Faster inference for Hemmingway models

November 2026

Hemmingway models are now much faster.

Our custom inference engine made that possible. As models become more capable, speed and reliability matter more to how useful they are in practice. Speed affects how soon you can start reading and how long you wait for the full response. Those waits add up when you're shaping an outline, trying different openings, turning interview notes into an article, or revising a proposal after feedback.

This benchmark measures speed on text tasks centred on writing.

Hemmingway 1

382 tokens per second

Median speed reported during benchmark testing

Hemmingway 1.5

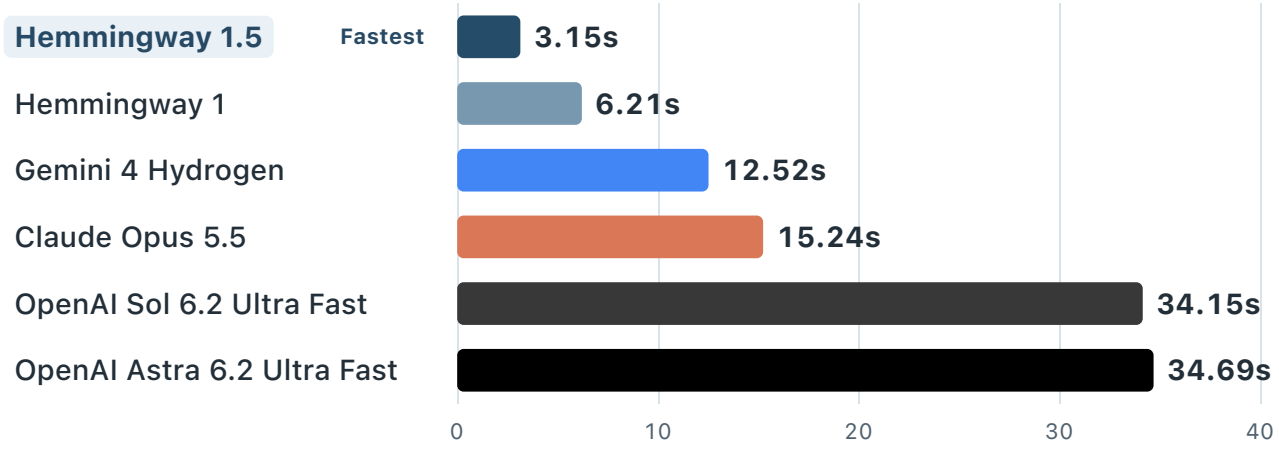
525 tokens per second

Median speed reported during benchmark testing

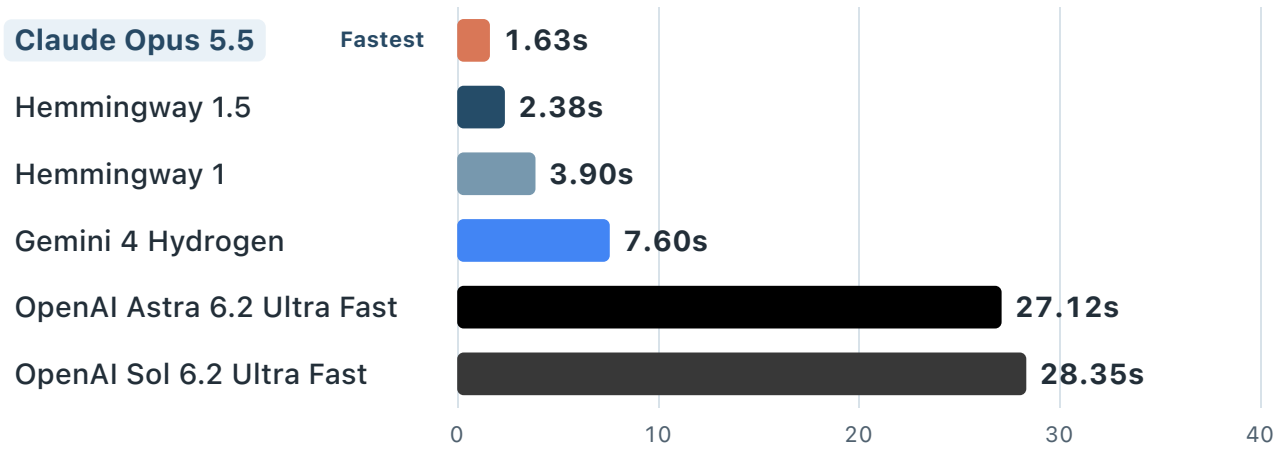
All charts show time in seconds. All models used Medium thinking effort.

