VOKABELN: AGENTIC ENGINEERING

THOMAS SIEVERING

die Teile benennen

WARUM BEGRIFFE HELFEN

- Zwei Menschen sagen "KI" und meinen oft unterschiedliche Dinge.
- Gute Fachsprache macht bessere Prompts und bessere Reviews.

DIVE DEEPER

ZUM NACHSCHLAGEN

Beispiel für Animationen

---

animations.dev/vocabulary

https://animations.dev/vocabulary
Begriffe, Animationen und zusätzliche Erklärungen.

PROJECT VOCABS

WORKSPACE

- **.pi**
  - **skills**
    auth.json
- **assets**
- **slides**

AGENTS.md

VOCAB.md

DIESE BEGRIFFE KLÄREN WIR

token · llm · model · harness · context · context window · skill · tool · agent

TOKEN

- Modelle lesen keinen Text wie wir, sondern Textstücke.
- Ein Token ist oft nur ein Bruchstück von einem Wort.
- Kosten, Geschwindigkeit und Kontextfenster hängen von Tokens ab.
- Mehr Tokens heißt nicht automatisch mehr Verständnis.

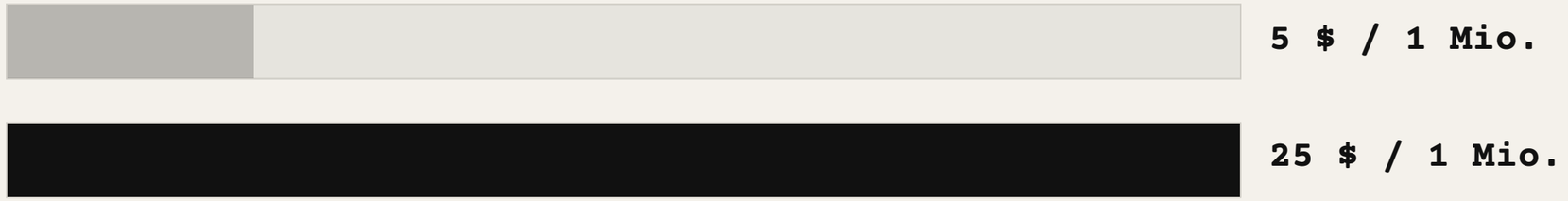
What's a token, actually?

Not a word. Not a character. Something in between.

