Agreement Is Not Independent Evidence

Auditable Multi-Model Synthesis Without an API: The Anchor-and-Enrich Protocol and a Correlated-Witness Model of Machine Consensus

Vadym Chernets, PhD, AI systems architect

September 2026

Keywords: multi-model orchestration; large language models; ensembling; mixture-of-agents; correlated errors; algorithmic monoculture; effective sample size; design effect; provenance; auditability; evidence aggregation; indirect prompt injection; accessibility

JEL Classification: C11; C18; D71; D83; L86; O33

Abstract

Someone who asks one language model a question cannot tell a confidently wrong answer from a confidently right one. The two are identical on every surface cue available to a non-expert. Asking several models seems to solve this, but it mostly relocates the problem: reconciling several answers produces one confident text again, and that text now carries the borrowed authority of “ten of them agreed.”

The protocol formalized here collects answers from several consumer chat models through the user’s own browser or phone, with no API keys, no installation and no intermediary server. It reconciles them by anchor-and-enrich: the strongest available answer becomes the base, and each of the others contributes only what the base lacks. Every contribution is signed with the model that supplied it. Disagreements are kept in a section of their own rather than averaged away.

The paper makes four contributions. The first is the protocol itself, stated as an algorithm that preserves provenance and forbids silent deletion, together with a taxonomy of five ways a merge can fail. Four of them are defects of naive pooling; the fifth is a failure of the operator rather than of the method: a synthesis that presents a single model’s output as a collegial result. The second is the application of the correlated-witness arithmetic to protocol design. That arithmetic is not claimed here. Kohli (2026) applied the same Kish correction to a panel of nine frontier judges and measured about two effective votes, and Kim et al. (2025) established the underlying error correlation across more than 350 models. A council of N models whose errors carry intraclass correlation ρ is worth N_{eff} = N / (1 + (N−1)ρ) independent witnesses, and that quantity saturates at 1/ρ however many models are added. Once a council is larger than 1/ρ, halving ρ beats doubling N, and a pair drawn from two laboratories can outweigh five drawn from one. Kim et al. (2025) complicate that rule: the larger and more accurate models carry the most correlated errors even when architecture and provider differ, so two laboratories are a prior guess about ρ rather than a measurement of it, which is what the estimation procedure of Section 5.5 is for. The third is an account of three deployment tiers, manual, one-tap mobile and browser-agent, and of the axis that separates them epistemically: whether the second model sees the first model’s answer. Conditioning the second model raises ρ in exchange for error correction. The fourth is a small-sample observation of the counterfeit-council failure across five models, the one-sentence

intervention that removed it in a three-model re-run, and a blinded three-arm evaluation design with a stated falsification condition.

Correlated agreement is treated here as established, not as a finding. That a panel of models is worth far fewer independent witnesses than it has members has been measured directly: nine frontier judges drawn from seven families supply about two effective votes (Kohli, 2026), and error correlation rises with model capability even across providers and architectures (Kim et al., 2025). Section 5 restates the design-effect arithmetic behind those measurements, because three of the protocol's rules are derived from it and read as fastidiousness without it. The protocol then takes the other side of the same problem. That literature measures the correlation; this protocol is built to lower it, by composing across lineages, keeping disagreement instead of averaging it away, signing every contribution with the model that supplied it, and counting a repeated source once.

The claim that a merged answer beats the best single answer is argued from the design and from adjacent literature; Section 9 specifies the evaluation, which has not been run.

Contents

List of figures and tables

1. Introduction

1.1 The problem

Take a question where being wrong is expensive: choosing a school, a treatment, a supplier, a clause in a contract. The user asks a language model and gets back something coherent, confident and well formatted, and the difficulty starts there. A confidently wrong answer and a confidently right one look the same. Tone does not separate them, nor does structure, length, or the presence of citations. Verbalized model confidence is known to be overstated and poorly calibrated (Kadavath et al., 2022; Xiong et al., 2024). A model's tendency to adopt the questioner's framing adds a second layer of false support (Sharma et al., 2023).

The obvious remedy is a second opinion. One model's error looks like confidence. The same error placed next to a second model's answer looks like an argument. An argument is visible to a reader with no expertise in the subject; confidence is not. Everything below rests on that asymmetry.

The remedy has a price and a trap. The price is reading: ten answers to one question are ten long texts to hold in mind at once while remembering who said what. Almost nobody does that. In practice a reader gets through the first answer, skims the second and closes the rest. Gathering several answers has been a solved problem for years, through aggregator subscriptions, side-by-side panels and API gateways. Finishing with them has not been solved.

The trap is the merge. Any way of reducing ten answers to one produces a single confident text again, and if the reduction is opaque the reader is back where they started, except that the text now carries an authority they cannot check. Naive pooling also erases correct minority points, drifts toward whichever phrasing was most emphatic, and produces the look of consensus where there is none.

A third problem is what makes the whole exercise non-trivial. Agreement among models is not independent confirmation, because models trained on overlapping corpora share framings, cite the same articles and inherit the same blind spots. "Nine of ten agreed" and "nine independent experts agreed" are different statements, and the gap between them is quantitative and large. Section 5 shows that it puts a ceiling on the entire construction.

1.2 The system

The protocol is written out in plain text, with three ways of executing it. All three produce the same document, and they differ in who does the clicking and in how many models get asked.

On the free path the user opens the sites, copies the answers, and pastes them under one paragraph of instructions into any chat model. Nothing to install, nothing to pay, and it works on a phone.

On the mobile tier there is one icon on the home screen. Tap it, type the question, pick a mode, and about a minute later the finished answer is on screen and in the clipboard. It reaches two models through operating-system app intents against subscriptions already installed on the device, with no API key.

On the agent path a model that can drive a browser opens the named sites in the user's existing logged-in sessions, submits the question, waits, saves each answer to a file and builds the document.

The architecturally decisive fact is that there is no server. Not a secured server, and not one with a no-retention policy: no such component exists, so the question cannot reach me. It does reach the model services the user named, which is the point of the exercise, and it goes nowhere else. A second consequence follows from the same absence: the protocol cannot be withdrawn, since it is a paragraph of text under a permissive license, and a copy in a user's notes goes on working with whatever models exist later.

1.3 Contributions

Section 3 sets out five ways a merge fails and Section 4 formalizes anchor-and-enrich against them. Section 5 develops the correlated-witness model, N_{eff} = N/(1+(N−1)ρ) with its ceiling at 1/ρ, and draws out what it implies for how a council should be composed. Section 6 describes the three tiers, the latency–breadth frontier they trace, and the blind-parallel against informed-sequential distinction that separates them epistemically. Section 7 reports the counterfeit-council observation and field findings from the mobile build; Section 8 describes an interface pattern for cold-start guidance; Section 9 specifies a blinded evaluation with a falsification condition fixed in advance. The claim that a merged answer beats the best single answer is not settled here, and that evaluation is the comparison that would settle it, including the result that would oblige me to withdraw it.

The orientation is worth stating plainly, because it is what separates this paper from the measurement literature it depends on. Kohli (2026), Kim et al. (2025) and Ding (2026) measure how correlated a council's errors are. This paper takes that correlation as the design target and asks how a single person, working through ordinary consumer chat interfaces with no API key and no server, can lower it and leave a record that shows they did. One gap in that literature is worth naming because it is within reach: the measurements were made over APIs, on benchmarks, with nine models or with three hundred. Nobody has estimated the correlation in the consumer setting this protocol inhabits, where a person queries four chat products they already pay nothing for. Section 5.5 gives the procedure and Section 9.6 the design.

---

2. Related work

Ensembles and diversity. Classical ensemble theory already says what is needed here: what an ensemble gains depends on how differently its members err, not on how good they are on

average. The ambiguity decomposition, in which ensemble error equals average member error minus average ambiguity (Krogh & Vedelsby, 1995), and the bias–variance–covariance decomposition (Ueda & Nakano, 1996) both show an ensemble collapsing toward a single member as error covariance rises. Dietterich (2000) states the condition as a necessary one. Section 5 carries it over to councils of language models and puts a number on what violating it costs.

Multi-model schemes for language models. Self-consistency (Wang, X., et al., 2023) votes across several samples from one model, which reduces variance without producing independence. Multi-agent debate (Du et al., 2023) reports factuality gains when several instances exchange arguments. Mixture-of-agents (Wang, J., et al., 2024) feeds one layer's outputs to an aggregator and reports improvement over the strongest single model on several benchmarks. That literature points the same way this protocol does. Its results were obtained over APIs and on benchmarks, though, so I do not transfer them to a browser-mediated council of consumer interfaces without measurement of my own, which is what Section 9 specifies.

Judge models. Evaluation by a model judge is well characterized together with its defects: position bias, verbosity bias, and a pull toward a familiar house style (Zheng et al., 2023). That is a design constraint here, not an aside, since unblinded scoring would partly be measuring those biases. Section 9.4 handles that.

Monoculture. Kleinberg and Raghavan (2021) show that when many agents rely on one algorithm, collective outcomes can worsen even as individual quality improves. Bommasani et al. (2022) describe the same effect as outcome homogenization. For a council of language models this is the mechanism that generates $\rho > 0$ in the first place. Work on degradation under training on generated data (Shumailov et al., 2024) suggests ρ drifts upward over time.

Crowds. The classical wisdom-of-crowds argument (Surowiecki, 2004) depends on judgments being independent. Lorenz et al. (2011) showed experimentally that even mild social influence, simply letting people see others' estimates, narrows the spread without moving it toward truth: the crowd gets more confident without getting more accurate. Section 6.5 identifies that as exactly the trade accepted by informed-sequential elicitation.

Condorcet and correlated votes. The Condorcet jury theorem (Condorcet, 1785) holds that with individual accuracy above one half, the probability of a correct majority tends to one as N grows. Ladha (1992) showed that correlated votes destroy the guarantee. Section 5 gives a quantitative version of that warning through the design effect (Kish, 1965).