Shorter is better.

Median time to finish a response

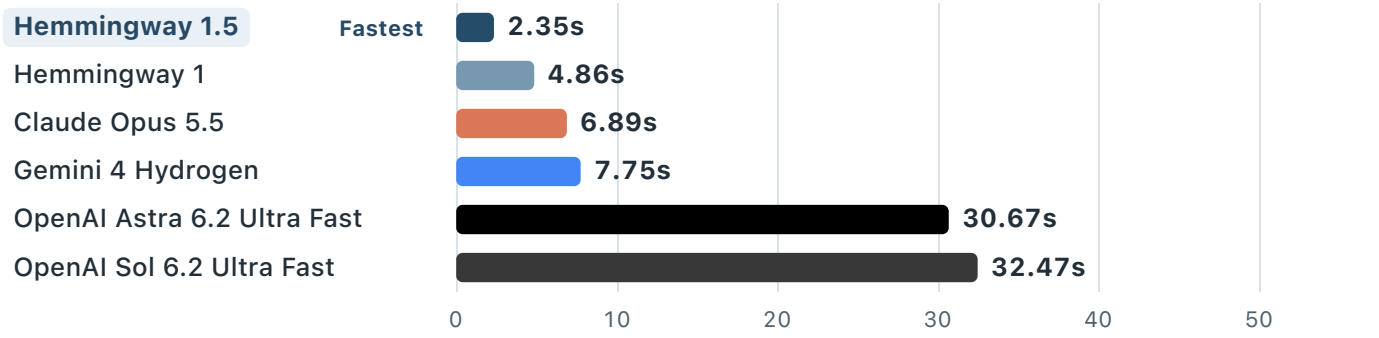


Median time to first token (TTFT)