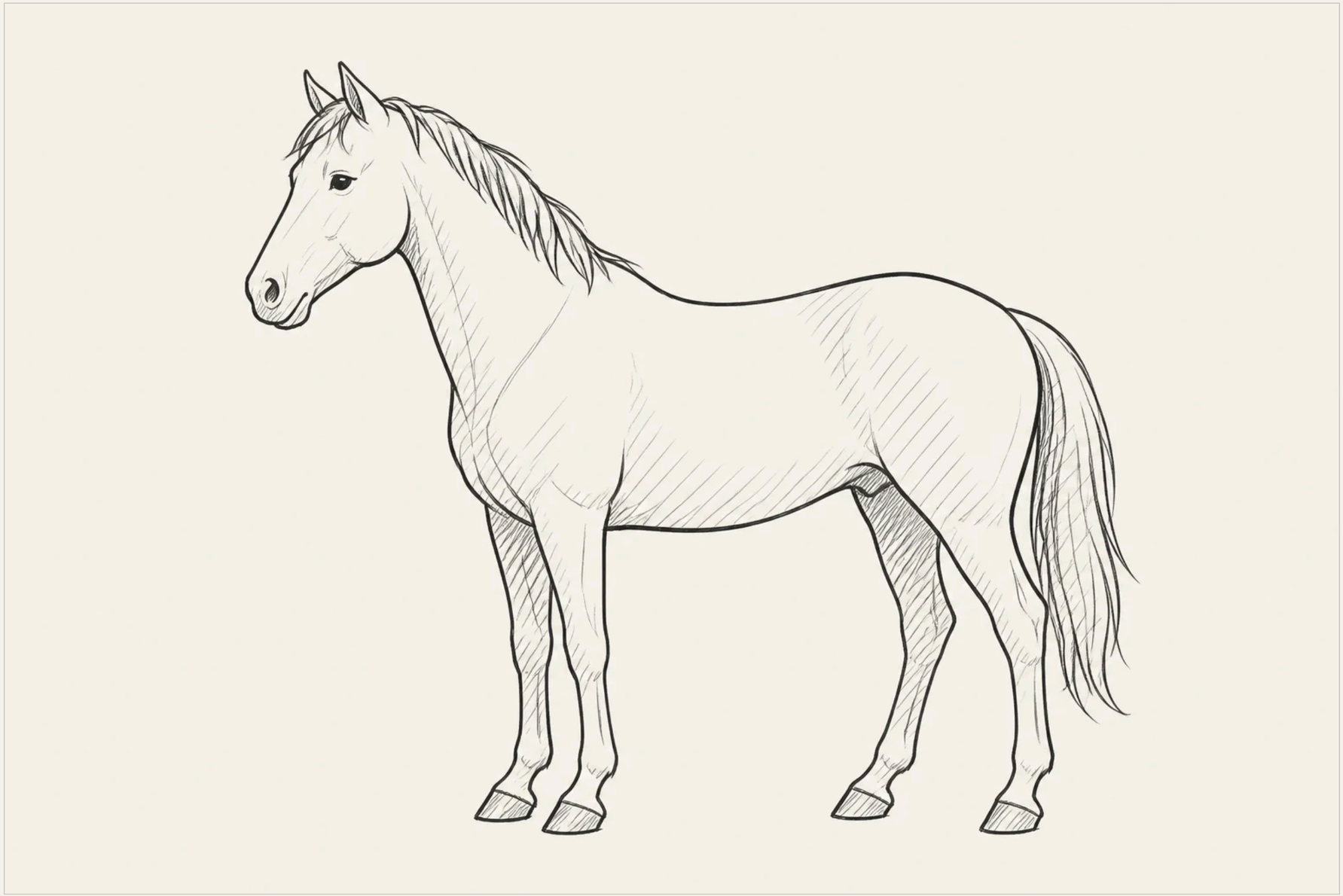
4 words ⟶ 8 tokens

OPUS 4.8

Token-Kosten



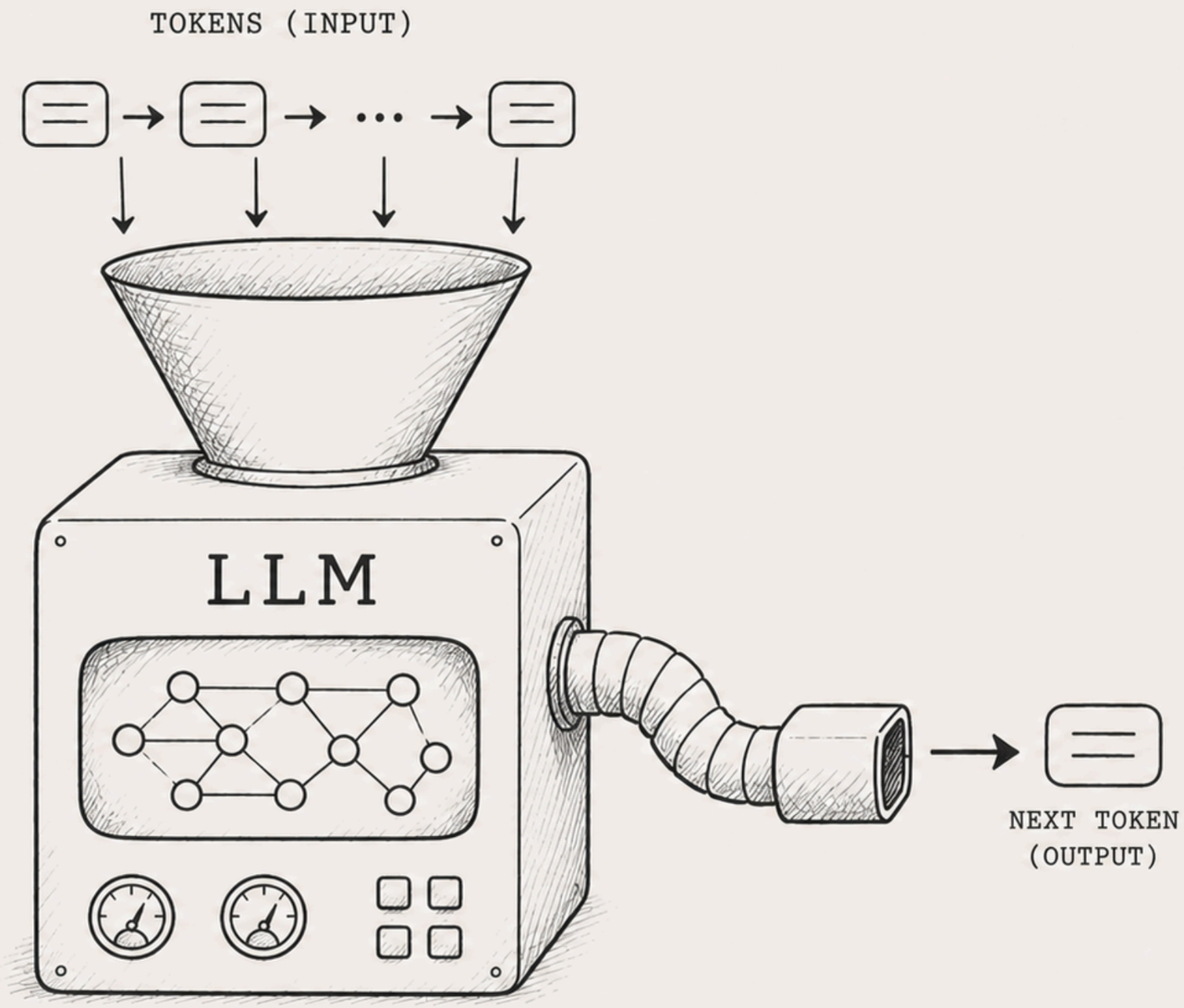
Output ist deutlich teurer als Input — lange Antworten kosten.



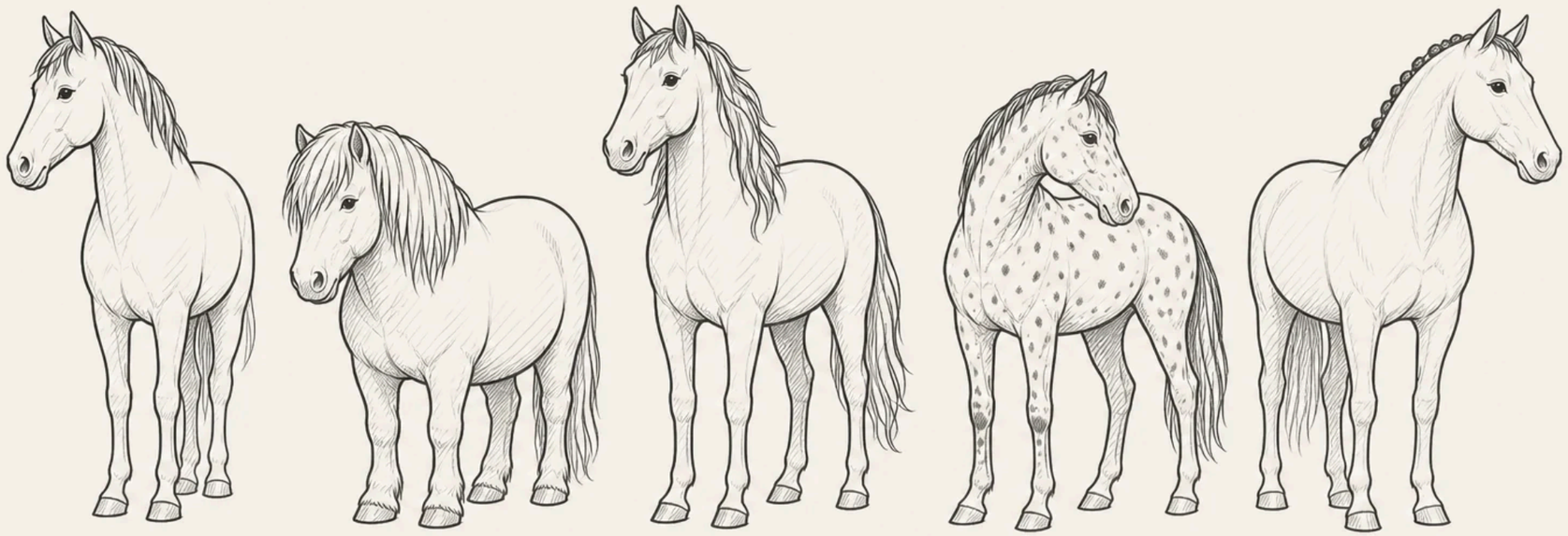
MODEL: ROHE FÄHIGKEIT, NOCH OHNE AUFGABE, GRENZEN ODER RICHTUNG.

LLM UND MODEL

- LLM: Large Language Model — ein Sprachmodell für Text, Code und Struktur.
- Es verarbeitet Eingaben und erzeugt passende Ausgaben in Tokens.
- Model: die konkrete Version eines LLM, z. B. Claude Opus 4.8 oder GPT-5.5
- Ein Model hat ein learning cutoff date



LLMS ARBEITEN TOKEN FÜR TOKEN: KONTEXT HINEIN, NÄCHSTES TOKEN HERAUS.