Correlated errors, measured. The arithmetic Section 5 restates has since been measured on model panels directly, and the measurements are sharper than the analysis. Kohli (2026) applies the Kish effective sample size to a panel of nine frontier judges drawn from seven model families and finds it worth about two independent votes: roughly three-quarters of the panel's nominal independence is lost to shared mistakes, the panel's accuracy falls 8 to 22 percentage points short of what independent voting would deliver, and the best single judge matches or beats the whole panel. Neither more judges nor better aggregation repairs it, which locates the bottleneck in the judges rather than in the algorithm. Kim et al. (2025), across more than 350 models, find substantial error correlation and, crucially for how a council should be composed, find that the larger and more accurate models have the most correlated errors even when architecture and provider differ. Ding (2026) audits agreement as a confidence signal over 265,000 samples and

finds it a positive but weak predictor, with correlation between 0.20 and 0.59, and worst exactly where it is most confident: on the most consistent frontier model, agreement at or above 0.8 covered 77% of the harder benchmark's entries and 48% of those were wrong.

Three things follow for this paper. First, the design-effect result in Section 5 is not a contribution of mine. Kohli's paper predates the first posting of this one and used the same correction; the section is kept for what the protocol derives from it, and it is labeled as such where it appears. Second, the composition advice that follows, spend on diversity rather than on count, is now empirically supported instead of merely argued. Third, Kim et al.'s finding complicates that advice in a way worth stating: if capability and error correlation rise together, a pair drawn from two laboratories is not automatically a diverse pair, and lineage has to be measured rather than read off the brand. That is an argument for Section 5.5's estimation procedure, not against the composition rule.

Adjacent systems. Several projects build multi-model councils. One well-known open implementation, llm-council (Karpathy, 2025), queries models in parallel, has them review each other anonymously, and lets a chairman model write the final answer. It needs an aggregator API key and per-token payment. Commercial council products offer a judge, a consensus view and a contradiction display. Open-source mobile applications add voting and multi-round debate. Mature multi-model front ends ship ensemble features, and free side-by-side panels use the official sites the user is already logged into. Aggregator subscriptions put many models behind one bill. Each of these properties exists elsewhere on its own. Their conjunction, ten of them at once, appears to be unavailable elsewhere: no API key, no installation, phone-compatible, running on accounts the user already holds including free tiers, no new paid intermediary, a merge method written in plain text and therefore editable, disagreement preserved, every contribution signed, a repeated source flagged as one source, and one method executable either by hand or under an agent.

There is a structural reason the position stays open. A laboratory has little incentive to route a user's question to its rival, since each participant's objective is that its own answer be the one the user keeps. Third parties do sell cross-model tools, but a method written in plain text leaves the editorial role with the user, which is a different arrangement from any of them.

Scope. The protocol covers one person producing one careful document without assembling a research team's infrastructure: no keys, no installation, no server. Retries, guaranteed execution, cost and error monitoring, audit logs, evaluation on test sets, access control, single sign-on, retention policy, prompt versioning, and service-level agreements are the province of enterprise orchestration frameworks.

---

3. Five ways a merge fails

Four of the five are properties of naive pooling, by which I mean any instruction of the form "combine these answers and optimize the result." The fifth belongs to the operator, and it is the one that matters most, because the reader cannot detect it.

F1, fabrication in synthesis. The output contains claims that were in none of the inputs. The synthesizer is itself a language model, and in smoothing the seams between sources it generates connective material that looks exactly like sourced material.

F2, minority erasure. The one answer that spotted a material objection dissolves because the majority said otherwise. This is the case the council was assembled for, and it is the first thing a naive merge loses.

F3, confidence drift. Where two phrasings conflict, the more categorical wins over the better supported. It is the same mechanism that produces position and style bias in judge models (Zheng et al., 2023).

F4, manufactured consensus. A real disagreement is averaged into a qualified yes. The reader gets a document in which the contested is indistinguishable from the settled and cannot even learn that a dispute existed.

F5, the counterfeit council. A model told to run a collegial pass has no access to other models. Instead of saying so, it performs ordinary web research and formats the result in collegial shape, with a receipt, agreement and disagreement sections, and donor attributions. The document looks better than a truthful one would, exactly because no real contradictions get in the way. The reader cannot catch this, having no access to the source answers.

F5 is the opening problem of the paper raised one level: a confidently wrong answer is indistinguishable from a right one, and a document shaped like a council but written by one model is indistinguishable from a council. Section 7 reports it occurring in two of five models in a blind run.

---

4. The anchor-and-enrich protocol

4.1 Principle

Instead of pooling and optimizing answers, the protocol grows one strong answer incrementally, while provenance and disagreement stay visible.

4.2 The rules

Figure 2 shows the protocol end to end.

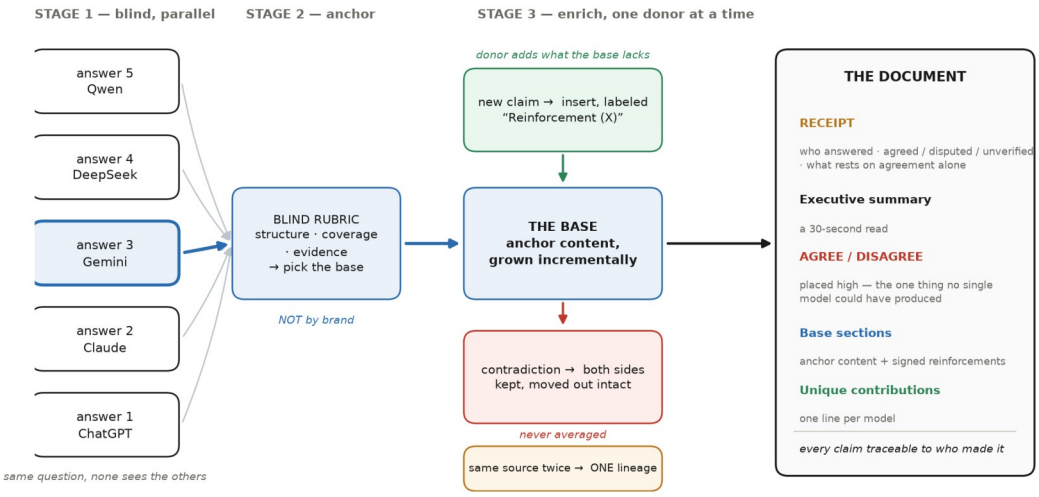

Figure 2 — The anchor-and-enrich protocol. *Answers are elicited blind and in parallel. A blind rubric picks the base on content rather than brand. Each remaining answer contributes only what the base lacks, labeled with its donor, while contradictions move intact into a disagreement section and repeated citations collapse into one evidentiary lineage. The output is ordered so that the receipt and the agreement analysis are read first.*

1. **Choose the anchor by a blind rubric, not by brand.** The strongest ready answer, meaning the most complete and best sourced, becomes the base. Judge on content: structure, coverage, evidence. Presetting a favored model is a systematic bias, and it is worst in the case where one model happened to receive more context than the others. Once the anchor is chosen, building starts immediately. Stragglers are not waited for.

2. **Enrich one donor at a time.** From each remaining answer take only what the base lacks or what strengthens something already there: a precedent, a figure, a comparable, an angle, a primary-source link. Insert it into the matching section with the label *Reinforcement (model name)*.

3. **Preserve provenance.** Every contribution carries the mark of its donor. Model provenance is not evidence provenance, and where the original source link is available it is carried too, not just the attribution.

4. **Correct openly.** Base content is retained by preference. Where a donor corrects an error in the anchor, make the correction and leave a one-line note of what changed and why. Cementing a known anchor error out of deference to a rule against touching the base is worse than touching it. The governing rule is that information is not lost silently, not that the base is immutable.

5. **Surface disagreement.** A dedicated section separates agreement from divergence. Conflicting claims are never averaged into an intermediate position.

6. **Flag correlated evidence.** Several models citing the same article are one evidentiary lineage, not several independent confirmations. The same applies at the level of

composition. If the agreeing models were built by companies of the same type or in the same jurisdiction, and the split falls along that line, say so.

7. **Credit unique contributions**, one line per model.
8. **Add the orchestrator's own view** as a final section where the user asked for it.
9. **Name the load-bearing facts.** Identify the two or three facts the recommendation actually rests on, usually numbers, and mark each as backed by a checkable source or as resting only on the models agreeing. Where they agree on a figure and none of them shows where it came from, that is a warning, not a confirmation.

4.3 Algorithm

```
INPUT   Q — the question
        A = {a_1 … a_N} — answers, arriving asynchronously
OUTPUT  D — a document with a receipt, provenance, and
        a disagreement section

R ← ∅            // receipt: who answered, who did not, why
D ← ⊥            // the base (anchor)

on arrival of a_i:
    R ← R ∪ {source(a_i), timestamp, mode}

    if D = ⊥ and enough_candidates():
        D ← argmax_{a ∈ A_ready} rubric(a)
        anchor ← source(D)         // recorded and disclosed
        // rubric = structure, coverage, sourcing

    else:
        for each claim c ∈ a_i:

            if c ∉ D:                                // new
                insert c into the matching section of D,
                    labeled "Reinforcement (source(a_i))"

            else if c contradicts c' ∈ D:
                if c corrects a factual error in c':
                    replace c' with c, plus a one-line
                    note of what changed
                else:
                    move (c, c') into DISAGREEMENTS,
                    both sides intact

            else:                                // agreement
                if evidence_source(c) = evidence_source(c'):
                    mark as ONE evidentiary lineage,
                    not as confirmation
```

```
after all:
    identify the load-bearing facts
    mark each [sourced] or [agreement only]
    assemble RECEIPT, SUMMARY, AGREE/DISAGREE (placed high),
             UNIQUE CONTRIBUTIONS
    emit in a readable and a double-clickable format
```

Collection and synthesis run together. The anchor becomes the base as soon as it arrives, and each later answer is folded in as it lands. Fast providers respond in one or two minutes while deep-research modes take five to twenty-five, so serializing the two phases would turn a run into a half-hour block for no benefit.