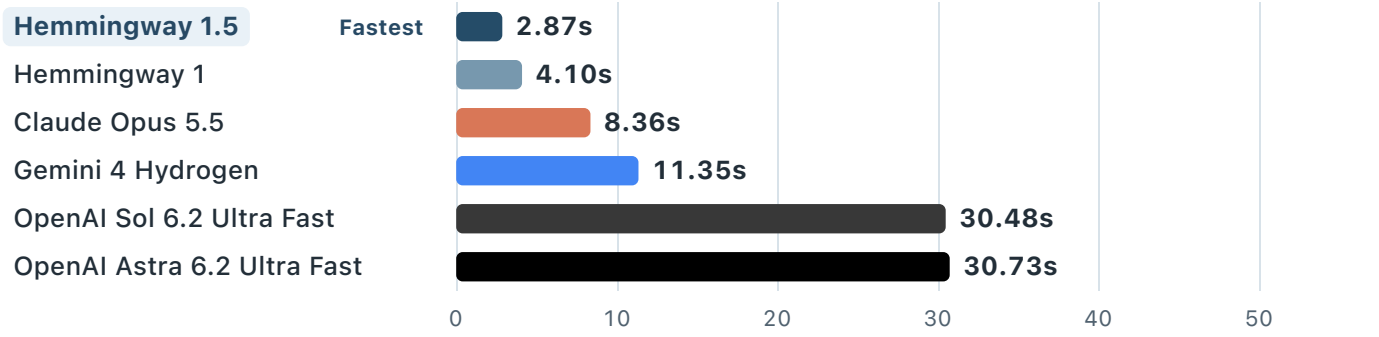
Short responses · median time to finish

Target ≈ 100 words



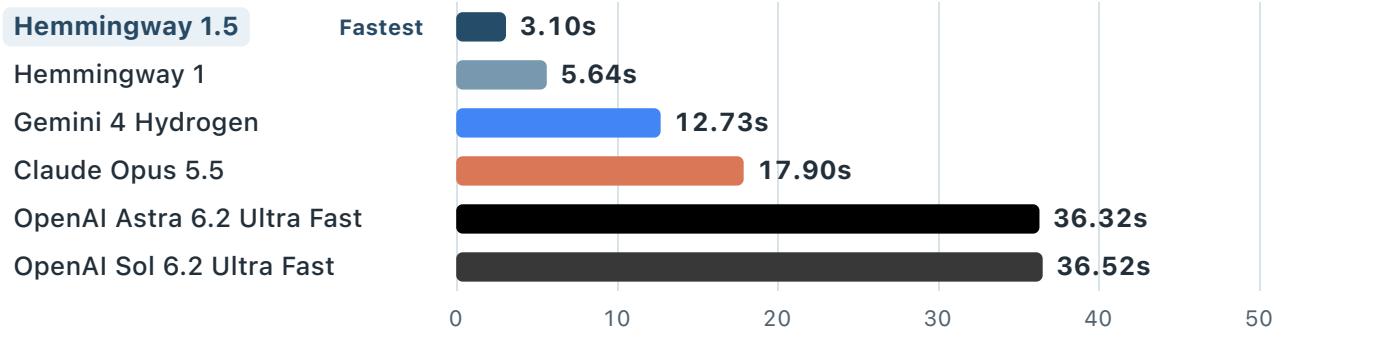
Medium responses · median time to finish

Target ≈ 250 words



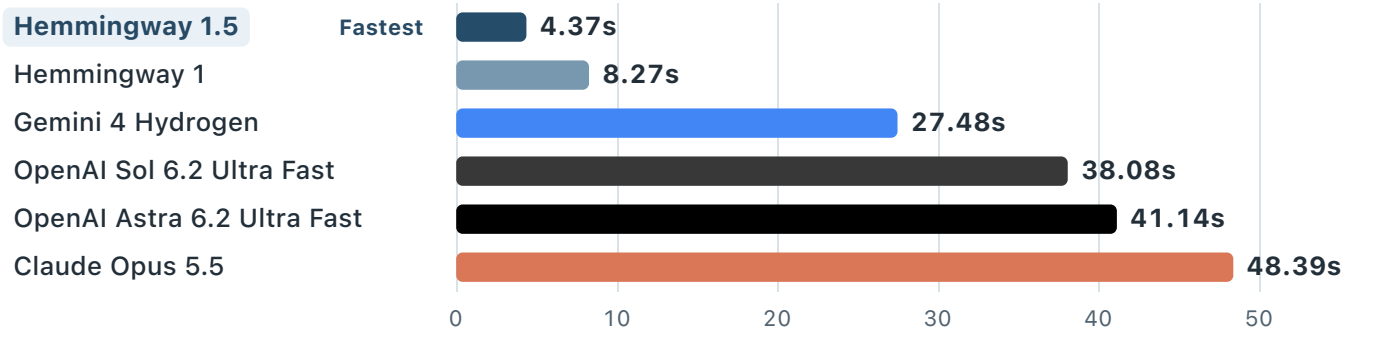
Long responses · median time to finish

Target ≈ 600 words



Very long responses · median time to finish

Target ≈ 1,400 words



Benchmark specifications

Test set We tested 20 English-language writing prompts across four target lengths and five categories: creative, academic and technical, journalism, everyday and civic, and business. Each length had one prompt from each category. We ran each prompt once per model setting, for 120 completed runs. S marks prompts that ask for short responses (100 words), M for medium responses (250 words), L for long responses (600 words), and V for very long responses (1,400 words).

Settings We started fresh chats at Medium thinking effort on every model. We ran the models in their web or desktop app interfaces.