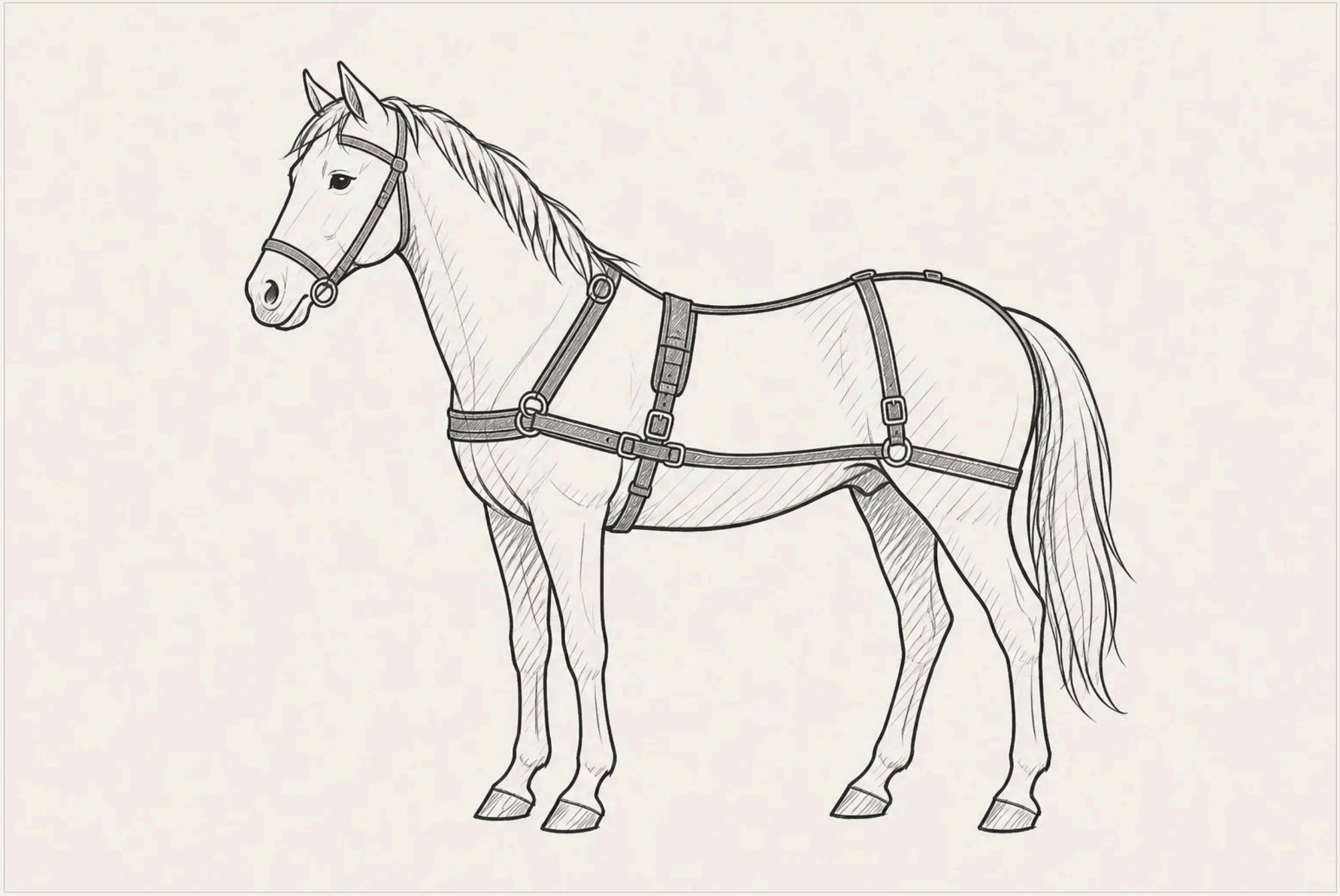


DIFFERENT MODELS: SCHNELL, STARK, AUSDAUERND, RUHIG — PASSEND ZUR AUFGABE STATT ABSOLUT BESTES MODELL.

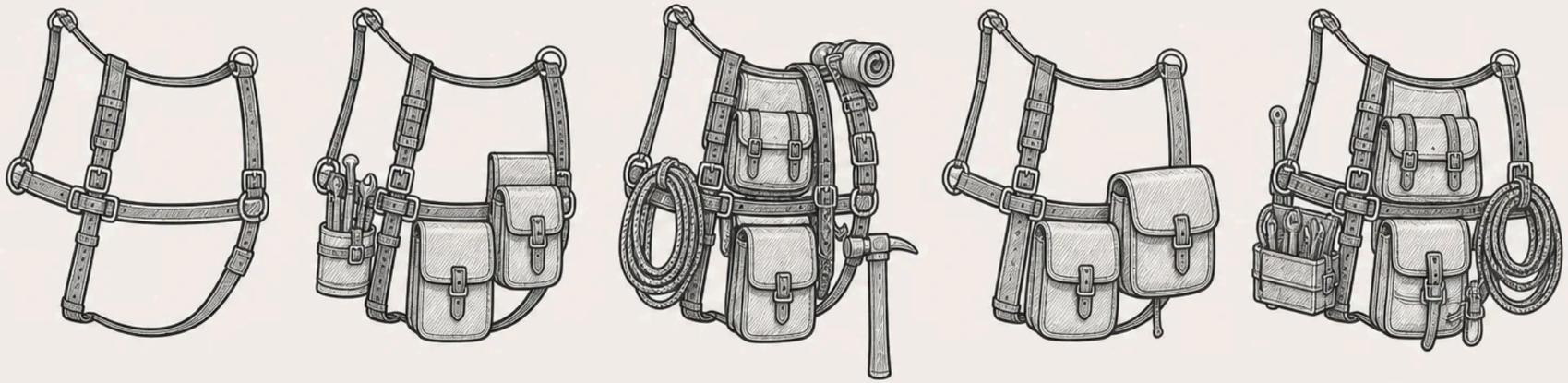
BEISPIELE

VERSCHIEDENE MODELLE

GPT-5.5 | Claude Haiku | Claude Sonnet | Claude Opus | Gemini Flash | Gemini Pro



HARNESS: TOOLS, KONTEXT, REGELN, PERMISSIONS UND FEEDBACK MACHEN MODELL-POWER STEUERBAR.

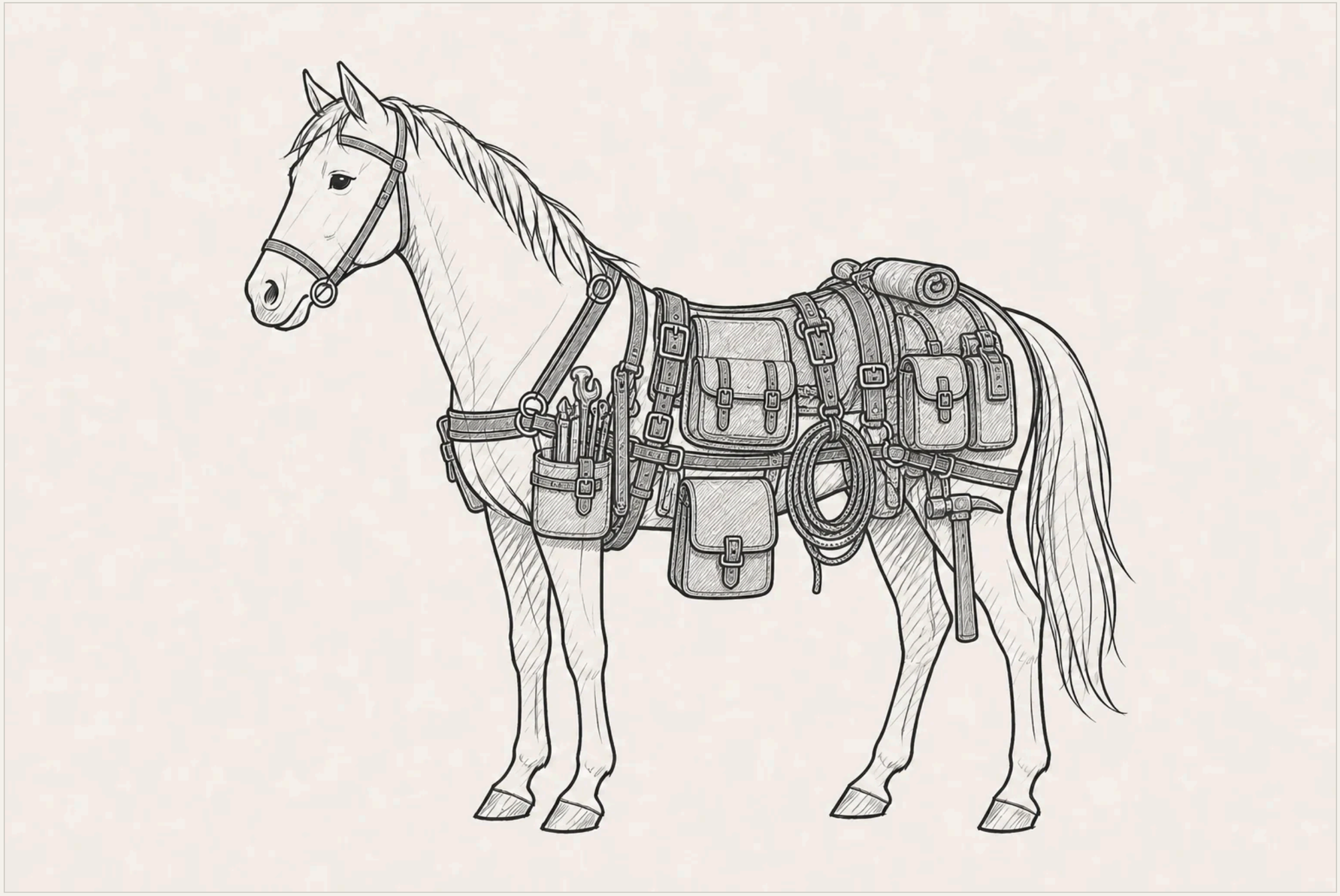


DIFFERENT HARNESSES: MINIMAL UND KONTROLLIERBAR ODER MÄCHTIG UND KOMPLEX.

