idk.

Uncertainty Is Infrastructure

Why the next layer of the AI stack is the one that says

"I don't know."

// idk, inc. · white paper v1.0 · august 2026 · idk.io

confidence in this document: 91% (calibrated)

¯_(°_°)_/¯

// ABSTRACT

The most expensive words in AI are the wrong ones said confidently.

Large language models have a structural defect: they deliver correct answers and fabricated ones in the same fluent, assured voice. The model that aces your benchmark and the model that invents a court citation are the same model, on the same day, at the same temperature. What is missing is not intelligence. It is **calibration** — a measurable, auditable link between how confident a system sounds and how often it is right.

This paper argues that calibrated uncertainty is not a model feature to be fine-tuned in, but a distinct **infrastructure layer** — as separable from the model as authentication is from the application. We introduce idk, a model-agnostic confidence engine that scores every AI output, attaches evidence, and abstains — returns a structured “I don't know” — when confidence falls below a caller-defined threshold. We describe the engine's architecture, its abstention economics (you pay for answers, not shrugs), and the market forces — regulatory, legal, and reputational — that make honesty a purchasable primitive.

> *Overconfidence is a bug. “I don't know” is a feature.*

// 01 · THE PROBLEM

Fluency has outrun honesty.

Every serious deployment of generative AI eventually meets the same failure: an output that was wrong, and — more importantly — *gave no signal that it might be*. Lawyers have filed briefs with invented case law. Support bots have promised refund policies that never existed. Coding assistants import packages that were never written. The failures differ in domain but share one anatomy: the error arrived dressed exactly like the truth.

This is not a bug being patched out by scale. It is a direct consequence of how these systems are built and rewarded:

→ **Training rewards an answer, any answer.** Next-token prediction optimizes for plausible continuation, not for truthfulness — and never for silence. There is no gradient toward “I don't know.”

→ **Benchmarks reward guessing.** On a multiple-choice eval, an abstention scores zero and a lucky guess scores full marks. Models are literally trained by test-taking incentives every student knows: never leave it blank.

→ **Human feedback rewards confidence.** Raters prefer assured, complete-sounding answers. Hedging gets down-ranked. Over thousands of updates, the model learns that swagger is what humans want — and swagger is what we get.

The result: a system whose **tone carries no information**. Confidence, the signal humans have used to weigh testimony for our entire history as a species, has been decoupled from accuracy and turned into a constant.

> *A wrong answer is a bug. A wrong answer that sounds right is a liability.*

// 02 · THE MISSING PRIMITIVE

Calibration: making 85% mean 85%.

A system is **calibrated** when its stated confidence matches its empirical hit rate: of all answers issued at 85% confidence, roughly 85% are correct. Calibration does not make a model smarter. It makes a model **legible** — it converts output from an assertion into a priced claim you can build policy on top of:

- → Above 95%: ship the answer straight to the user.
- → 60–95%: show the answer with its score and evidence.
- → Below 60%: abstain — route to retrieval, retry, or a human.

Note what this buys. Nothing about the underlying model changed, yet the deployment went from “hope it's right” to a **risk-managed pipeline** with an audit trail. That is the difference between a demo and infrastructure. Auth, payments, and observability all made the same journey: from feature every team hacked together badly, to layer everyone buys. Uncertainty is next, for three converging reasons:

- → **Regulation.** Emerging AI rules (the EU AI Act among them) require risk management, transparency, and human oversight for high-risk systems. An abstention log — *what did the system decline to answer, and why* — is the cleanest artifact of oversight yet invented.
- → **Liability.** Courts and customers increasingly treat AI output as the deployer's speech. The company that can show calibrated review of every output is in a different legal universe from the one that shipped raw model text.
- → **Agents.** Autonomous AI that acts — books, buys, files, deploys — cannot be reviewed answer-by-answer by a human. The only scalable control is a machine that knows when to stop and say idk.

// 03 · THE ENGINE

How idk measures what a model doesn't know.