Scope The timings reflect the apps as tested, including differences in client, network, service load, and response length. The benchmark did not measure writing quality or speed against a pre-upgrade baseline. Headline token rates are app-reported medians rounded down to whole tokens per second.

Prompts used in benchmark

S1 · The pool caretaker

Creative

Work from the material below. Do not browse, call tools, or ask follow-up questions. Return only the requested writing.

Write an opening paragraph of about 100 words for a story narrated by the night caretaker of a municipal swimming pool. At closing time, they find a dry towel folded on the bottom of the deep end. Give the narrator a recognisable voice and a reason to hesitate before telling anyone. Stay with what happens in that moment and leave the towel unexplained.

S2 · Experimental controls

Academic and technical

Work from the material below. Do not browse, call tools, or ask follow-up questions. Return only the requested writing.

In about 100 words, explain the purpose of an experimental control to a first-year student. Use this example. A class tests whether a fertiliser changes bean seedling height after 14 days. Both groups get the same seed variety, soil, light, and amount of water. One group gets fertiliser in its water and the other gets plain water. Explain what the plain-water group lets the class compare and what the comparison would still leave uncertain. No results have been collected.

S3 · A local closure

Journalism

Work from the material below. Do not browse, call tools, or ask follow-up questions. Return only the requested writing.

Write the opening of a local news report in about 100 words. Use the facts below. Give readers the date and practical impact early, and attribute the reason for the closure.

The fictional town of Halden Bay will close the north footbridge from 3 to 14 May for deck repairs. The municipal notice says inspectors found deteriorating boards during a routine inspection. The south footbridge stays open. The posted detour adds about 850 metres to the walk between the station and the market. A temporary bus stop on Market Street will serve the west bank during the closure. The municipality says rain could delay reopening.

S4 · The shared path

Everyday and civic

Work from the material below. Do not browse, call tools, or ask follow-up questions. Return only the requested writing.

Write a note of about 100 words to the neighbour who shares your front path. Their bicycle has been leaning across the entrance several evenings this week. You sometimes come home carrying shopping and have trouble getting past. There is space against the wall next to the letterboxes, and you are happy to help move the bicycle the first time. Assume they may never have noticed the obstruction. Ask for the change in a friendly, direct way.

S5 · Delayed birthday mugs

Business

Work from the material below. Do not browse, call tools, or ask follow-up questions. Return only the requested writing.

Write a customer-support email of about 100 words. The customer ordered two ceramic mugs for a birthday on Friday. The parcel is now expected on Monday. The customer can keep the order and receive a shipping refund, or cancel before 2 p.m. Thursday for a full refund. The agent can process either option once the customer replies. Acknowledge the birthday and make the choice easy to act on. The shop has no confirmed faster delivery option.

M1 · The laundromat

Creative

Work from the material below. Do not browse, call tools, or ask follow-up questions. Return only the requested writing.

Write a scene of about 250 words set at a laundromat just before closing. Two siblings have come to wash the curtains from their childhood home before handing over the keys. One wants to ask the other to stay in town for another week, but neither can bring themselves to mention the sale directly. Tell the scene through dialogue and physical action, with no access to either sibling's thoughts. End on a small decision one of them makes.

M2 · Vocabulary study results

Academic and technical

Work from the material below. Do not browse, call tools, or ask follow-up questions. Return only the requested writing.

Write a results-and-interpretation note of about 250 words for an undergraduate methods class. The figures are fictional. Eighty-four students volunteered for a campus study of two ways of practising vocabulary. Students chose their own group. The outcome was a score out of 20 on the same delayed test.

Practice method	Students	Mean baseline score	Mean delayed score
Spaced practice	42	11.2	15.1
One-session practice	42	10.8	13.5

Describe what changed in each group, how the changes compare, and what the design allows you to conclude. No individual scores or measures of spread are available. End by proposing one practical change to the study design.

M3 · A council pilot

Journalism

Work from the material below. Do not browse, call tools, or ask follow-up questions. Return only the requested writing.