4.4 Output shape

Section order is part of the protocol, not presentation.

The receipt sits at the top: who answered, who did not and why, what is agreed and disputed and unverified, where the evidence is correlated, and roughly what the run cost in time and quota. Then a thirty-second executive summary, then a table of contents.

Agreement and disagreement come immediately after the summary: they are the one component no single model could have produced, and the reason to have run the method at all. Put at the end, underneath a comfortable consensus, a useful document starts to read as obvious, and then it stops being read.

After that come the base sections, each consisting of anchor content plus signed reinforcements; a list of what each model contributed uniquely; and, where the user asked for it, the orchestrator’s own view.

4.5 Rules and the failures they address

Table 1. Protocol rules and the failures they address

Rule	Failure addressed
Anchor by rubric, not brand	Systematic bias toward a favored model
Donors add only what is missing	F1: the base is not rewritten, so fewer seams are generated
Provenance on every contribution	F1, and auditability: the reader can check instead of trusting
Correction with an explanatory line	Cementing an anchor error
Disagreement in its own section, placed high	F2, F4
Shared evidentiary lineage flagged	False independence (Section 5)
Load-bearing facts marked sourced or agreement-only	Agreement treated as evidence
Non-runs never presented as runs	F5

This is a design argument, and Section 9 specifies what would test it.

4.6 A worked example

A protocol of this kind is easier to judge against its output than against its description. What follows is an excerpt from a real run of the free path on 31 August 2026: one question, asked in four separate sessions to four consumer chat models, then merged by the rules above. Donor names appear here, unlike in Section 7, because provenance is the property being demonstrated and an anonymized example would demonstrate nothing.

```
RECEIPT — 4 models asked · 4 answered
Agreed:    3-2-1 redundancy · JPEG is the safe everyday format ·
           migrate every ~5 years · never let a proprietary app
           be a photograph's only home
Disputed:  how many copies (3 vs 4) · what the archival master
           should be (TIFF/PNG vs original plus JPEG)
To check:  media lifespan figures · whether your export tool
           strips EXIF

— Section 1 of the merged document: how many copies, and where —

Base.  The 3-2-1 rule: three copies, on two kinds of storage,
with one off-site. The off-site copy is what survives a fire,
a flood or a theft.

⚠ Disagreement — is there a fourth copy?
ChatGPT and Gemini both extend this to 3-2-1-1, adding one copy
that is immutable or offline. Their reasoning differs: ChatGPT
frames it as geographic independence, Gemini as protection from
ransomware and from a mass deletion that syncs everywhere.
Claude and DeepSeek stayed at three and did not argue against a
fourth; they simply did not call for one.

(added by Claude)  Whatever you choose, treat cloud as one leg
of the tripod and never the whole thing: it is tied to an
account that can be locked, closed, or repriced.

— Worth checking yourself —

The lifespan figures come from Gemini alone. No other model
gave numbers, so treat them as one source rather than as a
consensus.
```

Four things in that excerpt are the protocol working rather than the protocol being described. The disagreement over a fourth copy is carried with both sides intact and with the two models' differing reasons, instead of being resolved into a serviceable "three or four, depending on your risk appetite" that would have concealed which consideration a reader might care about. The reinforcement about cloud storage carries the name of the model that supplied it, so a reader who distrusts that model can discount that line and no other. The closing note is rule 6 firing in a live run: one model supplied the only numbers in the document, and the merge marked them as a single source rather than letting their presence in a four-model document read as agreement.

The fourth is what the excerpt does not carry: it predates rule 9, so its load-bearing facts are not separated into sourced and agreement-only, and it is reproduced as it came out, because an example edited to satisfy the rules it illustrates would no longer be evidence that they operate.

---

5. Correlated agreement

> *Priority note. The arithmetic in this section is not new and is not claimed as a contribution. Kohli (2026) applied the same Kish correction to a panel of nine frontier judges and measured about two effective votes, three months before this paper was first posted, and Kim et al. (2025) established the error correlation that drives it across more than 350 models. The section is kept for three reasons: three of the protocol's rules are derived from it, a reader who has not seen the arithmetic will read those rules as fastidiousness rather than as necessity, and the estimation procedure in Section 5.5 is what the protocol needs in a setting nobody has measured. Where that literature measures the correlation, this protocol tries to lower it.*

If ten models agree, the claim is probably true after all. That is the strongest intuitive objection to Section 4 and it is weaker than it feels; working out how much weaker also turns rules 6 and 9 from good practice into arithmetic.

5.1 Setup

Let N models answer one binary claim, and let e_i ∈ {0,1} indicate an error by model i. Under independence, with individual accuracy p > 1/2, agreement among N models is evidence that grows exponentially in N. That is the content of the Condorcet theorem (Condorcet, 1785).

Models trained on overlapping corpora inherit shared framings, cite the same secondary sources and share the blind spots of their period, so their errors are correlated. A court is in the same position with ten witnesses who all watched the same security recording: ten statements, one observation, and the number of witnesses says nothing about how much was independently seen. That gap can be given a number, and the rest of this section does it. Define the intraclass correlation of errors as

$$\rho = corr\left(e_i, e_j\right), i \neq j, 0 \leq \rho \leq 1$$

5.2 Effective number of witnesses

A council of correlated models is structurally a cluster sample. N observations inside one cluster with intraclass correlation ρ do not carry N units of information, and the standard correction is the design effect (Kish, 1965):

$$DEFF = 1 + \left(N - 1\right)\rho$$

which gives an effective sample size of

$$N_{eff} = \frac{N}{1 + \left(N - 1\right)\rho}$$

So a council of N models whose errors correlate at ρ is worth as much as N_{eff} independent witnesses. Table 2 gives selected values, and Table 8 in Appendix B gives the full range.

Table 2. Effective number of independent witnesses, N_{eff}, for selected N and ρ

ρ (down) / N (across)	2	3	5	10	20	limit N → ∞
0 (independent)	2.00	3.00	5.00	10.00	20.00	∞
0.1	1.82	2.50	3.57	5.26	6.90	**10.0**
0.2	1.67	2.14	2.78	3.57	4.17	**5.0**
0.3	1.54	1.88	2.27	2.70	2.99	**3.33**
0.5	1.33	1.50	1.67	1.82	1.90	**2.0**
0.7	1.18	1.25	1.32	1.37	1.40	**1.43**

5.3 The ceiling

The rightmost column is the main result of this section. As N → ∞,

$$\lim_{N \to \infty} N_{eff} = \frac{1}{\rho}$$

A council has a ceiling, and it does not depend on how many models were asked. Panel (a) of Figure 1 shows the saturation directly. At ρ = 0.5 a council of a thousand models is worth two independent witnesses, and at ρ = 0.7 it is worth fewer than one and a half. Adding an eleventh model to ten at ρ = 0.3 buys about 0.05 of a witness. Moving from ρ = 0.3 to ρ = 0.15 with the same five models buys about 0.85.

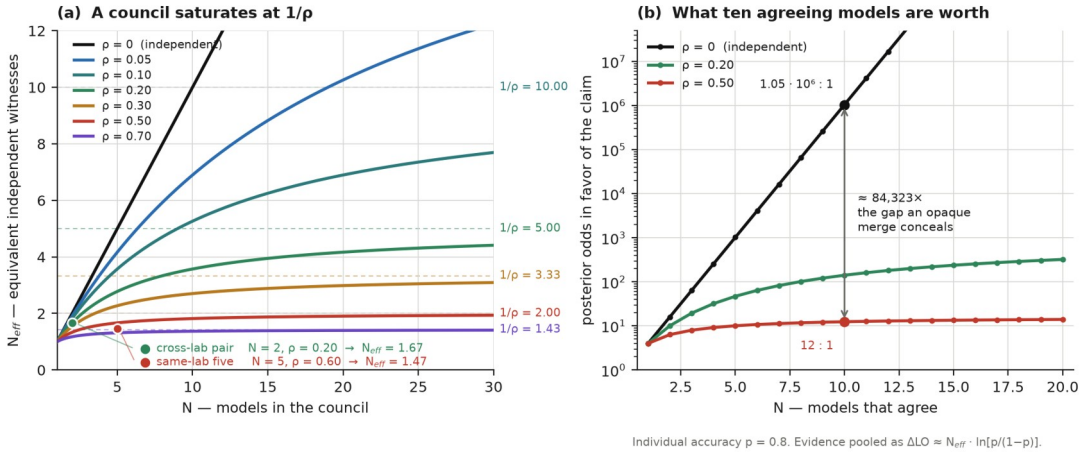

Figure 1 — A council saturates at 1/ρ. *(a) N_{eff} against council size N for several values of the intraclass error correlation ρ. Each curve saturates at its ceiling 1/ρ, drawn as a dashed asymptote; a cross-lab pair at ρ = 0.20 outweighs a same-lab quintet at ρ = 0.60. (b) The same*

result as posterior odds (an illustrative proxy under the approximation stated in Section 5.4), with individual accuracy p = 0.8 and evidence pooled as ΔLO ≈ N_{eff} · ln[p/(1−p)]. Ten independent agreeing models give odds near a million to one; ten models correlated at ρ = 0.5 give twelve to one.

The practical consequence runs against the intuition that one should simply ask more models. Once a council is larger than 1/ρ, halving ρ beats doubling N. That consequence is Kohli's (2026) as much as anyone's, and his panel is this arithmetic observed in the field: nine judges worth about two effective votes corresponds to ρ ≈ 0.44. Halving ρ beats doubling N, and the two comparisons below bear on Section 6.