The idk engine sits between any model and its consumer. Input: a prompt, an answer, and optionally the sources the answer should rest on. Output: a single calibrated confidence score, the evidence behind it, and a decision — **answer or abstain**. The score is fused from independent signals, because no single test for machine ignorance survives contact with production:

01 / SELF-CONSISTENCY	Resample the model several times and measure *semantic* agreement — do the answers mean the same thing, not just repeat the same words? Knowledge is stable under resampling; confabulation wobbles.
02 / EVIDENCE COVERAGE	Decompose the answer into atomic claims and check each against retrieved sources. Every unsupported claim drags the score down. An answer can be fluent, consistent — and still unsupported. That's a hallucination with good posture.
03 / MODEL SIGNALS	Where available: token-level probabilities, entropy over answer variants, and disagreement across different models asked the same question. Cheap, fast, and surprisingly informative at the extremes.
04 / CALIBRATION MAP	Raw signals are fused and mapped to real-world accuracy against a continuously updated, domain-specific calibration set — so the number that comes out has an empirical meaning, not a vibe. Legal 85% and trivia 85% are different curves.

Crucially, evidence is **fetched live**. At answer time the engine issues real web searches, retrieves current sources, and grounds every claim against the internet as it exists today — not as the model remembers it from training. Grounding raises the calibration ceiling: a current-events query that honestly caps near 50% ungrounded can clear 90% with fresh evidence attached, and every answer carries the links it stood on.

Below the caller's threshold, the engine does not merely refuse. It returns a **structured abstention**: machine-readable reason codes (no grounded evidence; unstable under resampling; future contingent; out of domain), the nearest verifiable partial answer if one exists, and a recommended fallback route. An idk from the engine is not a dead end — it is a routing instruction with a paper trail. Every decision, answered or abstained, lands in the **confidence ledger**: an append-only audit log built for the compliance officer who will, one day soon, ask exactly what your AI declined to say and why.

`// 04 · THE PRODUCTS`

One engine, three surfaces.

idk engine — the API

Calibration as a service. Two lines of integration: send `{prompt, answer, sources}`, receive `{confidence, evidence, abstain, reasons}`. Model-agnostic by design — we score outputs from any provider, open or closed. Developers choose thresholds per route: 0.99 for medication dosages, 0.60 for brainstorming.

idk verify — the firewall

Production middleware for teams already shipping AI. Verify audits every output in-flight, flags overconfident claims, blocks or reroutes below-threshold answers, and maintains the confidence ledger. Deploys as a proxy or SDK; on-prem for regulated industries. This is the product the EU AI Act sells for us.

idk ask — the showcase

A consumer answer app with one radical property: every answer shows its confidence score, and some answers are simply “idk.” Ask is deliberately small as a business and deliberately loud as a brand — it teaches the market to *expect* a confidence score next to every AI answer, the way padlock icons taught users to expect HTTPS. Every skeptical power-user it wins is a future enterprise champion.

`// 05 · THE BUSINESS`

Abstention economics: pay for answers, not shrugs.

idk prices on a simple, self-aligning rule: **metered checks, with abstentions free**. When the engine lets an answer through, that check bills. When it says idk, it doesn't. The incentive alignment is the point — we make money only when we are confident enough to stand behind an answer, which means our revenue is a bet on our own calibration. No other vendor in the AI stack has its income statement wired to its honesty.

→ **Curious** — free tier: 10k checks/month, community calibration sets. For developers and skeptics.

→ **Confident** — usage tier: verify middleware, per-route thresholds, domain-tuned calibration, ledger export.

→ **Certain*** — enterprise: on-prem or VPC deployment, compliance reporting, dedicated calibration engineering. (*Nothing is certain. We're the company that tells you that.)

The moat is the calibration data. Every check that flows through the engine — every prediction later confirmed or contradicted by outcomes, corrections, and human review — sharpens the calibration map. Confidence scoring is a business where the product measurably improves with every customer, and where the incumbent's error bars are always narrower than the challenger's. Models commoditize; ground truth about *when models fail* compounds.

`// 06 · THE FLYWHEEL`

Search once, answer forever.

A conventional AI answer is an expense: tokens are burned, the answer is shown, and the work evaporates. The next person to ask the same question pays for it again. idk inverts this. Every answer that clears the confidence threshold is **minted as an asset** — archived as a permanent, statically rendered, crawlable page with a canonical URL. Generation cost is paid once; the answer is served forever.

The address is the query itself, stored in a deliberately unique way: the search text is normalized and **truncated to a readable slug**, then made collision-proof with a short content hash — `/p/best-mechanical-keyboard-under-100-k3x9f2`. Deterministic slugs mean a repeat search resolves to the stored page before any model is invoked: the archive doubles as a zero-token cache. One person's curiosity becomes the next person's instant, free answer.