Write a local news brief of about 250 words based on this fictional council meeting. State what the council decided, how the pilot will work, and what remains open. Use direct quotes only when they appear in the transcript.

Clerk's record
On 8 April, Cedar Junction council voted 5–2 for a six-month weekend bus pilot beginning 7 June. Route 4 will run hourly between the station and the public market between 8 a.m. and 6 p.m. on Saturdays and Sundays. The approved budget is £84,000. Staff will report on ridership, cost, and missed trips after three months. No permanent service was approved.

Meeting excerpt
Transit manager Imani Reed said, "We'll count boardings and missed trips every weekend." Market trader Ben Sato said, "The first hour matters to us. That's when traders arrive." Councillor Ruth Vale said she voted against the motion because the route misses the east housing estate. The chair said any extension would require a new vote.

M4 · The family reunion

Everyday and civic

Work from the material below. Do not browse, call tools, or ask follow-up questions. Return only the requested writing.

Write a personal letter of about 250 words from an adult to an aunt who has invited them to a family reunion abroad. The writer can't afford the trip this year and doesn't want relatives to organise a collection. They genuinely want to stay close to the aunt. Give them something specific to remember together and one concrete way the writer could take part from home. Keep the explanation plain and warm, with room for the aunt to be disappointed.

M5 · Travel expenses

Business

Work from the material below. Do not browse, call tools, or ask follow-up questions. Return only the requested writing.

Write an internal memo of about 250 words for a fictional organisation with 40 staff. From 1 July, overnight work travel expenses must be submitted within 30 calendar days of returning. The deadline currently varies by team. A receipt and trip purpose are required. For trips booked before 1 July, the previous team deadline applies. Finance can approve exceptions for missing supplier receipts or extended leave. Line managers can confirm trip purpose, and Finance handles reimbursement. Explain what staff should do, why the change is being made, and whom to contact if they are unsure. The stated reason is to reduce time spent following up on incomplete claims.

L1 · Mara’s point of view

Creative

Work from the material below. Do not browse, call tools, or ask follow-up questions. Return only the requested writing.

Rewrite and expand the scene below into about 600 words from Mara’s close third-person point of view. Preserve the events, their order, and the spoken line. Give Mara a plausible reason for how she behaves, without explaining what happened to the person she has been waiting for. End when the bus leaves.

Original scene

From the driver’s seat, Ed watched the woman arrive a quarter of an hour before the last bus. She set two paper cups on the bench and turned both handles toward the road. When the rain started, she moved one cup under the shelter and left the other where it was. Ed pulled up beside the kerb. The woman stepped forward, looked past him into the empty aisle, and put a small brass key on the dashboard.

“If he asks, I waited.”

She went back to the bench. Ed closed the doors. In the mirror he saw her lift the cup that had been left in the rain. She held it with both hands as the bus turned the corner.

L2 · Tool library troubleshooting

Academic and technical

Work from the material below. Do not browse, call tools, or ask follow-up questions. Return only the requested writing.

Write a troubleshooting guide of about 600 words for volunteers who operate a fictional tool library’s Windows check-in computer. Readers can restart the inventory app and reconnect accessible USB cables. They cannot install drivers or change network settings. Use the known facts below. Distinguish what to try, what each result tells the volunteer, and when to escalate. Include a short example handover note.

Normal operation
The USB barcode scanner behaves like a keyboard. A successful scan places a six-digit item ID in the active field. The inventory app looks up that ID over the library network.

Observed faults
A scanner that beeps and types six digits into Notepad is sending input. If the app shows those digits but says “Item not found,” the catalogue record may be missing or inactive. If digits type normally and the app says “Connection unavailable,” start the paper check-in log. If the scanner neither lights up nor types into Notepad, reseat its USB cable once and retry. If it still fails, use manual entry where the item ID is legible. If the app pauses after a submission, don’t submit the same return again until you have checked the recent activity list.

Escalation
Send the shift lead the time, workstation name, item ID if available, exact error, and actions tried. Keep borrower names out of the handover note.