BEISPIELE

VERSCHIEDENE HARNESSES

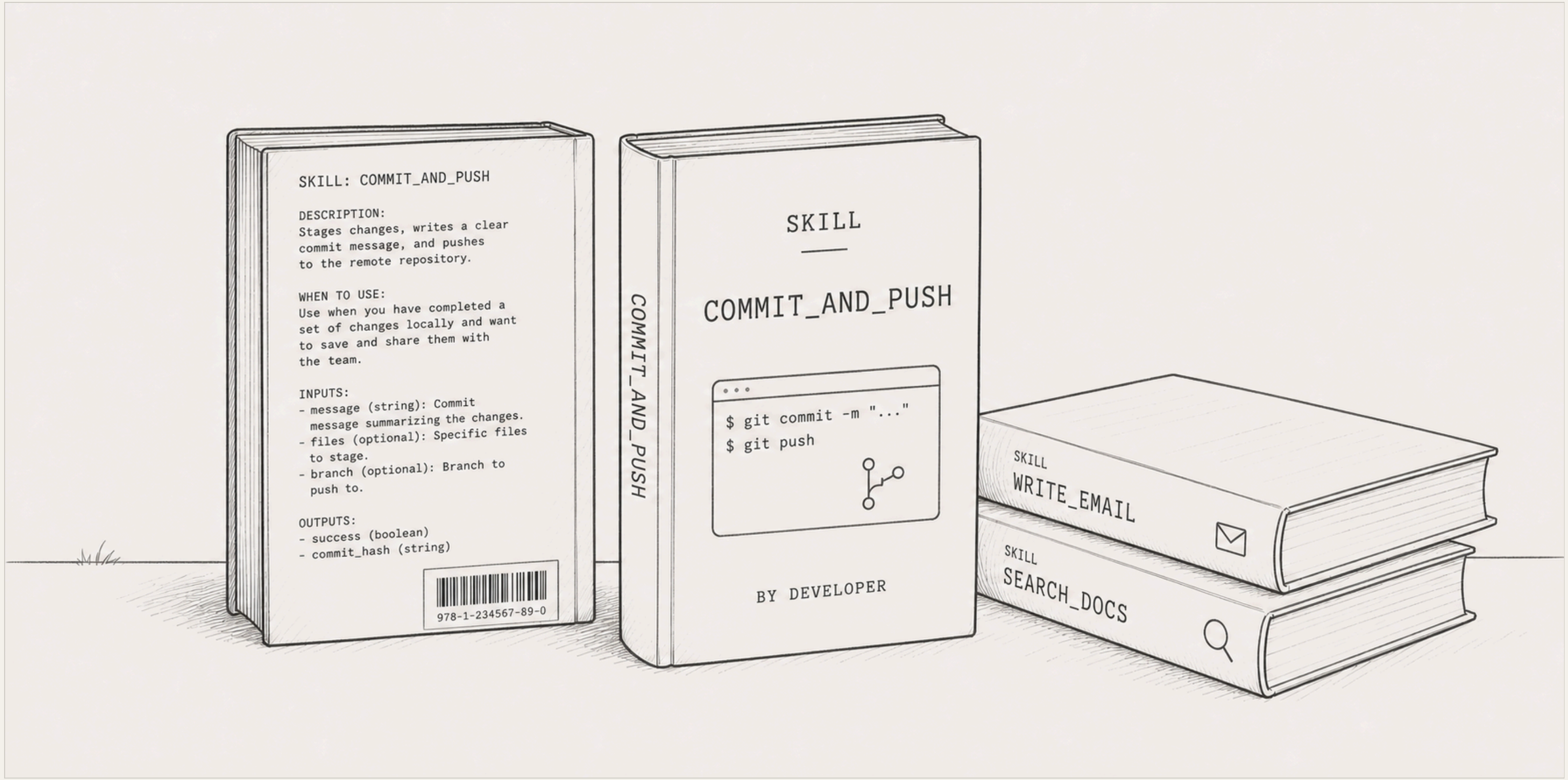
Claude Desktop
Claude Code
Codex
ChatGPT Desktop
ChatGPT iOS App
Pi
OpenClaw
OpenCode
Cursor
Copilot



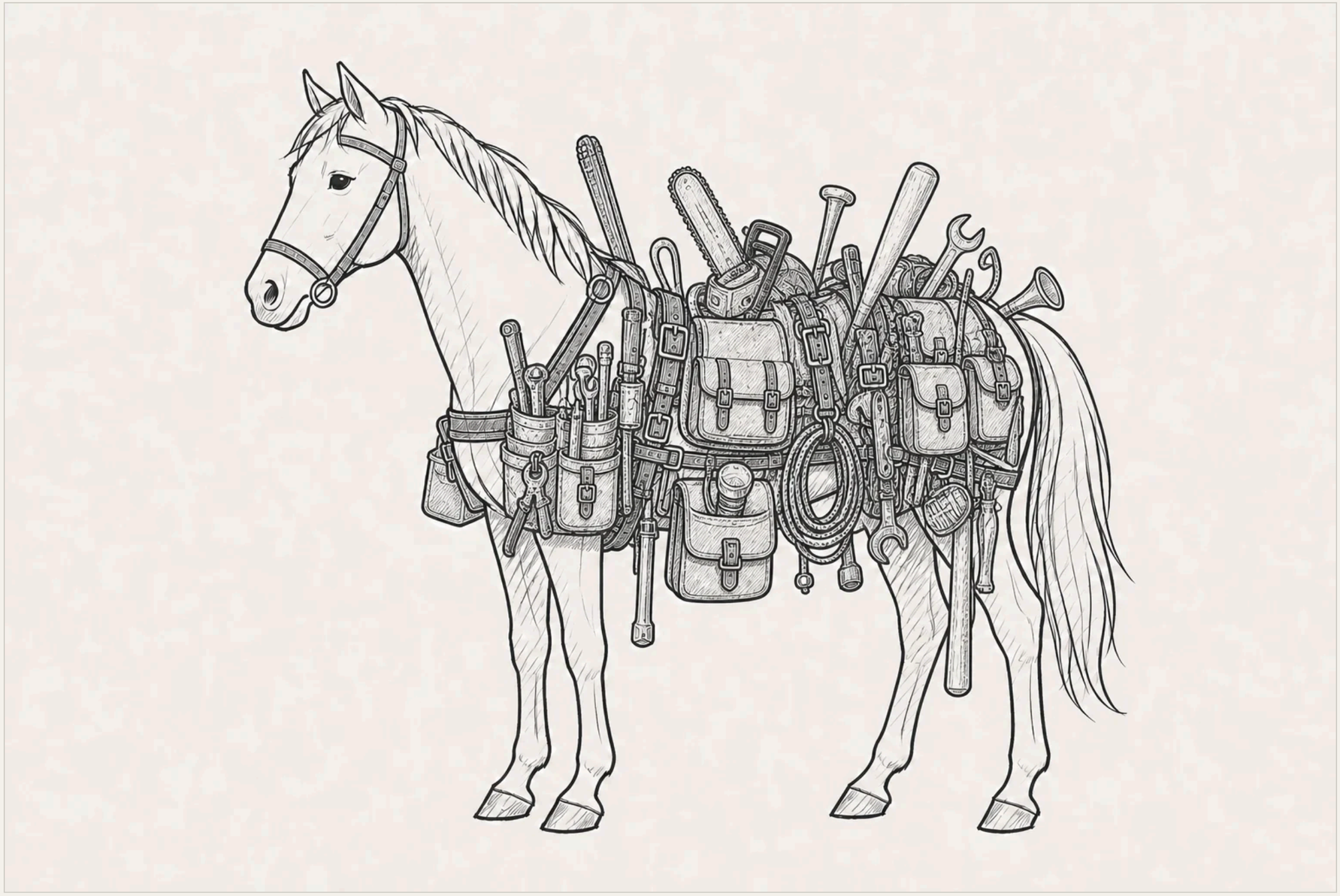
HARNESS TOOLS: JEDES TOOL ERWEITERT HANDLUNGSMÖGLICHKEITEN UND ENTSCHEIDUNGSFLÄCHE.

