Does a Glossary Stop the AI Drift? Eight Steps Until It Works

8 November 2026 · 12 min

> Let an AI model the same process several times, and the same step gets a different name every time. The obvious answer is a glossary. I measured it, in four test series – and at first the glossary achieved absolutely nothing.

In the last episode I showed that if you let the AI model the same process several times, you get something different every time. A single customer order had 32 different names. The obvious answer is a glossary. You tell the AI which words apply, and the problem is solved.

That is exactly what I did. And I was astonished when, at first, the glossary achieved absolutely nothing. This article sums up why, and what it takes for a glossary to really work. The video for this episode tells the same story with the screens from the lab (video). The documents and the glossary from the experiment can be downloaded; see the end of the article.

What drift is, and why it is not just chance

Drift means: the same documents, the same request to the language model, several runs. The structure often stays similar, but the words wander. A step is called "Check order" once, then "Validate order", then "Verify order".

Part of this is intended. Language models roll the dice a little when they pick the next word; the temperature controls that. You can set it to zero. I did, and still my pre-tests produced different sub-processes every time. This is not an isolated case: even at temperature zero, language models demonstrably give different answers to the same question (Ouyang et al., 2025). So stability is not something you can simply switch on.

The larger part of the drift has a different cause anyway. It lies in the documents.

The case: 4Health

The experiment runs on the "Sales & Order Management" project of the fictional company 4Health, a maker of sports bars with an online shop, a bar configurator and trade partners, with sites in five countries. The process is "B2C order-to-cash", from the order in the shop to the payment.

It comes with 28 documents of the kind that lie around in any company: a process description, the terms and conditions, an interface specification for the ERP, interview notes, e-mails from Munich and Italy, a training course from 2023, Excel lists. And as in any company, they contradict each other, above all in their words. The process description says "Order", the German e-mail "Bestellung", the

French manual "commande client", the ERP "Sales Order". There are six words for the credit note and seven for the dunning letter.

So the AI does nothing wrong. It picks a different word in each run, because nobody decided which one applies. That is exactly what a glossary is for: a list of terms, not of words. Terminology science has long kept these two levels apart: a concept is the thing, the designations are the words for it, and one concept can have several (ISO 704:2022).

Without a glossary: 190 names

First, the baseline. The AI builds everything itself: first the analysis with the activities, then the model with all sub-processes, about 50 activities per run. Five runs, without a glossary, with the documents only.

Over five runs, **190 different activity names** came together. Around the invoice alone there were 17 different activities. "Create invoice", "Generate invoice", "Invoice order" and "Create invoice document" mean the same work.

And as for the first word: in these five runs the AI used **65 different verbs**. Capture or receive, check or validate, create or issue, for the same tasks.

Eight steps until the glossary works

It took eight steps until the glossary had any effect. The first five built it. The numbers hardly moved.

Step	What was missing	What it took
1 · Word list	A list says which words exist, but not what they mean, nor which ones are wrong.	A real glossary with meaning, variants and languages
2 · Old glossary	The existing glossary came from a project for introducing BPM: lots of good words, just not those of order handling.	The terms of the business domain
3 · AI suggestions	Nobody writes a glossary by hand from over a hundred documents in four languages. The AI found 2072 suggestions, but one per word, not per term.	Merging
4 · Curation	Five rows for the credit note must become one term, with one name per language and the other words as synonyms.	Assigning each suggestion to a term, focused on the project
5 · Confirm	All terms were drafts, and drafts are not passed to the AI. One test series was therefore invalid.	Domain experts who confirm and reject
6 · Arrive	Of 473 terms, exactly three reached the AI.	Synonyms that bring a term into the prompt
7 · Verbs	The glossary governs nouns. The drift is in the verbs.	A verb rule
8 · Core objects	The three things the process revolves around were missing from the glossary.	Added by hand

Three steps deserve a closer look.

Step 6: The glossary does not arrive

Series B ran with 484 confirmed terms. Result: 181 instead of 190 different names. Practically nothing.

So I looked at what really reaches the AI. My lab tool does not give it the whole glossary, otherwise the request gets too long, only the terms the process mentions. And "mentions" meant: the word appears literally in the text of the process. If the process says "credit memo" but the term is called "credit note", it does not come along. Synonyms did not count. Of 473 terms, exactly three arrived: "invoice", "online shop" and "INVOIC", the name of an EDI message.

Since the fix, synonyms count too, and 31 terms arrive that fit this process. And one more thing: this must already happen in the analysis. That is where the AI first chooses what the steps are called. The design only takes the words over.

The lesson is uncomfortable because it is not visible in the user interface. A glossary can be complete, curated and confirmed and still have no effect if it does not find its way into the request.

Step 7: The verbs

Even a glossary that arrives has a gap. It governs nouns: what an order is, what a credit note is. But an activity name consists of a verb and an object, as modelling guidelines have long required (Silver, 2011; Mendling, Reijers & Recker, 2010). And the verb is where the AI drifts most.