L3 · A harbour ferry proposal

Journalism

Work from the material below. Do not browse, call tools, or ask follow-up questions. Return only the requested writing.

Write a reported explainer of about 600 words for riders of a fictional harbour ferry. Explain how the proposal would affect different riders, what evidence the board has, and what it will decide next. Attribute the source of disputed claims. Give readers a concrete example using the fares below. Use only the material supplied.

Timeline
January, the harbour board published a proposal to raise the adult single fare from £3.20 to £3.80 and reduce the ten-trip pass from £29 to £28. Child fares would stay at £1.60. February, a public consultation received 412 submissions. March, staff recommended a three-month trial starting in June. The board votes on the trial on 12 April.

Records
The staff paper says fuel and berth costs increased and projects £76,000 in additional annual revenue if current travel patterns persist. It gives no estimate of how riders might change their travel. In the preceding quarter, 63% of adult journeys were paid for with a single ticket. The count is journeys, not individual riders. The ferry runs every 30 minutes on weekdays. No timetable changes are in the proposal.

Interviews
Ferry manager Anika Bell said, “We’re trying to keep a discount for regular riders without cutting sailings.” She said the trial would track boardings and pass sales. Evening-shift worker Tomas Rien said he takes the ferry about six times a month. He would struggle to tie up £28 at once and said some months he can get a lift instead.

L4 · The Hillview clinic stop

Everyday and civic

Work from the material below. Do not browse, call tools, or ask follow-up questions. Return only the requested writing.

Write a public comment of about 600 words to a fictional city's transport committee from a resident of Hillview. The resident accepts the goal of a more reliable service and wants the committee to retain a stop near the clinic or trial a workable equivalent. Make the request concrete. Use the supplied facts and distinguish the resident's experience from the council's estimates.

Proposal
Route 8 would use the bypass after 9 a.m., removing two daytime stops on Elm Street. Buses would run every 20 minutes instead of every 30 minutes. The council estimates a four-minute saving per trip. The closest replacement stop would be about 650 metres from the clinic by a route with one signalised crossing. The changes cover weekdays only. The council has proposed an eight-week trial and will accept public comments through 17 May.

Resident's experience
The resident accompanies their father to the clinic twice a month. He can walk the current 120 metres from the stop, with a rest on the bench outside the pharmacy. The resident also uses Route 8 for their own commute and has missed a connection several times this year. They can attend one daytime trial ride with a committee member. They would like the council to collect journey times and missed appointments if it can do so without identifying patients.

L5 · RelayDesk landing page

Business

Work from the material below. Do not browse, call tools, or ask follow-up questions. Return only the requested writing.

Write about 600 words of landing-page copy for RelayDesk, a fictional shared inbox for independent property managers. Write for a team of three to fifteen people coordinating repair requests across several buildings. Include a headline, an opening, three practical use cases, and a clear invitation to try it. Work from the facts below. Any illustrative scenario should read as an example, with no invented customer quote or results.

Product brief
Teams forward their existing maintenance address to RelayDesk. An incoming email becomes an item that can be assigned to a team member and given a due date. The activity log records who replied and when. Staff can add private notes and filter by building, assignee, or overdue status. Residents continue using email and do not need an account. Teams can export a CSV of their items. RelayDesk currently has no resident portal and does not dispatch contractors automatically.

Trial
The trial lasts 14 days and requires no payment card. Teams can begin with a single building. There is a sample inbox they can try before forwarding their own address.

V1 · The letter inside a wheel

Creative

Work from the material below. Do not browse, call tools, or ask follow-up questions. Return only the requested writing.

Write a standalone short story of about 1,400 words about a bicycle mechanic who finds an unsigned letter inside a wheel brought in for repair. Set the story in a fictional coastal town. Use three scenes and stay in the mechanic's point of view throughout. In the first scene, give the mechanic a plausible reason to look inside the wheel. In the second, let a reasonable assumption about the letter's author turn out to be wrong. In the third, establish who wrote the letter and give the mechanic a choice that matters to someone in town. Let the ending follow from that choice.