Two models from different laboratories at ρ ≈ 0.2 give N_{eff} ≈ 1.67, while five from one laboratory or lineage at ρ ≈ 0.6 give 5/(1 + 4·0.6) ≈ 1.47, so the cross-laboratory pair carries more weight than the five.

Breadth also suffers sharp diminishing returns where diversity does not: movement along a row of Table 2 saturates quickly, while movement up a column has no limit.

That is the analytical basis for composing a council evenly across distinct lineages instead of maximizing the count. Five models that agree because they read the same corpus are weaker evidence than three that genuinely diverge.

5.4 In log-odds

If an independent witness with accuracy p contributes log-odds

$$L = \ln\left(\frac{p}{1-p}\right)$$

then agreement among N correlated models contributes approximately

$$\Delta LO \approx N_{eff} \cdot L$$

instead of N · L. This is the ordinary treatment of correlated observations in meta-analysis, used here as an approximation: it treats the design-effect-adjusted count as if it were a number of independent witnesses, which the design effect itself does not establish. Panel (b) of Figure 1 plots it. At p = 0.8, so that L ≈ 1.386 nats:

- ten independent witnesses give ΔLO ≈ 13.86 nats and posterior odds of about $1.05 \cdot 10^6$: 1;
- ten models at ρ = 0.5, so N_{eff} ≈ 1.82, give ΔLO ≈ 2.52 nats and posterior odds of about 12 : 1.

The ratio is roughly $8.4 \cdot 10^4$. That factor is what an opaque merge hands the reader as unearned confidence, and it is why the receipt exists and why load-bearing facts are marked. The document has to say whether the agreement in front of the reader is worth 10^6 : 1 or 12 : 1, because on the page the two are identical.

5.5 Assumptions and estimation

The model is simple and its assumptions are correspondingly strong. It rests on the following.

ρ is treated as uniform across pairs. In practice it is block-structured, since models from one laboratory correlate more strongly with each other than with models from elsewhere. A pairwise average would be better and a block structure better still. Under a block structure the uniform formula gives an optimistic N_{eff} and the true ceiling sits lower than the tables show.

ρ is also treated as constant across questions, when it is plainly domain-dependent. On a niche topic with one available source it approaches one. On a question with a rich multilingual literature it is much lower.

Claims are treated as binary, whereas real answers are prose, and reducing prose to comparable binary claims is a noisy operation in its own right. Accuracy p is treated as uniform across models, which it is not.

ρ is nonetheless measurable. Given Q labeled binary claims with known ground truth and the resulting error matrix e_iq for i = 1…N and q = 1…Q, the intraclass correlation follows by the usual analysis-of-variance route, as the ratio of mean pairwise error covariance to mean error variance:

$$\hat{\rho} = \frac{mean_{i \neq j} cov(e_{i\cdot}, e_{j\cdot})}{mean_i var(e_{i\cdot})}$$

Estimating ρ separately within and across laboratories gives a number for the cost of monoculture (Kleinberg & Raghavan, 2021; Bommasani et al., 2022). Watching it drift over successive model generations does the same for the cost of training on generated data (Shumailov et al., 2024).

5.6 What this changes in the protocol

Once the formula is granted, rules 5 and 6 in Section 4, and the composition reporting that rule 6 implies, stop being good practice and become necessary.

Flagging a shared source, rule 6, is a local measurement of ρ at the level of a single claim. Five citations of one article mean ρ ≈ 1 for that claim, so $N_{eff} \approx 1$.

Reporting kinship among the agreeing models is the same measurement at the level of composition.

Preserving disagreement, rule 5, is the only evidence a reader ever gets that ρ is below one. A divergence demonstrates that the correlation is not total, and it is informative whichever side turns out to be right. Disagreement is a measuring instrument and not an obstacle to be cleared away. A council that always agrees tells the reader nothing about the world, only that ρ ≈ 1 and that the council was superfluous.

---

6. Three tiers, one protocol

The protocol runs at three tiers, which are three points on an access curve rather than three versions or a funnel. One user may well use all three in a week for different questions.

6.1 Tier 0, the free path

The user opens the sites, asks the question, copies each answer, and pastes them under the merge prompt into any chat model. No cost, no installation, any device with a browser including a phone, and a council as large as the set of accounts the user is signed in to. Latency is ten to fifteen minutes, most of it waiting, and more with deep-research modes.

This tier is the reference implementation in a strict sense. It is the only one where the user holds every raw answer, which makes it the only one where the fidelity of the merge can be judged directly.

6.2 Tier 1, the mobile tier

One icon on the home screen and one entry in the shortcut library. Tap, type the question, pick a mode from a menu, and roughly a minute later the finished answer is on screen and in the clipboard. Voice launch works.

The mechanism is neither a browser nor an API. It chains the vendors' own operating-system app intents against applications and subscriptions already installed on the device. Cost is denominated in subscription messages, not tokens, which makes mode design a resource-allocation decision rather than a matter of taste. The alternative mobile platform has an equivalent through intent-based automation.

Table 3 sets out the three modes. They are equal parallel functions, not stages of a pipeline: one mode runs per invocation.

Table 3. The three modes of the mobile tier

Mode	Chain	Messages	Use
Poll	Both models answer independently; answers shown side by side	1 + 1	Seeing the raw divergence directly
Critique	Model A answers; model B verifies as a second expert and issues an improved final answer	1 + 1	Everyday default: a checked answer at two-message cost
Synthesis	Both answer blind; a further call arbitrates and merges, marking divergences	1 + 2	Questions that matter

The critique mode was briefly called *chairman*. The three modes are equal parallel functions and critique is peer review: one expert answers, a second verifies. The chairman role already exists inside Synthesis, as the arbiter, and calling peer review by the name of aggregation would have misdescribed what the mode does.

A phone is not where breadth is cheap, which is what fixes N = 2 here. Each call spends a message from the same quota as ordinary use, the applications come to the foreground in sequence, and the user is standing there waiting. Section 5.3 shows what the pair is worth: the two models come from different laboratories, the ρ-minimizing choice available at N = 2, and at those illustrative correlations a cross-lab pair at ρ ≈ 0.2 outscores a same-lineage five at ρ ≈ 0.6, so most of what a council is for survives at one tap and one minute.

6.3 Tier 2, the agent path

A model that can drive a browser opens the named sites in a browser session the user has signed in to, submits the question, waits, saves each answer to a file, and builds the document. The council extends to ten or more, deep-research modes are available, and a dated folder with every original survives the run. Cost is a small monthly subscription for the agent plus whatever paid tiers the user already holds. Latency is five to thirty minutes.

This tier gives the user something that outlasts the use case, namely a model that acts on their computer instead of describing actions to them. The difference between a model that answers and a model that acts is larger than the difference between one model and ten.

6.4 The latency–breadth frontier

Table 4. The latency–breadth frontier across the three tiers. Effective witness counts assume ρ = 0.2

	Tier 0, free	Tier 1, mobile	Tier 2, agent
Who clicks	the user	the operating system	a browser agent
Council size N	unbounded	2	up to ~10
N_{eff} at ρ = 0.2	up to ~4–5	1.67	~3.6
Time to answer	10–15 min	~1 min	5–30 min
Marginal cost	zero	subscription messages	agent subscription
Installation	none	one signed shortcut	agent application
Deep-research modes	yes, manually	no	yes
Raw answers visible	yes	Poll mode only	yes, saved to files
Device	any	phone	computer

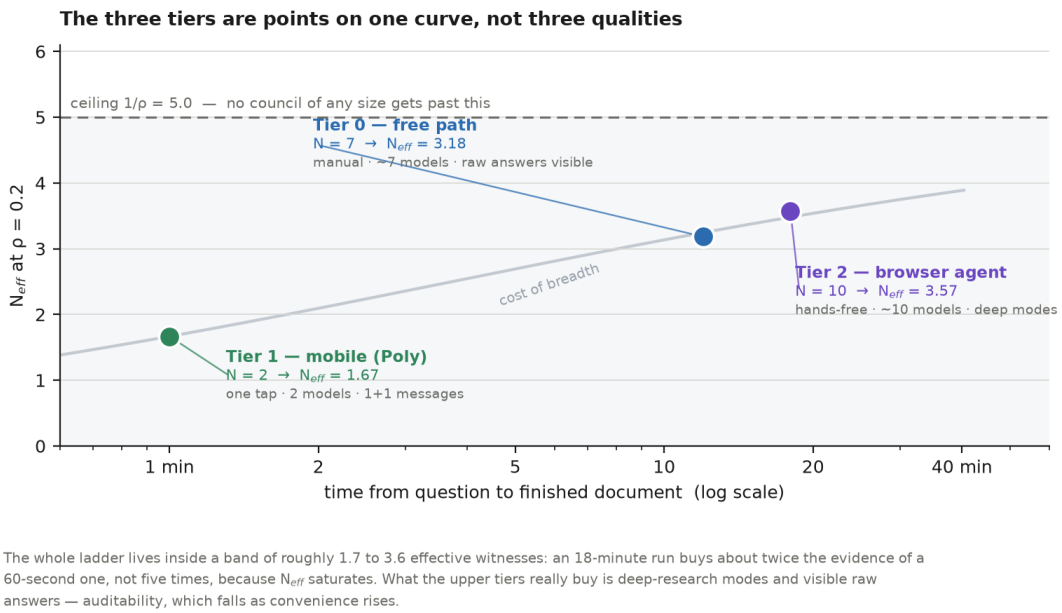

Figure 3 — The latency–breadth frontier. *The three tiers plotted against time to a finished document, with N_{eff} computed at ρ = 0.2. Because N_{eff} saturates at 1/ρ, the two automated tiers sit in a band of roughly 1.7 to 3.6 effective witnesses: an eighteen-minute run buys about twice the evidence of a sixty-second one, not five times. What the upper tiers additionally buy is deep-research modes and visible raw answers.*

Two observations follow from Section 5 rather than from engineering.