TOOL

- Ein Tool ist eine klar definierte Fähigkeit außerhalb des Modells.
- Beispiele: Datei lesen, Web suchen, Code ausführen, E-Mail entwerfen.
- Das Modell entscheidet, welches Tool sinnvoll ist; die Harness führt es aus.



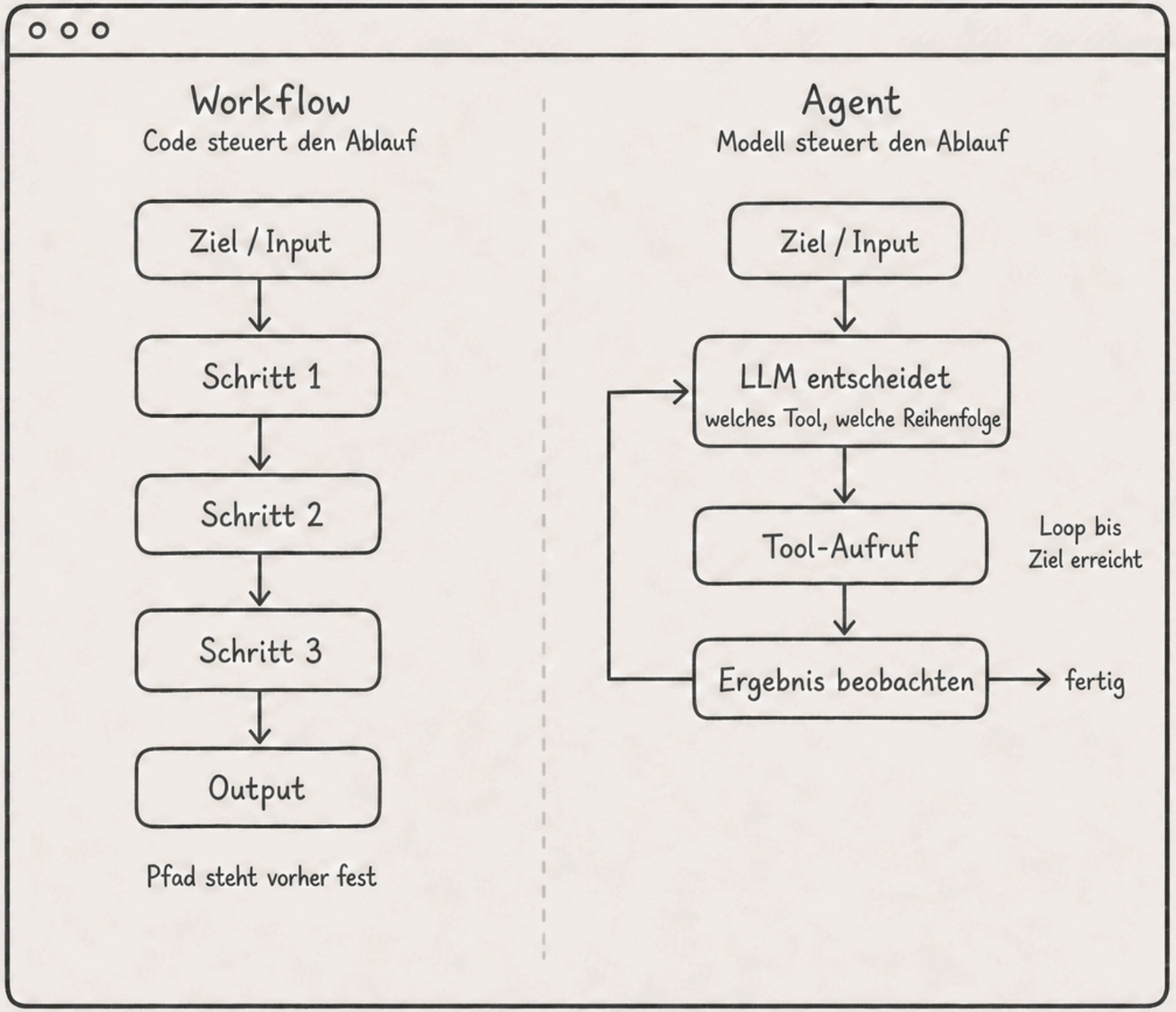
SKILL: EINE WIEDERVERWENDBARE, BESCHRIEBENE FÄHIGKEIT, DIE DER AGENT BEI BEDARF LÄDT.

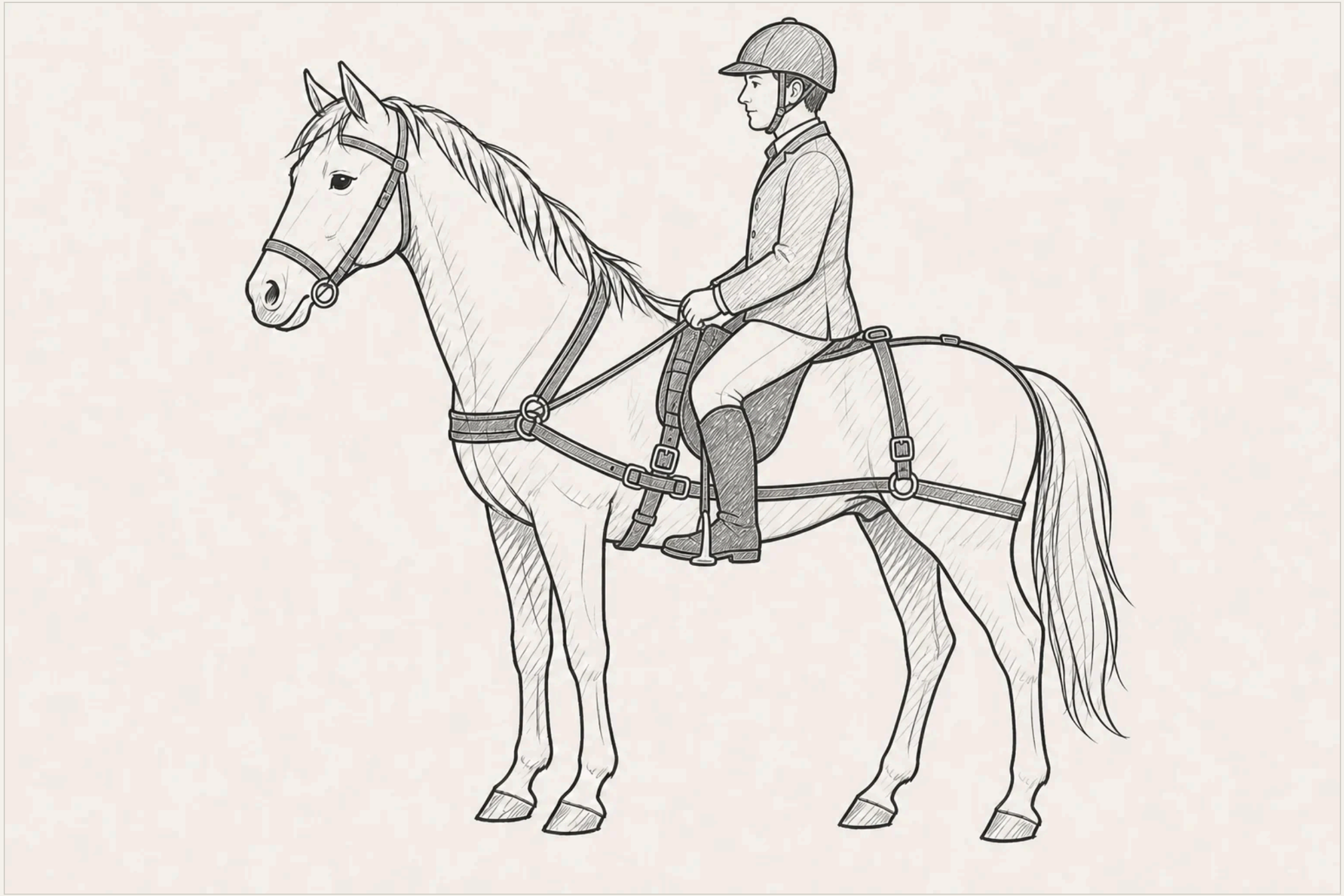


TOO MANY TOOLS: TOOLING-FLÄCHE KANN VERHALTEN VERSCHIEBEN UND KONTROLLE ERSCHWEREN.

