WHITE PAPER · AI TRUST & SAFETY

Grounded by Construction

How Bizer guarantees its AI agent never fabricates your numbers

A white paper on Bizer’s number anti-hallucination system

Executive summary

> Bizer's AI business partner, **Miles**, reads your invoices, books, and records and answers questions about your money. For a tool like this, one failure mode matters more than any other: **inventing a number.** A business assistant that confidently reports a price, total, or invoice that isn't real is worse than no assistant at all — it quietly corrupts the owner's decisions.
>
> Large language models, left to themselves, *can* do exactly that. When asked for something that isn't in the data, a model will often produce a plausible-looking answer rather than admit it doesn't know. No amount of careful instruction reduces this to zero, because instructions are a *request* to a probabilistic system, not a *guarantee*.
>
> So we did not rely on instructions. We built a **deterministic verification layer that sits outside the model** and inspects every answer before it reaches you.

Every figure Miles shows you must either be quoted verbatim from a source it actually read, or be the checked result of arithmetic over numbers that were. Any number it cannot ground is blocked before you ever see it.

Because the check is plain arithmetic and string matching — not model judgment — a number that isn't in your data **cannot be displayed.** This document explains how that works, exactly what it does and does not promise, and how the guarantee is kept honest over time.

1. Why "usually right" isn't good enough

A language model generates text by predicting what is likely to come next. That is extraordinarily useful for drafting, summarizing, and reasoning — and it is precisely the wrong property for reporting financial facts. Asked "what did I pay for the ribeyes?", a model that hasn't found ribeyes in the file may still write a confident dollar amount, because a confident dollar amount is what such an answer usually *looks like*.

Better prompting ("only state figures that appear in the source", "never guess") moves the needle. It does not close the gap. A prompt is an instruction the model may follow most of the time; "most of the time" is not a standard you can put under a business owner's financial decisions.

The conclusion that shaped our design: **trust in the numbers must be enforced by construction, not asked for in instructions.** The model can draft the answer. It cannot be the final authority on whether a number is real.

2. The incident that set the bar

The system was driven by a concrete failure we treat as a permanent test case.

The invoice that wasn't there. An owner uploaded a supplier invoice — a table of real line items — and asked what they paid for "bison ribeyes." There were no bison on the invoice. Instead of saying so, the agent fabricated two line items at **$100.00** each, under invoice numbers **#123456** and **#123457** — none of which appeared anywhere in the file. The real document contained Certified Angus *beef* ribeyes at **$1,014.13** and **$971.36**, and a genuine combined total of **$10,187.30**. The correct answer was simply: *"There's no bison in this file; the ribeyes I see are beef at $1,014.13 and $971.36 — did you mean those?"*

This single incident maps to a complete family of failures: inventing a figure, doing bad arithmetic, attaching a real figure to the wrong item, and caving to pushback by inventing specifics to agree. The system addresses each one.

3. The core principle

Everything follows from a single, non-negotiable rule:

> **Miles must never be the *source* of a fact about your data.**

Concretely, every figure in an answer must be one of exactly two things:

- **Quoted** — it appears, verbatim, in something the system actually retrieved on this turn (a file you uploaded, your own message, or the result of a live lookup).
- **Computed** — it is the exact result of arithmetic over numbers that are themselves quoted, *with the components shown* so the math can be re-checked.

Anything else is **ungrounded**, and ungrounded content is blocked before display.

4. How it works: a verification gate outside the model

The heart of the system is a **gate**: a deterministic checker that runs after the model has drafted its reply and before anything is shown to you. It is ordinary, auditable logic — number parsing, exact string comparison, and integer arithmetic — with no model judgment involved.

```
+--------------------------------------------------------+
|  Evidence gathered this turn                           |
|  - the file/photo you uploaded                         |
|  - your own message                                    |
|  - results of every live lookup the agent ran          |
+--------------------------------------------------------+
                        |
the model drafts an answer        | checked against
                        v         v
+--------------------------------------------------------+
|  THE GATE  (deterministic - no model judgment)         |
|  Every figure must be QUOTED or COMPUTED.              |
+--------------------------------------------------------+
      |                                    |
all grounded                       something ungrounded
      |                                    |
      v                                    v
shown to you            ask the model to correct (bounded retries)
                                           |
                               still ungrounded?
                                           v
                        an honest reply with NO unverified number
```

Three properties matter:

- **It is enforced in code.** The gate runs on Bizer's servers, inside the path that produces every answer. The model cannot talk its way past it, and the client cannot skip it.
- **It does not depend on the model behaving.** A number that isn't in your data fails an exact comparison. That is not a matter of the model "trying" to be accurate; it is arithmetic.
- **It fails safe.** When the gate rejects an answer, the model is given a chance to re-ground it (for example, to quote the real figure or to say it can't find one). If it still can't, Bizer shows a plain, honest message that contains **no** unverified number — never the fabrication.