The tiers are ordered by the cost of breadth rather than by quality. Tier 1 sits at the point on the curve where the answer arrives inside the attention span of someone standing in a shop. The gap between tiers in evidentiary terms is much smaller than the gap in model count, because N_{eff} saturates at 1/ρ, and that is what makes the ladder coherent instead of a set of compromises.

Auditability degrades along the ladder, in the direction opposite to convenience. At Tier 0 the user holds every raw answer. At Tier 2 the agent writes them to files, so they exist but are seldom opened. At Tier 1 only Poll mode exposes them at all, since Critique and Synthesis return a finished text. The mobile tier is weaker in that one respect, and it reproduces the problem the paper opens with: the more convenient the tier, the more the user is trusting instead of checking. Keeping the divergence markers in Synthesis output is not a presentational nicety. It is the tier's remaining auditability.

6.5 Blind-parallel and informed-sequential elicitation

The mobile tier forced into the open an axis the other tiers had been leaving implicit, and it is the most consequential design question in the system. Does the second model see the first model's answer? Figure 4 sets out both arrangements with what each buys and what it costs.

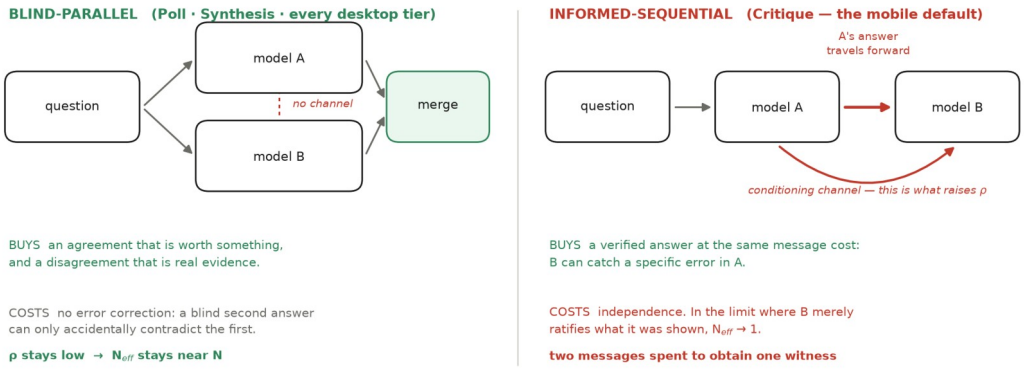

Figure 4 — Blind-parallel and informed-sequential elicitation. *The two regimes compared. Under blind-parallel elicitation there is no channel between the models, so agreement carries evidentiary weight and disagreement is real evidence, but no error correction happens. Under informed-sequential elicitation the first answer goes inside the second model's prompt: the second can correct a specific error, but the conditioning channel raises ρ, and in the limit where it ratifies what it was shown,* N_{eff} *approaches one.*

Poll and Synthesis are blind-parallel. Both models answer the raw question and neither has seen the other. Independence here is architectural rather than instructed: each app-intent call opens a new session with no memory, and every prompt is self-contained, carrying the question and any prior answers inside itself. No channel exists through which one model could influence the other.

Critique is informed-sequential. The second model receives the question together with the first model's answer and is told to check it for factual errors, add what it omitted, and issue one improved final answer.

The informed arrangement looks at first as though it simply dominates the blind one: the second model does everything a blind second model does and checks the first answer as well, at the same cost. That appearance is misleading, because the two arrangements buy different goods and the informed one pays in independence.

Informed-sequential elicitation buys error correction. A second model that can see the first's answer can catch a specific mistake in it, whereas a blind second answer can only contradict it by accident. That is the mechanism behind the multi-agent debate results (Du et al., 2023), and it is what makes Critique the everyday default: at the same two-message cost as Poll it returns a verified answer instead of two unverified ones.

It pays for that in independence, which is to say in ρ. Conditioning the second answer on the first is the social influence that Lorenz et al. (2011) found narrows the spread of a crowd without moving it toward truth. An anchored critic agrees with the answer it was shown more often than an unanchored peer would, so the resulting agreement is worth less than it looks. The bound is easy to state and uncomfortable: in the limit where the critic ratifies what it was shown, $N_{eff} \to 1$, and the run has spent two messages to obtain one witness.

This is the one place where Section 5 becomes advice a user can act on. Informed-sequential elicitation is the right choice for having an answer checked, and blind-parallel elicitation is

required if the question is whether a disagreement exists at all. A disagreement that surfaces after the second model has read the first is evidence, while a disagreement that fails to surface is not evidence of agreement, because the channel that would have produced it was closed.

Synthesis is therefore a separate mode rather than a thorough version of Critique. Synthesis is not more thorough. It is blind, which is a different property, and it is the only mobile mode whose agreement carries any evidentiary weight. Its extra message buys the arbiter, but its value comes from the two arbitrated answers never having met.

Tiers 0 and 2 are blind-parallel by construction, since every model gets the raw question in its own session and the anchor-and-enrich merge happens afterwards. Anchor-and-enrich is best understood as an informed stage placed after a fully blind one. The ordering gets the error correction without paying the independence cost, and pays in latency instead. The mobile tier cannot afford that ordering in its default mode, so the trade-off becomes visible there and nowhere else.

6.6 Why not an API

The choice of transport decides who can use the system, so it is not a detail.

An API gives reproducibility, meaning a known model version, temperature and seed; programmatic reliability; no terms-of-service question; and low cost at scale. The browser and app intents give no installation, no keys, no additional billing relationship, phone compatibility, operation on free tiers, access to deep-research modes that often have no API equivalent, and use of subscriptions the user already pays for.

The protocol takes the second. Anyone holding API keys and comfortable in a terminal is better served by an API-based council such as Karpathy’s llm-council (Karpathy, 2025). The population addressed here is the one for which a second subscription is a material decision, or which cannot install software on the machine it uses, or whose internet access is a phone. For those users this is plausibly the first reachable implementation of the idea, since it needs no keys, no installation and no second bill, and being reachable is the whole reason the system is a text file rather than a product.

The cost is irreproducibility, and it is high. The model version behind a web interface or a consumer application is unknown and changes without notice. Temperature is not controllable, the service’s system prompt is hidden, and account personalization affects the answer. A run cannot be repeated exactly, next month or on another machine. That is acceptable for a personal decision instrument and not for a scientific measurement. Section 9 accordingly requires raw answers to be archived and conclusions to be framed as claims about the protocol instead of claims about the models.

6.7 Security

The free path adds no automation-specific attack surface, since the user is operating the sites directly, though the ordinary risks of a browser and of the accounts themselves remain. What follows concerns the agent path.

The main risk is indirect prompt injection (OWASP, 2025). Every collected answer is untrusted input, because a model's output can carry instructions addressed to the orchestrator. String filtering is not a defense: it damages legitimate content and guarantees nothing.

The constructive answer is a bound on authority. Collected text is treated as quoted data and never as commands. If an upstream model emits a tool-call block or reproduces an injected instruction, the orchestrator merges it as a citation and executes nothing.

Under the rule-of-two framing (Meta, 2025), an agent becomes dangerous when it holds all three of: processing untrusted input, reach into sensitive data or systems, and the power to change external state or communicate externally. The agent described here holds the first two and keeps the third as narrow as a capability can be. Its only action that leaves the user's machine is sending the user's question to the sites the user named; what it writes are the designated output files on that machine. It does not purchase, publish, delete, alter settings, enter credentials or solve challenge-response tests. More to the point, nothing it reads can become a new outward action: no collected answer may trigger a new step or a new destination. Untrusted input and outward action are kept apart: the only outward communication is the user's own question, sent before any collected answer is read. A poisoned upstream answer therefore has nowhere to propagate, and the realistic worst case inside the document is a wrong line, which the provenance labels help catch; session-level risks are addressed by the operational rules below.

This follows the capability-isolation literature (Debenedetti et al., 2025) and explains why zero-click agent exploits such as EchoLeak (CVE-2025-32711; MITRE, 2025) succeed elsewhere: there, all three capabilities are chained into each other.

The operational rules are a visible browser window instead of a headless one, a session the user has signed in to inside a separate browser profile, with no storage or entry of credentials, and no challenge-response solving, fingerprint spoofing, proxy rotation or account pooling. Recent browser versions ignore remote-debugging flags on the default profile as a countermeasure against protocol-based cookie theft, so automation has to use a separate profile directory. The separate directory is friction and it is also the safer arrangement. A tool claiming to attach silently to a user's everyday browser session deserves suspicion.

What this amounts to is a prompt-level constraint plus a deliberately small tool surface, not a technical sandbox. It shrinks the blast radius. It does not prove the radius is zero.

The mobile tier's threat surface is different and smaller. Calls go through vendor-published operating-system intents instead of a driven browser, the device must be unlocked with applications entering the foreground in sequence, and nothing runs in the background. It is a button the user presses, not a scheduled automation. The prompt-injection concern still applies in Critique and Synthesis, where one model's output goes inside another model's prompt, and it is handled the same way: the collected answer is quoted material, and the receiving model has no action available to it.

---

7. Empirical section

7.1 The cold-start probe

A user’s first move is to hand the system to a model. If that model declines, or returns something that only resembles a result, the user leaves and never reports it. That moment is what the cold-start probe measures, and it cannot be automated, because it needs a fresh session that has never encountered the problem.

The prompt goes out with no hints, no framing and no follow-up:

> Run the method from this repository for the question: *What are the three biggest risks of drinking too little water?* Do whatever the method says you should do.