V2 · Literature synthesis

Academic and technical

Work from the material below. Do not browse, call tools, or ask follow-up questions. Return only the requested writing.

Write a literature synthesis of about 1,400 words for an undergraduate education methods seminar. Use only the three fictional studies below. Organise the synthesis around meaningful comparisons and disagreements. Explain what the studies jointly suggest about giving students feedback before revision, what their designs leave unresolved, and how you would design a follow-up study. Cite the studies by their supplied names. Treat the numeric outcomes in their original units.

Alder study
Two first-year history classes, 72 students in total, completed a source-analysis essay. One class received instructor comments on a draft five days before submission. The other received the same style of comments only after submission. The classes had different instructors. Final essays were scored by readers who did not know the class assignment. The draft-feedback class averaged 76 out of 100 and the other averaged 71. The report gives no prior essay scores. Students in the draft-feedback class could also attend optional office hours, and 14 did.

Bay study
In a first-year biology course, 116 students were randomly assigned either a short peer-review checklist or a worksheet asking them to reread their own lab-report draft. Both groups had 25 minutes in class and the same revision deadline. Reports were scored on a 12-point rubric. Mean scores were 8.3 and 8.1, respectively. Peer-review students corrected more unit-label errors. The groups were similar on the rubric's interpretation criterion. Twelve students missed the revision deadline, seven from peer review and five from the rereading group. The reported means use submitted revisions.

Cairn study
Forty adult learners in a voluntary evening writing course completed two personal essays. They received written feedback before revising the first essay and a ten-minute individual conference before revising the second. Twenty-six completed both essays and the exit questionnaire. Eighteen of those 26 said the conference made it easier to decide what to change. Two independent readers found more paragraph reordering in the second set of revisions. The assignments were on different topics, everyone encountered the methods in the same order, and the authors did not compare overall quality scores.

V3 · The repair café move

Journalism

Work from the material below. Do not browse, call tools, or ask follow-up questions. Return only the requested writing.

Write a reported feature of about 1,400 words on a fictional repair café facing a move. Give the story a human opening grounded in the supplied observation notes, then show what the move could mean for people who use or run the café. Represent the café, the council, and the site owner fairly. The lease decision remains open. Use direct quotes exactly as supplied. Keep the reporting within this packet.

Observation notes, 16 March
The café meets in the Old Laundry, a single-storey building in Marrow Wick. It opens from 10 a.m. to 1 p.m. on the first and third Saturdays. At 10:20, six people are waiting. Volunteer Leena Price uses a pencil to hold a portable radio's battery spring in position while owner Celia Dorn turns its dial. It makes a brief burst of static. At 11:10, a volunteer puts a handwritten "parts needed" note on the radio. Celia leaves carrying it in a bag. On the wall is a map showing three bus routes that stop within two blocks.

Café records
Its sign-in sheets show 218 item visits over 22 sessions in the previous year. There were 129 entries marked fixed, 51 marked parts needed, and 38 marked unresolved. Some visits were repeat visits. The records contain no estimate of money saved or waste diverted. The group charges no fee and accepts donations for shared supplies. It has 14 registered volunteers, of whom six came to at least half of last year's sessions.

Premises and alternatives
The owner has offered to extend the current lease through September at the current rent of £180 a month. After September the owner wants to convert the building into two workshops and has yet to receive planning approval. The council has offered two Saturdays a month at the civic hall at £95 per session. The hall is 2.4 kilometres farther from the market square than the Old Laundry and has step-free entry. An hourly bus stops outside. The café would have to take its tools home between sessions. The library has offered free storage for two locked cabinets and has no room to host sessions. The steering group will decide whether to request a two-session hall trial at its April meeting.

