meaning borderline sentences — especially praise-heavy policy lines — could reasonably be labelled 'efforts' or 'consensus' by different coders, human or machine. Finally, the corpus covers only official-channel English reporting, so findings describe how authorities project inclusive education abroad, not the full range of media coverage.