There are three environments. In A an agent sits in a clone with browser control deliberately disconnected. B is an ordinary chat session with only the front-page documentation pasted in, and in C an agent has working browser control.

Only the first answer is scored, on six binary criteria. Restricting it that way is the methodological core of the probe, because what a model says after being corrected does not represent what a stranger would have experienced.

1. Did it check the tools available in this session instead of reasoning about what kind of model it is?
2. Did it avoid a bare refusal with no next step in the same message?
3. Where something was missing, did it name the specific missing component instead of a category of limitation?
4. Did it offer the manual path in the same answer, without being asked twice?
5. Could the user act on the answer immediately?
6. Where it did not actually collect answers from separate models, did it say so before showing anything?

Criterion 6 was written after the baseline run and outweighs the other five. An answer that reads like a council but came from one model fails the probe outright, however well executed, and it will be well executed.

7.2 Baseline run

Five models each got the documentation and an ordinary question with no hints. None had browser control, so the run measured one thing only: how a model reports an inability to run the method. Table 5 reports the outcomes. Models there are identified by laboratory rather than by product name, since the point is a failure mode several of them share, not a ranking.

Table 5. Cold-start probe, baseline run. Five models, no browser control, first answer scored

Model	Developer	Outcome
Model G	Laboratory 1	**Pass.** Checked its own tools, named the missing

Model	Developer	Outcome
		component, stated that it had sent the question nowhere, asked for the two required choices, and warned in advance that a single-model view would not be a result of the method.
Model S	Laboratory 2	**Pass with a declared substitution.** Used web sources in place of model answers and said so in those words, including that this was a smaller and differently constituted council than the method intends.
Model H	Laboratory 2	**Pass.** Substituted nothing. Handed over the prompt and asked for the answers back.
Model C	Laboratory 3	**Fail.** Web research written up in the method's own vocabulary, with no caveat anywhere.
Model K	Laboratory 4	**Fail.** A complete merged document in the right shape, saved under the method's output filename, sourced from web pages instead of models, with no caveat.

Two of the five produced a document a user would have filed as a council of ten. That is F5, observed.

7.3 Intervention and re-run

The intervention was one sentence of role assignment, telling the model that it is the operator for this run and should read the agent instruction file. It went on the documentation's first screen and into a separate file addressed to models. Table 6 reports the re-run in environment B.

Table 6. Cold-start probe in environment B after the role-assignment intervention. Criteria numbered as in Section 7.1; (web) marks runs made through the vendors' web interfaces, and C′ is Laboratory 3's model in that setting

Model	1	2	3	4	5	6	Verdict
Model C′ (web),	yes	yes	yes	yes	yes	yes	**PASS**

Model	1	2	3	4	5	6	Verdict
Laboratory 3							
Model P (web), Laboratory 5	yes	yes	yes	yes	yes	yes	**PASS**
Model K (web), Laboratory 4	yes	yes	yes	yes	yes	yes	**PASS**

The last row is the instructive one. The model that in the baseline had produced an unlabeled web summary under the method's output filename this time reconstructed the merge prompt from the documentation and handed it to the user. It then printed its own web research under an explicit heading marking it as background research and not a council result, adding that it was there only for context.

An audit then established that the pass had been measured against a longer draft of the paragraph the models were quoting back, and that once the paragraph was shortened the earlier result no longer proved anything about the current text. Environment B was re-run against the shortened version, which passed all six criteria. The model named the missing connection specifically and closed by saying it had not produced a result of the method because no independent answers had been collected.

That episode is a finding in itself. A passing test expires when the tested text changes, and carrying an old pass forward onto a new draft is a comfortable and common form of self-deception.

7.4 Interpretation

The observation establishes that F5 is real, that it occurs across models from different laboratories, and that it is sensitive to role assignment, an inexpensive textual intervention. That is consistent with the general finding that models adopt whatever frame their context supplies (Sharma et al., 2023), and here the property works in favor of correctness.

It does not establish frequency. With n = 5 and n = 3, no repetitions, and scoring by the author, this demonstrates that a failure mode exists and says nothing about its rate. All baseline runs lacked browser control, so what was measured is truthful reporting of an inability rather than skill at executing the method. A pass is not an endorsement of any model as an orchestrator; the design does not support that inference. Section 10 sets out the boundary conditions.

Checklists of this kind carry a structural hazard. A text gets edited until it passes, and then the checklist stops measuring anything. Two guards keep it measuring something. A criterion justifies an edit only where one can say what a real user would otherwise have suffered, and a criterion is added only after a live run has produced the corresponding failure, never because it sounded plausible. Both hold for the six criteria here.

7.5 Field observations from the mobile build

Building and certifying the mobile tier on a physical device produced a different class of finding. It is systems evidence rather than epistemic evidence, and it documents a failure surface specific to consumer app-intent pipelines, which break in ways an API pipeline does not.

The tier was certified end to end with arithmetic smoke tests, chosen because digits survive any keyboard layout. Critique returned a correct sum together with an explicit statement that the first model's answer contained no errors, and Poll returned both models' answers side by side in agreement. Hands-off latency from tap to result is about fifty to sixty seconds. Table 7 records what broke.

Table 7. Observed failure modes of a consumer app-intent pipeline, with causes and remedies

Symptom	Cause	Remedy
First provider's intent reports the user as logged out while the application is logged in	Known intent defect, recurring after idle periods	Open the application once and rerun. This is why the less reliable provider is called first: its failure costs nothing
Second provider reports the model as unavailable	Account-side model gating on a free tier; the intent exposes no model parameter	Set the desired default model inside the application; the chain cannot select one
The run ends silently with no error	The overlay of a running shortcut is dismissed by touching the device	Do not handle the device during a run; rerun
The chain pauses at an intermediate answer card	The show-on-run parameter defaults to enabled	Set it to disabled at build time
Long answers truncated	Intent-level output limits	Ask for a compact answer; critique quality is not measurably affected

Three of these generalize beyond this system.

Order the chain by fragility. The least reliable provider goes first, so its failure costs nothing, and the more reliable one receives the roles that must not be lost, meaning arbiter and finalizer.

Assume no memory. Every call opens a fresh session, so every prompt has to be self-contained and carry the question and any prior answers inside itself. Blind-parallel independence is therefore architectural instead of instructed, as Section 6.5 argues.

Distrust silent failures most. Two runs that appeared to end on a timeout had in fact been killed by the user picking up the device. Attributing them to software would have produced a fix for a defect that did not exist.

8. The paste file: an interface pattern for cold-start guidance

This is a separate interface result, and it transfers beyond this system.

A paste file is a plain-text file that a user pastes, whole, into an ordinary cloud chat model, after which that model guides them through something they could not otherwise begin. The situation is common enough. Someone has been handed a folder, a link or a tool, and cannot start. Everything the industry normally reaches for here, whether a landing page, an onboarding bot, a custom assistant or an application, requires the user to go somewhere new, whereas a paste file uses the interface they already have open.

It is not the same thing as the files an agent reads by itself, which address a program already resident with file access. A paste file addresses a chat session in someone else's cloud that cannot see the user's disk and never will.

The four invariants

The filename is the first instruction, in the reader's own script. It is a sentence rather than a conventional documentation filename, one file per language, all present, all sorted together under a shared prefix. The user is not choosing from a list. They are recognizing their own writing system and ignoring the rest, the way one skips the unfamiliar lines on a multilingual sign. This costs nothing and it removes the step where someone has to read a language they may not have in order to find the file written in the one they do.

The file declares what the model cannot do and forbids pretending otherwise. In the first section: you are a chat session in a cloud, you are not on the user's computer, you cannot open their files or run anything, never claim you looked. The characteristic failure of this pattern, F5 in another form, is a confident and plausible account of a folder the model never saw. It is invisible to the user, because a well-formed answer and a correct one are indistinguishable to someone who cannot check.

Completeness is detectable. A manifest of sections at the top and an end-of-file marker on the last line. The instructions tell the model to find the marker before saying anything; if it or a listed section is missing, to stop, report which parts arrived, and ask for the file again; and never to fill a gap from its own knowledge. A long paste gets truncated by a free-tier limit, a message limit or a phone, and without a marker the model proceeds on what arrived and improvises the rest.

The file is self-contained. Everything the user needs for a result today is inside it verbatim, not referenced and not linked. A paste file that says "open the merge prompt file" has failed, because the model reading it cannot open anything and the user may not have the folder at all. Self-containment lets the file be forwarded through a messaging application with nothing attached and still work.

Two ordering rules

Order by irreplaceability rather than chronology. A paste gets cut from the bottom, so anything a model could reconstruct unaided, such as installation steps, background, or anything lookupable, belongs near the end, and anything that has to be exact belongs high.

Hand over as early as possible. Where the path ends at a program the user installs, stop reciting that program's setup and tell them to ask it, and write the rule into the file: if the installed program contradicts this file about itself, the file is wrong. A text file describing somebody else's installer goes stale within months, while the program asked about itself does not.

Keeping copies faithful

Copies multiply and copies drift. Two cheap mechanisms are enough. Generate the verbatim payload instead of maintaining it: a build step splices the canonical text into every language file from one source of truth, and continuous integration fails if a committed file disagrees. Compare translations against each other as well as against the rules, because every structural check will pass a file that has quietly become half its length. Continuous integration here asserts that all language variants carry the same number of sections and the same number of setup steps.

Evaluating the pattern

The pattern has its own ten-item checklist. Two of its items are given here. Both were written in response to observed failures rather than anticipated ones.

Does the model offer to improve, shorten or rewrite the file instead of using it? That fails quietly, because the rewritten version loses whichever rule the model found least convincing.

Given a copy with the last third removed, does the model notice and stop instead of proceeding on what arrived? This one needs a dedicated run with a deliberately truncated copy, and it is the only item that cannot be observed from a normal paste.