-> OPEN CLAW

CHATBOT VS. AGENT

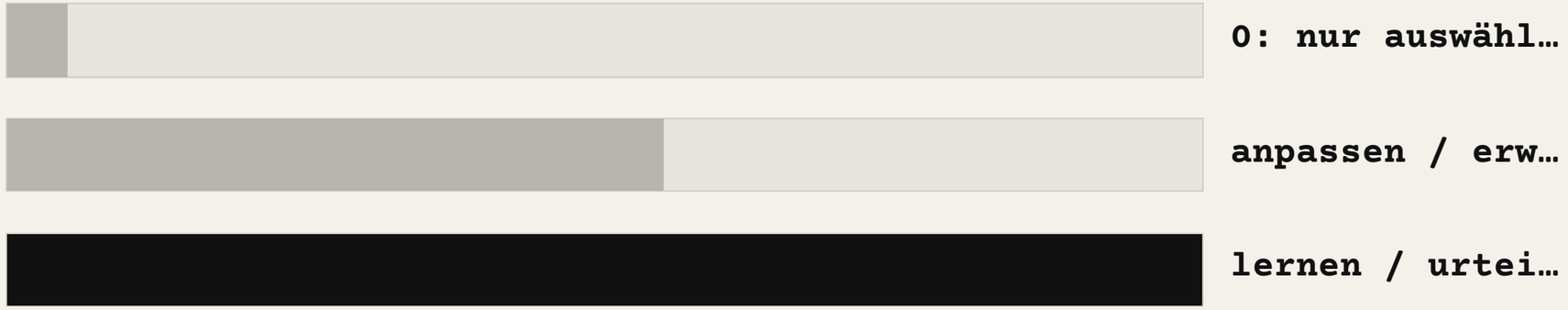

Workflow
Code steuert den Ablauf
Ziel / Input
Schritt 1
Schritt 2
Schritt 3
Output
Pfad steht vorher fest
Agent
Modell steuert den Ablauf
Ziel / Input
LLM entscheidet
welches Tool, welche Reihenfolge
Tool-Aufruf
Loop bis
Ziel erreicht
Ergebnis beobachten
fertig



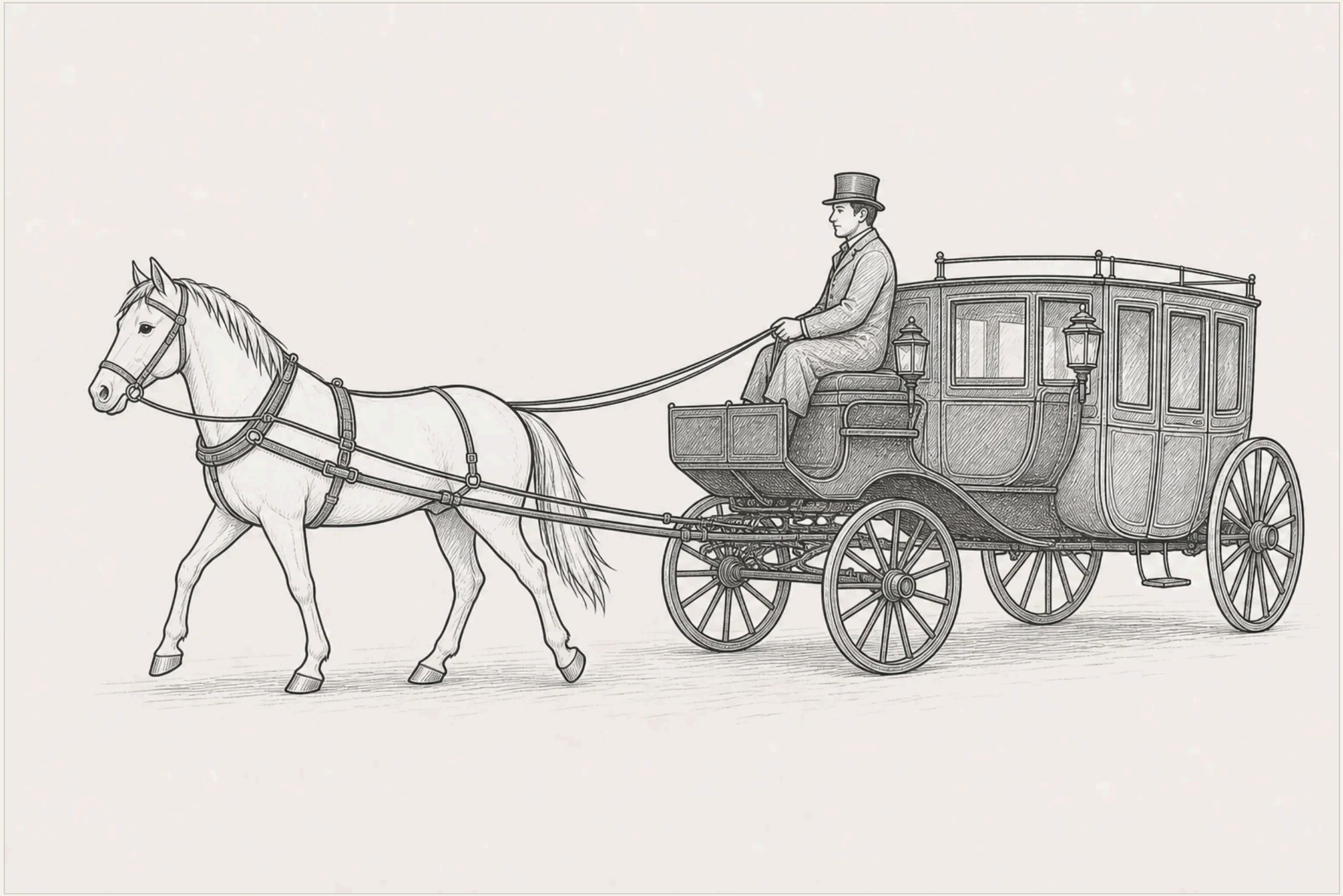
RIDER/USER: DER MENSCH GIBT RICHTUNG, SETZT GRENZEN UND BEHÄLT KONTROLLE.