Every block is recorded, so the rate at which the model *attempts* an ungrounded answer is itself monitored as an early-warning signal.

5. What the gate guarantees

The gate eliminates a specific, complete set of failure modes. In plain terms:

Failure	What it looks like	How the gate handles it
Invented figure	A price or ID that exists nowhere in your data	Blocked — it isn’t in any source
Bad arithmetic	A total that doesn’t add up from real line items	Blocked — the shown components must sum exactly
Right number, wrong item	A real amount attached to the wrong product	Blocked — the figure must sit with the item it’s claimed for
Capitulation	Flipping a figure just because the user pushed back	Re-checks the records and reports what they say; won’t invent specifics to agree

A useful way to read the table: the agent can no longer **invent**, **miscalculate**, or **mis-attribute** any figure about your data.

6. Defense in depth

The gate is the hard guarantee, but it sits on top of three reinforcing layers.

1. Faithful retrieval — get the real numbers in front of the model. A figure can only be quoted if it was actually seen. When you upload a document or photo, Bizer reads it at **high fidelity**, so dense material like a supplier invoice — with its small per-unit prices and weights — stays legible rather than being blurred into an approximate “gist.” Questions about your records are answered by querying those records directly, not by reading whatever summary happens to be on screen.

2. Generation discipline — make grounding checkable. Miles is instructed to quote the figures it relies on, to show the line items behind any total it calculates, and — importantly — to value **accuracy over agreement**: when an owner insists it’s wrong, it re-checks the data rather than simply caving.

3. The deterministic gate — the guarantee. Described above. This is the layer that turns “should be accurate” into “cannot display an unverified number.”

Underpinning all three is a deliberate definition of **what counts as evidence**: the files you upload, your own messages, and the live results of looking up your records.

7. Trustworthy without being annoying

A verification system that frequently blocks *correct* answers would be its own kind of failure. Our design treats false rejections as a real cost and minimizes them:

- The gate is **conservative by construction.** It flags only high-confidence cases. Where a claim is genuinely ambiguous — a list of items, a figure with no clear subject, a number we can't confidently classify — it deliberately holds back rather than risk blocking a good answer.
- When it does block, the **first response is to ask the model to correct itself**, which usually yields the real figure. The plain fallback message appears only if re-grounding fails.
- The worst case the owner experiences is therefore a **clarifying question** ("I can't confirm that figure — point me to the exact record") — never a confidently wrong number.

We accept this trade deliberately: a blocked-but-correct answer costs a sentence; a displayed-but-fabricated number costs trust.

8. Assurance and non-regression

A guarantee is only as good as its durability. The incident in Section 2 — and the full family of failure modes in Section 5 — are encoded as **automated tests built from the real case.** They run on **every change** to the system. If a future update were to weaken any guarantee, the build fails before it can ship. The protection cannot silently erode.

9. Honest limits

For a trust document, candor about the boundaries matters as much as the guarantees.

- **It eliminates fabrication, not source error.** If a document *you* provide contains a typo, Miles will faithfully report the typo — but it always tells you which source a figure came from, so the error is visible and traceable, never invented by the agent. "Garbage in" is out of scope by design; silent invention is not.
- **It governs figures about *your* data.** The guarantee is about figures concerning your records — not general knowledge about the outside world, which is handled separately and with its own sourcing.
- **Rare false rejections err toward asking.** When the checker is unsure whether two forms of a number are equal, it treats the figure as unverified and asks rather than guesses. The cost is an occasional extra question, never a wrong answer.

We consider stating these limits plainly to be part of earning trust, not a caveat to bury.

10. Why this is different

It is worth being precise about what makes this more than “a careful prompt.”

- **Prompting** asks the model to be accurate. It reduces fabrication; it cannot guarantee the absence of it, because the model remains the final author of the number.
- **Retrieval** (putting the right documents in front of the model) improves the *inputs*, which helps — but the model can still misread, miscompute, or mis-attribute what it retrieved.
- **Bizer’s gate** verifies the *output*. It removes the model from the position of final authority over any figure: a number survives to your screen only if deterministic logic confirms it is real. That shift — from persuasion to enforcement, and from trusting the model to verifying its result — is the whole point.

11. The guarantee, restated

> Bizer’s agent cannot show you a number about your business that isn’t either taken directly from your data or computed, with its components, from numbers that are. This is enforced in code, outside the model, and protected by tests built from a real failure — so the protection holds as the product evolves.

For a business owner, that means the figures Miles puts in front of you are figures you can act on.

Glossary

- **Grounded** — backed by evidence the system actually gathered this turn.
- **Quoted figure** — a number that appears verbatim in a source that was read.
- **Computed figure** — a number derived by checked arithmetic over quoted numbers, with the components shown.
- **Evidence** — the files you upload, your own messages, and live lookups of your records.
- **The gate** — the deterministic check, run outside the model before any answer is shown, that blocks ungrounded figures.