→ **Mint.** Above-threshold pages are rendered to static HTML — title, truncated meta description, structured data, sources, confidence score, timestamp — and written to the archive under their slug.

→ **Index.** The archive feeds an auto-generated sitemap. Crawlers see fast, canonical, information-dense pages: programmatic SEO where the editorial calendar is the user base's curiosity.

→ **Compound.** Search-engine traffic lands on archived answers; those visitors ask new questions; new pages are minted. Content, traffic, and calibration data all grow as a function of usage, not budget.

The quality gate is what makes this defensible: **abstentions are never indexed**. Below-threshold answers do not get URLs. Content farms publish everything and hope; idk publishes only what it can stand behind, every page wearing its confidence score and archive date. It is an SEO strategy structurally incapable of spam — the index admits nothing that failed τ.

> *Every question answered once is a page the internet never has to ask for again.*

`// 07 · THE CONCEPT LAYER`

Resolution before search: merging Nuclei.

For the bounded class of concepts with single-term names, idk no longer searches at all. It **resolves**. The engine now sits on top of Nuclei.si, a canonical concept layer: a register of single-term .si authorities (451 held, 65 serving at the time of writing) where each node serves exactly one signed, content-addressed JSON-LD Concept Record at a constructible address. A query that normalizes to a registered term — `physics`, `zakat`, `catholicism` — goes straight to `https://<term>.si/api/v1/concept`: no ranking, no source selection, no guessing which of ten links is authoritative.

The two systems close each other's open problems. Retrieval on the open web fails three ways — address ambiguity, unverifiable quotation, provenance collapse — and those are precisely the uncertainties idk's evidence signal must price in. A Nuclei record removes all three by construction: the address is the concept, the bytes verify against a SHA-256 digest over the record's RFC 8785

canonical form, and corpus method is a required, machine-readable field. idk verifies in the client — canonicalization, digest comparison, Ed25519 signature where the platform allows — and labels exactly what was checked. In return, idk gives the concept layer what its own specification concedes it lacks: a consumer at scale, and the evaluation telemetry (hit rate, steps saved, token delta, verification failures) its evaluation section calls for.

→ **The evidence ladder.** Every query now descends four rungs: the archive (our own past answers — free); the concept layer (signed canonical records — verified); the live web (fresh but unsigned); the model prior (scored ruthlessly). Confidence calibrates per rung.

→ **Honest weighting.** A signature binds authorship, not truth — Nuclei's own words. Records raise grounding confidence for definitional and conceptual content; they do not mint certainty, and coverage is bounded to single-term concepts.

→ **Graph traversal in the UI.** Each record carries typed edges to neighbouring concepts. In idk, edges render as one-click searches — moving through the concept graph without ever returning to a ranked results page.

Resolution before search; verification before trust.

// 08 · THE ROAD

From scores to stop-signs.

→ **Phase 1 — Score** (now): engine API and ask in public beta. Calibrated confidence on question-answering across open domains; first domain packs (legal, support, code).

→ **Phase 2 — Gate**: verify in production. Policy engine — thresholds, fallback routes, human-in-the-loop queues — plus the confidence ledger and compliance exports.

→ **Phase 3 — Govern**: calibration for agents. Pre-action confidence checks (“how sure are you before you click buy?”), abstention-aware planning, and fleet-level uncertainty dashboards: a single pane showing everything your AI declined to do, and why.

The end state is simple to describe: a world where no consequential AI action ships without a number attached, where “the system wasn't sure” is a logged, auditable event rather than a post-mortem finding — and where the three most useful words in computing are finally the three humblest.

// CODA · THE MANIFESTO

We taught machines the three hardest words.

Overconfidence is a bug.

File it. Fix it. Ship the patch.

“I don't know” is a feature.

The most underrated string in computing.

Trust ships with a number.

No score, no deploy.

Calibration beats charisma.

85% should mean 85%.

Abstain, then escalate.

A shrug with a plan attached.

Honesty compounds.

Every admitted unknown buys the next answer credibility.

idk, inc. · hello@idk.io · idk.io

This document describes a product architecture and design targets. Illustrative thresholds and examples are exactly that — illustrative. We told you we'd be honest about uncertainty; that includes our own.