MODEL · HARNESS · USER

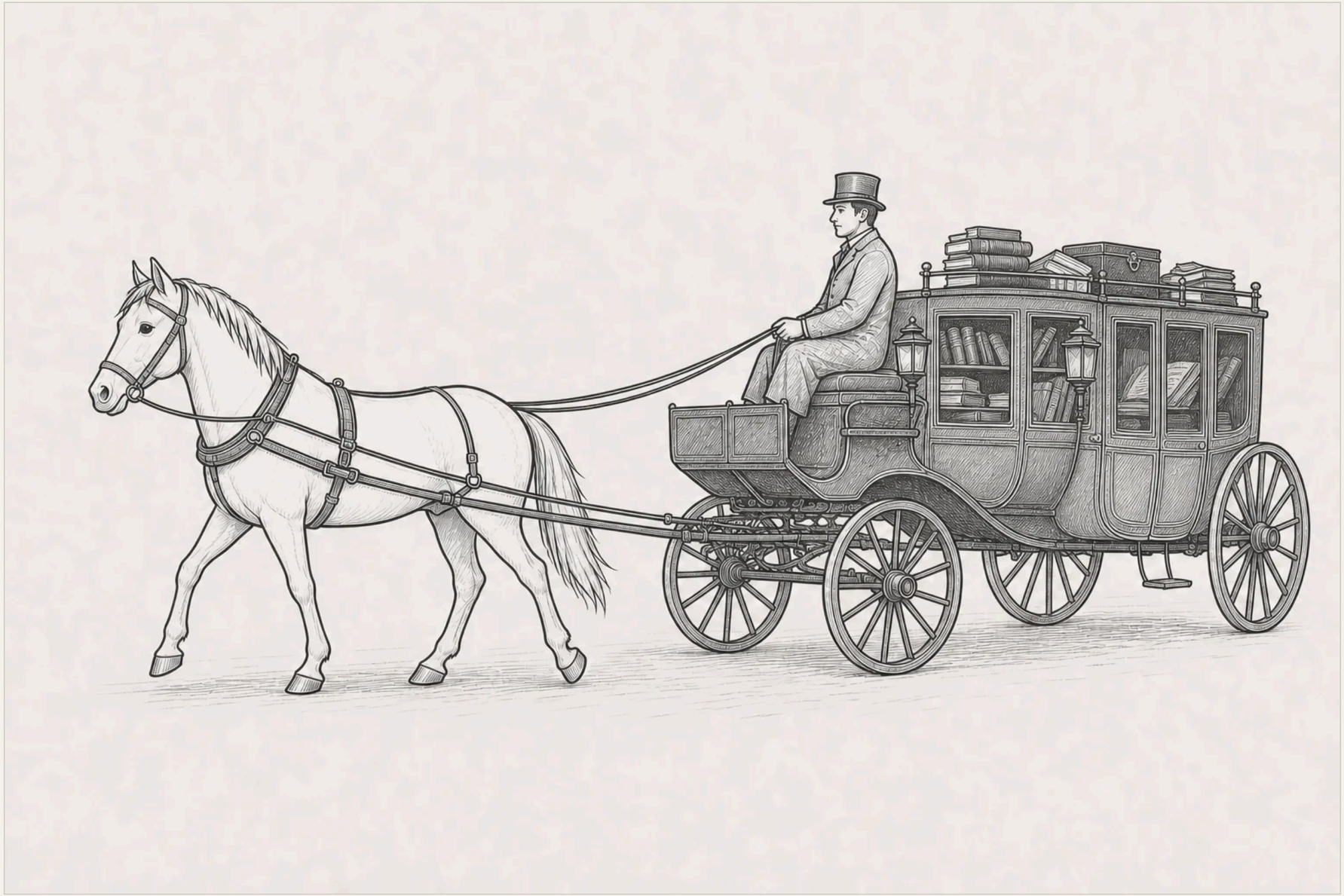
Wo haben wir Einfluss?



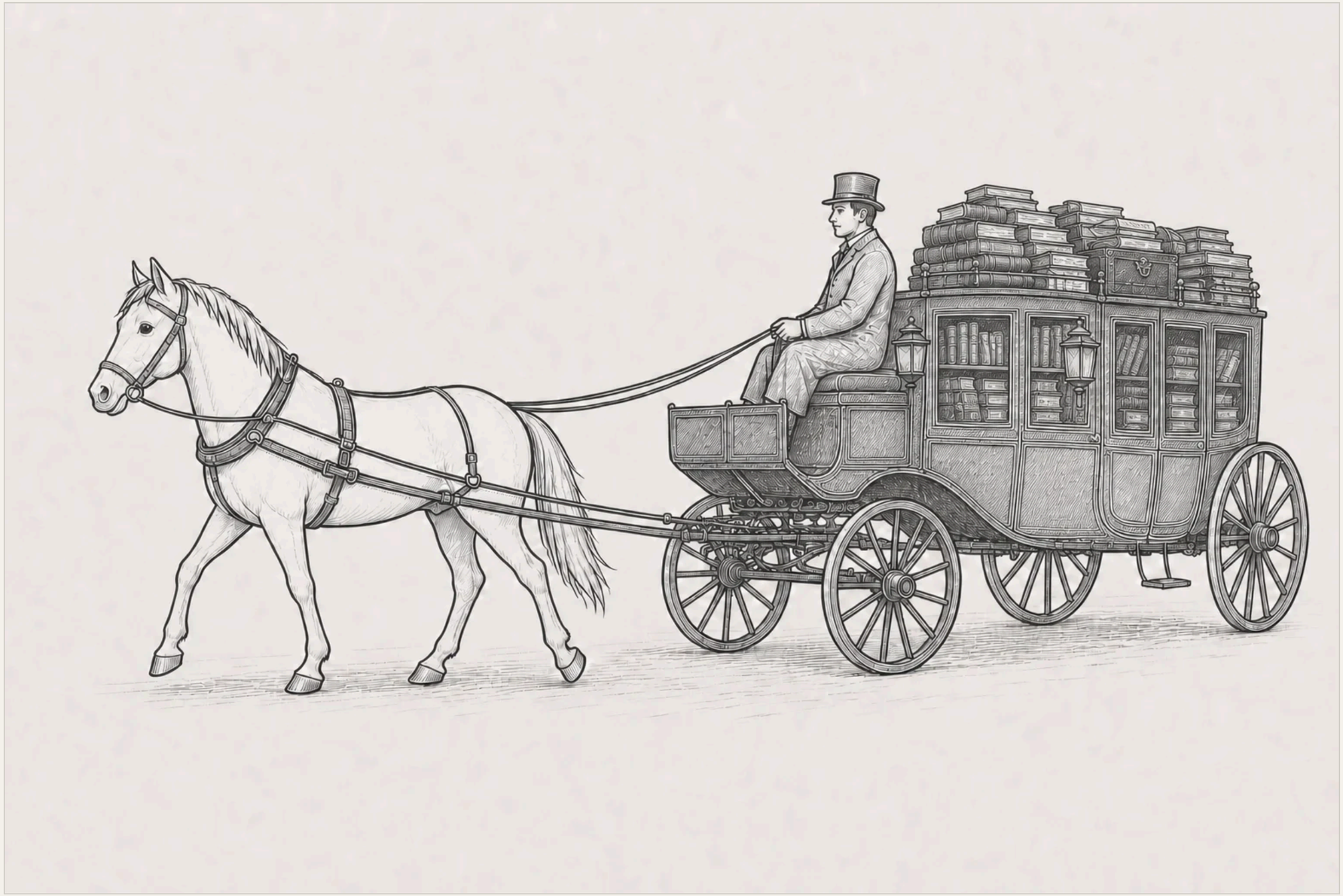
Modell: Wir wählen aus. Harness: Wir konfigurieren Tools, Rechte und Kontext — und können manche Harnesses erweitern. User: Wir verbessern Fragen, Checks, Mental Models und Qualitätsurteil.