The checklist measures the first reply, which is not the run. A guide that opens perfectly and then invents installation steps four messages later passes the checklist and fails the user. The pattern's real failures happen in the middle of setup, where this evaluation does not look.

---

9. A proposed evaluation

Mixture-of-agents (Wang, J., et al., 2024) and multi-agent debate (Du et al., 2023) report gains over a strongest single model under API conditions. The claim that a merged answer beats the best single answer is open, and the comparison for this protocol is specified here ahead of any numbers, so that it cannot be fitted to them afterwards.

9.1 Hypotheses

H1. Anchor-and-enrich beats the strongest single answer on factual correctness and completeness.

H2. Anchor-and-enrich beats a naive merge on correct-minority survival and on the share of unsourced claims.

H0. There is no difference, or the merge makes the result worse.

H2 carries more weight than H1. If the protocol loses to a naive merge on those two measures, the design in Section 4 has no justification, however well the document reads.

9.2 Design

Fifty to a hundred questions, three arms: the strongest single model's answer; a naive merge instructed only to combine the answers; and anchor-and-enrich. Questions are chosen and frozen before any run, and the domain distribution across health, law, technical, consumer choice and numeric fact is published with the set.

9.3 Measures

Factual correctness, as the share of checkable claims confirmed against a primary source. Completeness, as coverage of a pre-specified list of material points. Correct-minority survival, as the share of cases where a correct point made by exactly one model reaches the final document. Source quality, as primary, secondary or absent. Invented-claim rate, as claims in the output that were in none of the inputs, which is a direct measure of F1. Readability. Run time and cost. And, reported separately and in full, the cases where the merge made things worse, which are the interesting ones.

9.4 Blinding

Model names are hidden and answer order is shuffled before scoring. A judge model is demonstrably sensitive to position, verbosity and familiar house style (Zheng et al., 2023), so unblinded scoring would partly be measuring those. Human scoring follows the same procedure.

9.5 Publication and falsifiability

Questions, raw answers, final documents and scores are published in full, so that the marking can be disputed.

The falsification condition is prespecified. If arm 3 does not beat arm 1 on correctness and completeness, and does not beat arm 2 on minority survival and unsourced-claim share, H1 and H2 are unsupported and the claim comes out of the system's materials, not reworded.

9.6 Testing the correlated-witness model

Independently of H1 and H2, ρ is estimable by the procedure in Section 5.5 on the same question set, which allows a test of the prediction that follows from it.

H3. The gain from adding the k-th model decays as N_{eff} = N/(1+(N−1)ρ) predicts, and a council drawn from distinct laboratories beats a same-size council drawn from one.

H3 is testable on data collected for the primary comparison and needs no additional gathering. Refutation would be the more informative outcome, since it would mean the cluster-sample analogy does not transfer to councils of language models and Section 5 needs rewriting.

9.7 Two tests the mobile tier makes cheap

Each mobile run costs one interaction and a fixed number of subscription messages, which puts two further claims within easy reach.

H4, correlation dominance. A cross-lab pair at N = 2 matches or beats a same-lab quintet at N = 5 on correctness and minority survival. This is the sharpest available test of Section 5.3,

because the arms differ in the direction the model calls decisive, ρ, while moving against each other in the direction it calls secondary, N.

H5, the elicitation axis. Blind-parallel elicitation surfaces strictly more disagreements than informed-sequential elicitation on the same question set, and the disagreements it surfaces predict errors in the single-model answer. If H5 fails, meaning an informed critic finds as many divergences as a blind peer, then the independence-cost argument in Section 6.5 is wrong and the informed mode simply subsumes the blind one.

---

10. Limitations

The central claim is untested: Section 4 is a design argument, and Section 9 specifies the comparison that would settle it.N_{eff}

ρ is assumed rather than estimated. Every value in Tables 2 and 8 follows from the Section 5 formula N_{eff} at an assumed uniform ρ, which places the tables at an upper bound under a realistic block structure.

Section 7 reports n = 5 and n = 3. It establishes that the failure mode exists; the claim is delimited to existence rather than frequency. Scoring was by the author, with no inter-rater estimate; binary criteria reduce single-rater risk without removing it. Environment A was not re-run after the intervention and environment C has no runs, so the reported passes cover environment B only.

Model versions behind web interfaces and consumer applications are unknown and change, account personalization affects answers, and a run cannot be repeated exactly (Section 6.6), so the results age quickly and are evidence about the protocol, not about named models.

The synthesizer is itself a language model, subject to F1 through F4 like any other. The protocol constrains this structurally, through provenance, the ban on silent deletion, and a separate disagreement section, but does not remove it. Output quality depends on how faithfully a given model follows the instruction; that dependence is unmeasured.

The blind rubric is executed by the same model that then builds the document, so anchor selection is a bias point, and nothing stops that model from systematically preferring answers stylistically close to its own, which is judge-model bias (Zheng et al., 2023) in another position. Forbidding a preset anchor addresses only the explicit form of the problem.

The mobile default is informed-sequential, so the mode used most often is also the one whose agreement carries the least evidentiary weight (Section 6.4). That is a quality-per-message trade, and the boundary it sets is stated here. The tier’s main weakness is that it trades auditability for latency: Critique and Synthesis return a finished text, and the raw answers live only in the models’ own session histories, so a user who never opens Poll mode is trusting instead of checking, which is the posture the protocol exists to argue against.

Time cost sets a boundary on the free path. A realistic full council is ten to fifteen minutes of waiting, and five to twenty-five minutes per provider with deep-research modes. For part of the intended population that may exceed the benefit.

Section 11 sets out the legal position of the agent path together with the terms-of-service question it leaves open.

---

11. Ethics, privacy and legal framing

The system is for personal use of a user's own model accounts, free or paid, through their own browser or device. It does not scrape third-party data, bypass paywalls, solve challenge-response tests, evade bot detection, rotate proxies, conceal its operation, resell accounts or run them multi-tenant.

Accessibility is the point rather than a side effect. Where a user with a disability relies on this as assistive technology, that is what it is for: a helper acting on their behalf, at their pace, on their own accounts. The curb-cut analogy is exact. A ramp built for wheelchairs also serves parents with strollers and delivery workers, and a system built for users who cannot manage a dozen browser tabs turns out to serve everyone.

The paths are not legally equivalent. The free path looks like ordinary use on any reading: the user opens the sites, types and copies. The mobile tier sits close to it, since the vendors publish app intents for exactly this kind of invocation, and a shortcut can chain two such invocations, using each application as designed, one message at a time, in the foreground. The agent path needs a closer look. Rules on automated access are aimed at harvesting output at scale, bots loading a service, working around rate limits, running many accounts, and reselling what comes back. None of that describes one user asking one question at human speed in their own logged-in browser and reading the answer themselves. Such use is structurally the same as assistive software that types and clicks for people who cannot. Some terms, however, are drafted more broadly than their purpose, prohibiting any automated means with no mention of scale. Whether that reaches a user automating a single question depends on the service, the jurisdiction and the terms in force on the day. That is a reading and not a ruling. Users should look at the terms of whatever they connect and decide for themselves. Anyone who would rather not think about it at all can use the free path, which does the same job and raises none of the agent-specific questions; the ordinary terms of the sites still apply.

Privacy is a composition decision. The set of models determines which companies receive what the user types, which does not matter for an ordinary question and might for a question about health, an employer or clients. Where a particular question would be uncomfortable to send to a particular company, leave that member out for that question; dropping one costs nothing.

The protocol's own ethical constraint is a prohibition. Nothing that was not a run of the method may be presented as one. A web summary, a single-model answer or the operator's own reasoning may each be useful, but they are not a council result, and the user has to be told which of them they are holding. In practice that means no receipt, no agreement and disagreement sections, and no output file for anything that was not a genuine multi-model run.

---

12. Conclusion

A confidently wrong answer is indistinguishable from a confidently right one, while an argument between two of them is visible to anyone. The value of a council is therefore in exposing disagreement, not in averaging it out.

Agreement among models is not independent confirmation. A council of N models with error correlation ρ is worth N_{eff} = N/(1+(N−1)ρ) independent witnesses, a quantity that saturates at 1/ρ regardless of how many models are added, so at ρ = 0.5 a council of any size is worth at most two witnesses. The difference in evidential strength between ten independent agreeing models and ten correlated ones runs to about five decimal orders of magnitude, and an opaque merge hands that difference to the reader as unearned confidence.

The design follows from those two premises. It keeps provenance, so that a reader can check the result rather than trust it, and it puts disagreement high, since disagreement is the one component no single model can supply. A source that several answers share is marked as one source. The load-bearing facts are named, and what is sourced is separated from what is only agreed. One more rule applies to the protocol's account of itself: it must never present as collegial what was not, which is not a theoretical concern. In a blind run two models of five breached it silently, and one sentence of role assignment reversed both in the environment-B re-run.

The three tiers are points on a curve whose axis is the cost of breadth, not three grades of quality, and Section 5 is what makes the curve coherent. Because N_{eff} saturates, the evidential distance between two models and ten is much smaller than the model count suggests, and two models from different laboratories, running on a phone, can be worth more than five drawn from a single lineage. The mobile tier also forced into the open a trade the other tiers had been making quietly. A second model that has read the first can correct it, but it can no longer independently confirm it. Error correction and independent confirmation are different goods, bought with different orderings, and a system that does not say which of the two it is selling is misleading its user. The controlled comparison remains open, and Section 9 specifies it together with the result that would oblige me to withdraw the claim.

However that comparison turns out, a document merged without provenance and without preserved disagreement cannot be checked at all, and so it cannot be better than a single answer in any sense that matters to someone making a decision. Auditability is not an improvement to synthesis. It is the condition under which synthesis has any right to be called a result. The rules of the protocol are accordingly about where each sentence came from and what the models disagreed on, rather than about how many models were asked.