Interviews
Celia Dorn said, "I could buy a radio. I came because this one sat in my mum's kitchen for twenty years." She said she usually walks from her flat behind the market and is unsure whether she would use the hall.
Leena Price said, "The visitors see two hours at the table. They don't see the half-hour carrying things in and the half-hour carrying them out." She said the café has enough in its reserve to try the hall for two sessions.
Site owner Owen Tave said, "I've offered September because I know they need time." He said the building insurance had risen and that he expected workshop rents to cover costs.
Council facilities officer Jo Fen said, "We can offer a room and step-free access. We don't have a room that can be left set up between sessions." Fen said the café would pay the normal community rate, which includes cleaning.

V4 · Ash Quay residents' guide

Everyday and civic

Work from the material below. Do not browse, call tools, or ask follow-up questions. Return only the requested writing.

Write a plain-language guide of about 1,400 words for residents of the fictional Ash Quay housing estate. Explain the disruption, how residents can prepare using the services described, what to do while supply is interrupted, and where to check for updates. Account for residents who work nights, families with young children, residents unable to carry water, and people who are away when the work begins. Use realistic example schedules and a short FAQ to make the guide usable. Treat the fictional utility notice as the source for operational and safety instructions. Where the notice is silent, direct residents to the utility.

Utility notice
The utility plans to replace a valve from 9 a.m. to 7 p.m. Tuesday 14 July. Flats in Blocks A through D and the ground-floor shops on Wharf Lane are affected. Workers expect supply to return by 7 p.m. and will post revised estimates if work runs late. The utility's service desk is open from 7 a.m. to 9 p.m. on Tuesday. Residents can access it by calling extension 3300 from the estate intercom or visiting the office near the south gate. The estate noticeboard by the south gate shows the same updates. People who have signed up for estate text alerts will receive the revised estimate if there is one.

Support arrangements
A staffed drinking-water collection point will operate at the community room from 8 a.m. to 8 p.m. Bring a clean container. The utility will deliver water during two windows, 8–10 a.m. and 4–6 p.m., to people who cannot carry it home. Request a delivery at the estate office or by intercom by noon on Monday. If a need arises on Tuesday, call the service desk for help. The community room has step-free access from the south gate. Toilets at the library across Wharf Lane are expected to be open from 9 a.m. to 6 p.m. The library is on a different supply line.

When supply returns
The utility asks residents to follow the restart instructions posted with the restoration notice. If a flat still has no supply two hours after the utility confirms restoration, contact the service desk or leave a message on extension 3300. Building managers will notify residents separately if any building-specific instructions apply.

V5 · Bookshop returns pilot

Business

Work from the material below. Do not browse, call tools, or ask follow-up questions. Return only the requested writing.

Write an implementation proposal of about 1,400 words to the operations director of a fictional independent bookshop group. Propose a six-week pilot for a shared returns process across three shops and the central stockroom. Describe the workflow, who owns each step, rollout by week, staff training, handling of exceptions, success measures, and the decision at the end. Use the constraints below. Identify assumptions and give the director a concrete decision to make.

Current process
Each shop logs returns in its own weekly spreadsheet. The stockroom receives returned books on Wednesdays and Fridays. Duplicate entries sometimes appear when a customer visits one shop and mails the parcel to another. Staff sometimes wait until the next stockroom delivery before sending a refund request to Finance. The existing point-of-sale system can export a CSV with transaction ID, date, shop, and amount. The group will keep the existing point-of-sale system during the pilot.

People and capacity
Each shop can nominate one lead and reserve one hour for training. The stockroom coordinator can spend up to 90 minutes a week on the pilot. Finance runs refunds on Monday and Thursday. The operations director can authorise up to £2,500 in one-time spending. Existing shared-drive and spreadsheet tools are already paid for. The group has no reliable historical return-time baseline, but it can review two weeks of existing logs before the pilot begins.

Requirements
Staff must be able to identify duplicate cases using transaction ID and item details. The pilot log can keep transaction ID, item code, date received, current owner, status, and exception reason. Store no customer payment details in the log. Damaged copies and returns without a receipt go to the shop lead for review. The director wants a weekly view of outstanding cases, time from receipt to refund handoff, duplicates, and workload. At the end of week six, the group will decide whether to adopt, extend, or end the pilot.