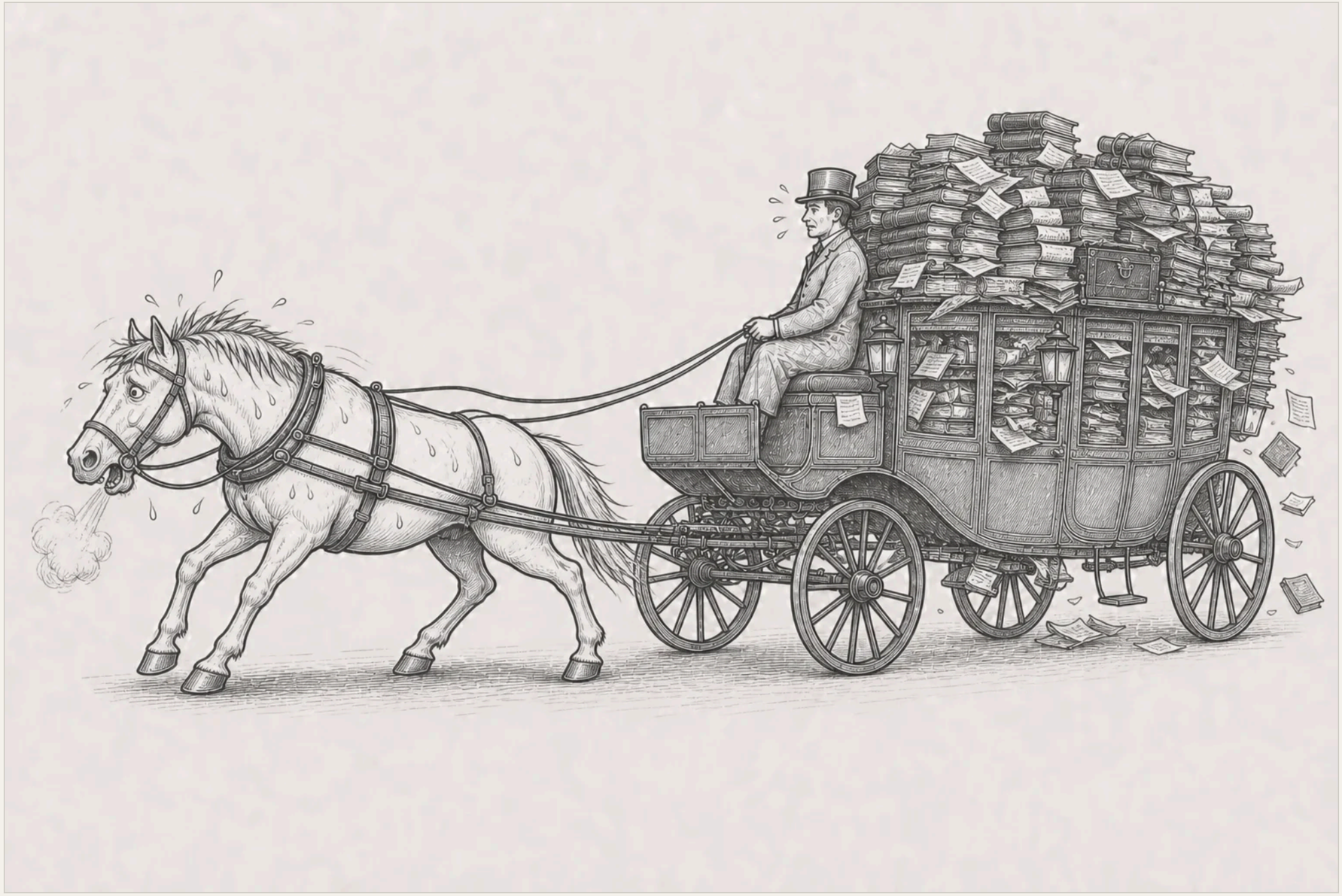
CONTEXT WINDOW: DIE KUTSCHE ENTHÄLT NUR DAS, WAS DAS MODELL GERADE MIT SICH TRÄGT.



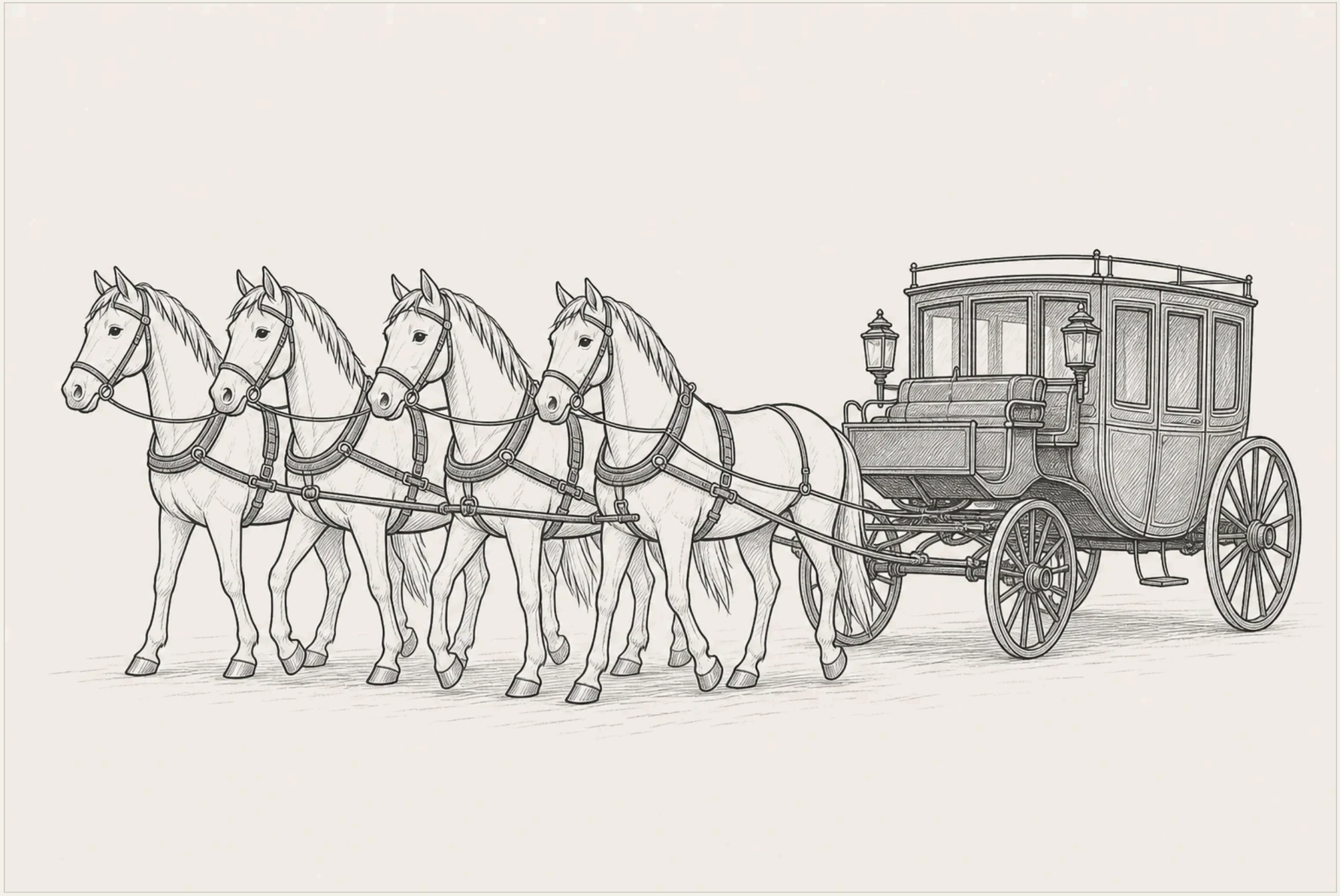
LOW CONTEXT LOAD: GENUG ORIENTIERUNG, NOCH PLATZ ZUM DENKEN UND PRIORISIEREN.



MEDIUM CONTEXT LOAD: MEHR DETAILS HELFEN, ABER AUSWAHL UND REIHENFOLGE WERDEN WICHTIGER.



DUMB ZONE: MEHR KONTEXT MACHT DAS SYSTEM SCHWERFÄLLIGER STATT SCHLAUER.



ORCHESTRATION: MEHRERE MODELLE ODER AGENTEN ARBEITEN KOORDINIERT — OFT MIT GETRENNTEN KONTEXTFENSTERN.

FRAGEN ANYONE?