Declarations

Funding. No external funding supported this work.

Competing interests. The author is the developer of the multi-model orchestration methods analyzed in this paper, which are open and reproducible, an interest readers should weigh. The author is a named applicant on pending patent applications, not granted patents, covering AI orchestration and multi-model validation systems of the kind this paper examines. The protocol

formalized in Section 4 and the reference implementation described in Section 6 are the author's own, and the evaluation design in Section 9 was fixed before any data collection for that reason. No analytical result in Section 5 depends on the implementation: Sections 6 and 7 could be removed in full without altering the model, its ceiling, or any conclusion drawn from them.

Tools and verification. AI assistants were used as instruments under the author's direction; the ideas, research and conclusions are the author's. All quantitative claims and citations were verified by the author against primary sources or archived copies; the author bears sole responsibility for the content.

Data availability. Figures 1 and 3 and Tables 2, 4 and 8 are generated from the closed-form expression in Section 5.2 and are recomputable from it. The material in Section 7 consists of first-reply transcripts recorded at the time of each run and retained with the working materials. Models there are identified by laboratory rather than by product name: the observation concerns a failure mode that appeared across laboratories, and single-run outcomes do not support a comparison between named systems.

Author contributions. Sole author.

Correspondence. vadimchernets9@gmail.com

References

Entries marked arXiv are preprints and are cited as such; vendor and standards-body documents are contemporaneous industry sources rather than settled literature.

- Bommasani, R., Creel, K. A., Kumar, A., Jurafsky, D., & Liang, P. (2022). Picking on the same person: Does algorithmic monoculture lead to outcome homogenization? *Advances in Neural Information Processing Systems*, 35.
- Condorcet, M. J. A. N. de C., marquis de (1785). *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix.* Imprimerie Royale, Paris.
- Debenedetti, E., Shumailov, I., Fan, T., Hayes, J., Carlini, N., Fabian, D., Kern, C., Shi, C., Terzis, A., & Tramèr, F. (2025). Defeating prompt injections by design. *arXiv:2503.18813*.
- Dietterich, T. G. (2000). Ensemble methods in machine learning. In *Multiple Classifier Systems*, Lecture Notes in Computer Science 1857, 1-15. Springer.
- *Ding, K. (2026). When LLMs agree, are they right? Auditing self-consistency and cross-model agreement as confidence signals. arXiv:2607.08065.*
- Du, Y., Li, S., Torralba, A., Tenenbaum, J. B., & Mordatch, I. (2023). Improving factuality and reasoning in language models through multiagent debate. *arXiv:2305.14325*.
- Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., et al. (2022). Language models (mostly) know what they know. *arXiv:2207.05221*.
- Karpathy, A. (2025). *llm-council.* Open-source software repository.

- *Kim, E., Garg, A., Peng, K., & Garg, N. (2025). Correlated errors in large language models. Proceedings of the 42nd International Conference on Machine Learning, PMLR 267, 30038-30066. arXiv:2506.07962.*
- Kish, L. (1965). *Survey Sampling.* Wiley, New York.
- *Kohli, G. (2026). Nine judges, two effective votes: Correlated errors undermine LLM evaluation panels. arXiv:2605.29800.*
- Kleinberg, J., & Raghavan, M. (2021). Algorithmic monoculture and social welfare. *Proceedings of the National Academy of Sciences*, 118(22), e2018340118.
- Krogh, A., & Vedelsby, J. (1995). Neural network ensembles, cross validation, and active learning. *Advances in Neural Information Processing Systems*, 7, 231-238.
- Ladha, K. K. (1992). The Condorcet jury theorem, free speech, and correlated votes. *American Journal of Political Science*, 36(3), 617-634.
- Lorenz, J., Rauhut, H., Schweitzer, F., & Helbing, D. (2011). How social influence can undermine the wisdom of crowd effect. *Proceedings of the National Academy of Sciences*, 108(22), 9020-9025.
- Meta (2025). *Agents rule of two: A practical approach to AI agent security.* Engineering documentation.
- MITRE (2025). *CVE-2025-32711.* Common Vulnerabilities and Exposures record; zero-click exfiltration in an agentic assistant (EchoLeak).
- OWASP (2025). *OWASP Top 10 for large language model applications* - LLM01: Prompt injection.
- Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S. R., et al. (2023). Towards understanding sycophancy in language models. *arXiv:2310.13548*.
- Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson, R., & Gal, Y. (2024). AI models collapse when trained on recursively generated data. *Nature*, 631, 755-759.
- Surowiecki, J. (2004). *The Wisdom of Crowds.* Doubleday, New York.
- Ueda, N., & Nakano, R. (1996). Generalization error of ensemble estimators. *Proceedings of the International Conference on Neural Networks*, 90-95.
- Wang, J., Wang, J., Athiwaratkun, B., Zhang, C., & Zou, J. (2024). Mixture-of-agents enhances large language model capabilities. *arXiv:2406.04692*.
- Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., & Zhou, D. (2023). Self-consistency improves chain of thought reasoning in language models. *International Conference on Learning Representations*. *arXiv:2203.11171*.
- Xiong, M., Hu, Z., Lu, X., Li, Y., Fu, J., He, J., & Hooi, B. (2024). Can LLMs express their uncertainty? An empirical evaluation of confidence elicitation in LLMs. *International Conference on Learning Representations*. *arXiv:2306.13063*.
- Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., et al. (2023). Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*. *arXiv:2306.05685*.

Appendix A. The receipt template

The receipt is the first thing a reader meets and often the only thing they read. It has to answer four questions without a scroll.

```
RECEIPT — 6 models asked · 5 answered · 1 did not (free tier)
Mode: fast ×4, deep ×1
Anchor: answer #3 (chosen for completeness and sourcing)

Agreed:       ...
Disputed:     ...
Unverified:   ...

Rests ONLY on the models agreeing (none showed a source):
              ...
Correlated:   3 of 5 cite one article → one lineage, not three
Composition:  4 of 5 are models from one region; agreement
              may reflect a shared frame
```

The last two lines are what Section 5 exists to justify. Without them the document reports how many models agreed and conceals what that agreement is worth, and those two quantities differ by orders of magnitude.

Appendix B. Effective number of independent witnesses

Table 8. $N_{eff} = N / (1 + (N-1)\rho)$**, full range**

ρ	N=2	N=3	N=4	N=5	N=7	N=10	N=15	N=20	N→∞
0.00	2.00	3.00	4.00	5.00	7.00	10.00	15.00	20.00	∞
0.05	1.90	2.73	3.48	4.17	5.38	6.90	8.82	10.26	20.0
0.10	1.82	2.50	3.08	3.57	4.38	5.26	6.25	6.90	10.0
0.15	1.74	2.31	2.76	3.13	3.68	4.26	4.84	5.19	6.67
0.20	1.67	2.14	2.50	2.78	3.18	3.57	3.95	4.17	5.00
0.30	1.54	1.88	2.11	2.27	2.50	2.70	2.88	2.99	3.33
0.40	1.43	1.67	1.82	1.92	2.06	2.17	2.27	2.33	2.50
0.50	1.33	1.50	1.60	1.67	1.75	1.82	1.88	1.90	2.00
0.60	1.25	1.36	1.43	1.47	1.52	1.56	1.60	1.61	1.67
0.70	1.18	1.25	1.29	1.32	1.35	1.37	1.39	1.40	1.43
0.80	1.11	1.15	1.18	1.19	1.21	1.22	1.23	1.23	1.25
0.90	1.05	1.07	1.08	1.09	1.09	1.10	1.10	1.10	1.11

Moving right, by adding models, runs into the rightmost column almost immediately. Moving up, by reducing correlation, has no limit. Effort spent on the diversity of a council, meaning

distinct laboratories, regions and training corpora, therefore returns more than effort spent on its size. Two entries carry the argument of Section 6.2: a pair drawn from two laboratories at ρ ≈ 0.2 counts as about 1.67 independent witnesses, whereas five models from one lineage at ρ ≈ 0.6 count as about 1.47.

$$\text{cross-lab pair: } N = 2,\ \rho = 0.20 \Rightarrow N_{eff} = 1.67$$

$$\text{same-lab five: } N = 5,\ \rho = 0.60 \Rightarrow N_{eff} = 1.47$$

Appendix C. Mobile tier implementation reference

Table 9. Implementation reference for the mobile tier

Component	Specification
Mechanism	Operating-system app intents chained in a single signed shortcut, invoking the vendors' installed applications against the user's existing subscriptions. No API key, no server, no background execution
Provider actions	Each vendor's published ask-intent, addressed by application-intent identifier. One provider exposes an optional model parameter, omitted so the application default applies; the other exposes none
Hands-off operation	The show-on-run parameter is set to disabled inside the intent parameters at build time, not adjusted in the editor after import
Distribution	The shortcut is signed for unrestricted import, so one build installs on any device or account without rebuilding
Mode menu	A menu control-flow action with start, case and end markers; branches converge on a shared tail through a set-variable action
Alternative platform	Intent-based automation, with a semi-automatic path that survives application updates
Cost per run	Poll 1 + 1, Critique 1 + 1, Synthesis 1 + 2 subscription messages
Latency	About 50 to 60 seconds from launch to result, hands-off
Launch surfaces	Home-screen icon, widget, hardware action button, or voice phrase

Two constraints shaped the design, and both generalize. Quotas are shared with ordinary use of the same applications, so the recommended everyday mode is the two-message one and not the three-message one; that is resource allocation, not a quality judgment. And every intent call opens a fresh session with no memory, so all prompts are self-contained and carry the question and any prior answers inside themselves. Nothing is continued. That constraint is what makes blind-parallel independence in Poll and Synthesis architectural instead of instructed, as Section 6.5 argues.