So the AI also gets a short verb list: one binding verb per kind of work, "Check" instead of "Validate" or "Verify", "Create" instead of "Generate" or "Issue". And the tool corrects an activity's verb if it is outside the list. The idea is not new; research on the automatic refactoring of activity labels works with exactly this split into verb and object (Leopold, Smirnov & Mendling, 2012).

Series B2: only **13 different verbs** instead of 65, and the names drop from 190 to **135**. The first really big step.

With a side effect: suddenly a lot was simply called "Create". "Create shipment" instead of "Ship order". A business activity becomes the mere creation of a record. So the list now explicitly allows business verbs that make sense in this process: ship, produce, pick, pack.

Step 8: The core objects

With the verbs under control, the drift moved on, to the objects. Is it "ERP order" or "sales order"? "Production" or "production order"? "Shipment" or "order"?

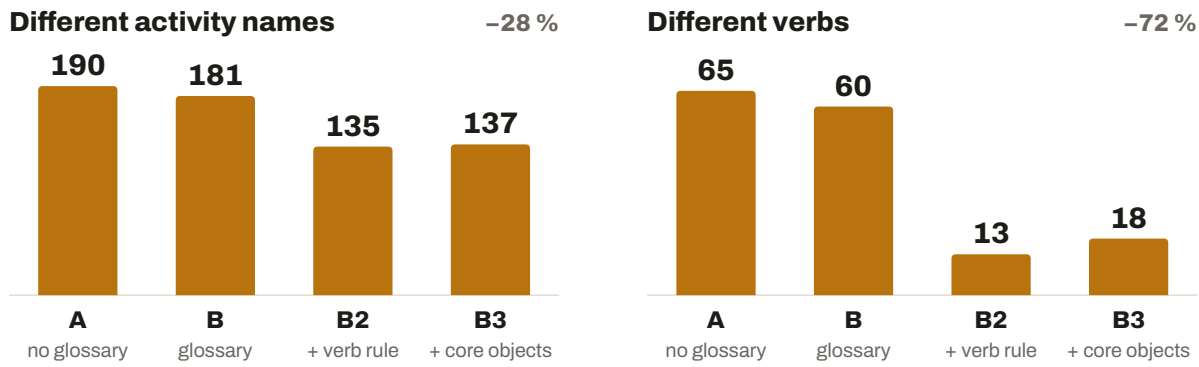
This is exactly where the glossary should have helped. I checked: these terms were missing. The curated glossary had 484 terms, but not the three things this process revolves around. So I added them: "sales order", "production order" and "parcel", and defined "ERP order" as a rejected term.

Series B3: "ERP order" no longer appears in any run, "shipment" only twice. The number of names stays at 137, but now they are the right words.

The lesson: the core objects of a process must be in the glossary. An AI that finds thousands of terms does not necessarily find the three most important ones.

What changes

Four test series of five runs each, the same process, the same model, the same documents:



Different activity names and different verbs over five runs each. The biggest jump comes with series B2, the verb rule.

The numbers in detail:

Series	Glossary	different names	different verbs	events the same in all 5 runs	flows the same in all 5 runs
A	none	190	65	6	1
B	484 confirmed terms	181	60	6	3
B2	plus synonyms and verb rule	135	13	7	4
B3	plus the core objects	137	18	10	8

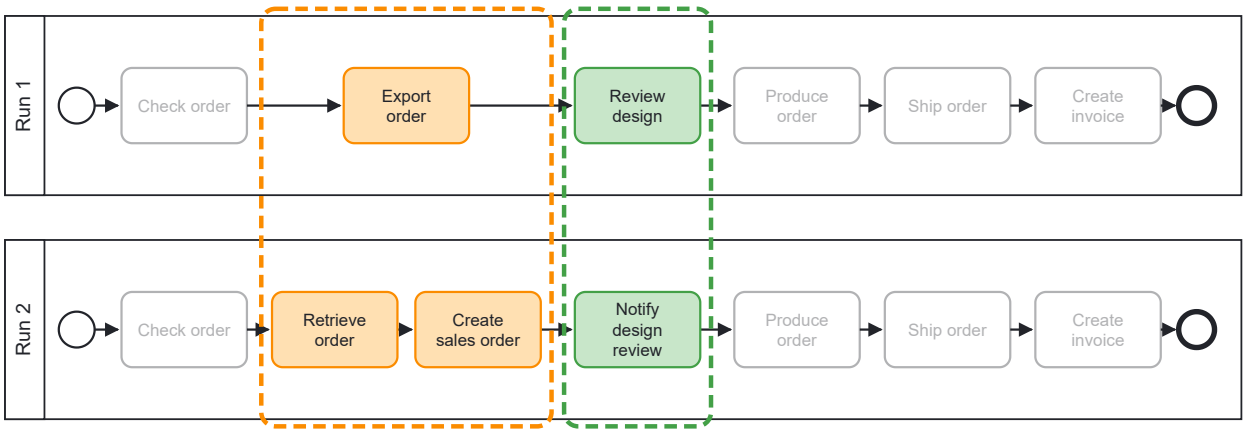
The names drop by a good quarter, the verbs from 65 to 18. So the biggest effect did not come from the glossary, but from the verb rule next to it. And what stays the same in all five runs increases: ten events instead of six, eight flows between the steps instead of one.

On top of that, every correction is traceable: which word was replaced, by what, and on the basis of which glossary entry.

So is it stable now? No, not completely. But much calmer, and the words that remain are those of the company.

What the glossary does not solve

First, how the steps are cut. On the top level, without a glossary, there was actually the least movement: nine different names, five of them the same in all runs. With the glossary there were sixteen. What changes there is no longer the word, but the cut.



Two of the five runs of series B3, top level. Faded: the same in both. Orange: the hand-over to the ERP is one step in one run and two in the other. Green: in one run the process reviews the design itself, in the other it only triggers the review.

Is the hand-over to the ERP a step of its own, or two? Does the process review the design itself, or only trigger the review? That is not a question of words. A confirmed analysis has to fix it, so the AI does not reinvent it in every run. The classic guidelines of business process modelling already separate the consistency of terms from the question of what a model represents and at what level of detail (Becker, Rosemann & von Uthmann, 2000).

Second, the work of the domain experts. The AI finds the terms and suggests groups, but it makes mistakes, and not the same ones every time. And, as we saw, it misses precisely the most important ones. A human has to confirm and complete.

Try it yourself

The material from the experiment can be downloaded from bpmn.ch (bpmn.ch/v/2): the 28 documents of 4Health, the glossary with 487 confirmed terms including synonyms, rejected words and sources, the verb list, a guide and a score sheet. The documents are in four languages; the guide is in German.

With it, the core of the experiment can be repeated with any tool: five times without a glossary, five times with the glossary and the verb list, and then count how many different names and verbs come out. The numbers from a chat are not directly comparable with those from the video, because there a whole model with sub-processes was built. The difference between the two series is.

The glossary is included exactly as it was used in the experiment, not polished afterwards. Whoever goes through it will also find the AI's mistakes, and with them the best evidence of why, in the end, a domain expert has to confirm.

MESTRO, my lab tool, implements steps 3 to 8: AI suggestions with sources, the curation proposal, confirmation, synonyms in the prompt from the analysis onwards, the verb rule and the traceable correction. It serves here as an example of where the work lies, not as a product comparison.

Conclusion

Does a glossary help against drift? On its own: almost nothing. A full, confirmed glossary hardly moved the number of names in my experiment, because it did not reach the AI.

It works when several things come together. The terms reach the AI along with their synonyms. There is a rule for the verbs, not only for the nouns. And the core objects of the process are in it. Then the names drop by a good quarter, and the verbs from 65 to 18.

So a glossary is necessary, but not sufficient. What it does not govern is how the steps are cut. For that you need an analysis that someone has confirmed.

The good news: the AI does most of the work. It reads the documents, finds the terms and suggests groups. But the decision is up to a human.

And a question for you: how many words does your company have for "order"?

Note on the experiment: process "B2C order-to-cash" of the fictional company 4Health, model gpt-oss, output in English, five runs per series, November 2026. In each run the AI rebuilds the analysis, then the model with all sub-processes; evaluated automatically over all activities, events and flows. Series B2 and B3 ran with a newer version of the tool (synonyms in the prompt, verb rule). All documents are fictional. MESTRO is my lab tool and is being developed at ELCA, where I work; it serves here as a reference implementation, not as a product comparison.

Sources & further reading

- Video for episode 2 "Does a Glossary Stop the AI Drift?" on the channel "Stephan Fischli · Agentic Orchestration": English version, German original
- Documents and glossary from the experiment on bpmn.ch: bpmn.ch/v/2
- Episode 1 "Text in, model out?", and the article "Text rein, Modell raus? Fünf Stufen, an denen die KI scheitert" (in German)
- J. Becker, M. Rosemann, C. von Uthmann: *Guidelines of Business Process Modeling*. In: W. van der Aalst, J. Desel, A. Oberweis (eds.): Business Process Management, LNCS 1806, Springer, 2000, pp. 30–49.
- ISO 704:2022: *Terminology work — Principles and methods*. International Organization for Standardization, 2022.
- H. Leopold, S. Smirnov, J. Mendling: *On the refactoring of activity labels in business process models*. Information Systems 37(5), 2012, pp. 443–459.
- J. Mendling, H. A. Reijers, J. Recker: *Activity labeling in process modeling: Empirical insights and recommendations*. Information Systems 35(4), 2010, pp. 467–482.
- S. Ouyang, J. M. Zhang, M. Harman, M. Wang: *An Empirical Study of the Non-determinism of ChatGPT in Code Generation*. ACM Transactions on Software Engineering and Methodology

34(2), 2025.

- B. Silver: *BPMN Method and Style*, 2nd edition, Cody-Cassidy Press, 2011. https://methodandstyle.com/