

Projekt Wilhelm

Ein prüfbarer Compliance-Wissensspeicher, gebaut vom biteme AI OS

88 Tage von der Materialsammlung zum auditierbaren Wissensspeicher, in dem jede Aussage einmal gegen die Primärquelle geprüft wurde. Was Sie hier sehen, ist die sichtbare Spitze eines mehrschichtigen KI-Systems, in dem elf Modelle in fest definierten Rollen zusammenarbeiten.

73EVIDENCE CARDS
69 VERIFIZIERT**88** Tage20 VERSIONEN
25.05.–20.08.**167**PRIMÄRQUELLEN
AM TEXT GEPRÜFT**12**OPERATIVE
WERKZEUGE**~2.200**MODELL-LÄUFE
GESCHÄTZT

biteme.digital, Wien · Fact Sheet, Stand 25.08.2026 · Gezählte Werte stehen für sich; alle mit ~ markierten Zahlen sind ausgewiesene Schätzungen
· erstellt vom biteme AI OS

★ Drei Dinge, die den Unterschied machen

Fünf Dokumente Prüfarbeit je Karte

Rund **414 Arbeitsdokumente** stehen **73 ausgelieferten Karten** gegenüber. Sieben der zwanzig Versionen waren reine Korrektur-Releases: mehr Zeit im Nachprüfen als im Hinzufügen. Genau daher kommen die **69 verifizierten** Karten.

Das richtige Modell für jede Aufgabe

Beschaffung und deterministische Prüfung laufen über **Abo-Modelle ohne Kosten pro Lauf**, Reichweite über spezialisierte Modelle, das teure Urteil nur dort, wo es zählt. Der protokollierte Abschlusstag kostete **2,75 USD** extern.

Was Sie sehen, ist die Spitze

Die 73 Karten sind das sichtbare Ergebnis. Darunter arbeiten eine **orchestrierte Modell-Maschine** und rund **2.000 direkte Architektur-Läufe** im Hauptsystem. Der reale Intelligenz-Einsatz ist um ein Vielfaches größer als das Endprodukt.

1 Was Wilhelm heute ist

Der geprüfte Bestand, Stand 20.08.2026

73

EVIDENCE CARDS
AKTIV

69

DAVON VERIFIZIERT

26

KARTEN ARCHIVIERT

12

OPERATIVE
WERKZEUGE

14

THEMEN-PAKETE

167

PRIMÄRQUELLEN-PDFS

786

MARKDOWN-DATEIEN

1.465

DATEIEN GESAMT

398 MB

BAUMGRÖSSE

100 %

KARTEN MIT
HERKUNFTS-
FRONTMATTER

Die 73 Karten sind der Kern. Jede trägt Frontmatter mit Quelle, Prüfdatum, Prüfer, Vertrauensgrad und Ablaufdatum: das ist der Unterschied zwischen einem Wissensspeicher und einer Sammlung von Notizen. Die **26 archivierten** Karten sind kein Ballast, sondern der Beleg, dass das System altert und aufräumt. Eine abgelöste Karte wird nicht gelöscht, sondern archiviert, und die Nachfolgerin benennt ihre Vorgängerin.

2 Der Weg: 20 Versionen in 88 Tagen

v1.0 (25.05.) bis v1.12.1 (20.08.)

25.-28.05.

Aufbau

Neun Versionen in vier Tagen. Reorganisation, Vollaussattung, acht Werkzeuge, drei Audit-Wellen, Eval-Suite. Ergebnis: Mietkunden-Tauglichkeit.

14./15.06.

DACH

Deutschland und Schweiz dazu, nach zwei unabhängigen Gegenprüfungen durch eine getrennte Instanz.

24./25.07.

Q3-Abschluss

Medien-Vertikale, Kernfakten-Refresh, Einführung des Primärquellen-Gates, Integritäts-Reparatur.

19.08.

Refresh-Welle

Fristen-Welle plus modellübergreifender Gegencheck, dann zwei Nachprüfungs-Releases am selben Tag.

20.08.

Vertragswissen

Neun neue Karten (AVV, NDA, AGB, Haftung), zwei Werkzeug-Gerüste, Gegencheck, eine Bestandskarte sauber abgelöst.

Bemerkenswert am Verlauf: Von den 20 Versionen sind **sieben reine Korrektur- oder Verifikations-Releases**. Das Projekt hat mehr Zeit mit Nachprüfen verbracht als mit Hinzufügen. Genau das erzeugt die Qualität von 69 verifizierten Karten.

3 Die Arbeit dahinter: rund 414 Arbeitsdokumente

Briefings, Befunde, Synthesen, Prüfprotokolle, Belegmappen

RECHERCHE- UND PRÜF-WELLE	ARBEITSDOK. (.MD)	BELEGDATEIEN	ZEITRAUM
Aufbau-Recherchen	60	244	25.05.–14.06.
Audit-Welle 3	14	0	27.05.
Eval-Suite	5	0	27./28.05.
Deep Audit Q3	67	3	24./25.07.
Abschluss-Welle Q3	69	12	25.07.–20.08.
Refresh-Welle 08/2026	147	37	19.08.
Vertragswissen-Welle	47	149	20.08.
weitere (Coverage, Primärquellen, Strategie, Spec)	5	0	27.05.
Summe Arbeitsdokumente	414		

Das Verhältnis ist die eigentliche Aussage. Rund 414 Arbeitsdokumente auf 73 Karten: auf jede ausgelieferte Karte kommen etwa fünf Dokumente Prüf- und Belegarbeit.

Die **149 Belegdateien** der Vertragswissen-Welle sind ein eingefrorener Quellen-Snapshot: 39 Primärquellen mit kryptografischer Prüfsumme und Inhaltsprobe, damit kein Modell auf einer leeren Seite eine Fundstelle erfindet.

4 Der Intelligenz-Einsatz hinter Wilhelm

Wilhelm-bereinigt, projektfremde Läufe entfernt

MODELL-INTERAKTIONEN FÜR
WILHELM GESCHÄTZT

~2.200

SCHICHT 1 • DIE MASCHINE

200+

orchestrierte Worker-Läufe

Automatisch delegierte Beschaffungs- und Prüfaufträge über fünf Worker-Familien. Um projektfremde Kundenläufe bereinigt.

SCHICHT 2 • DIE ARCHITEKTIN

~2.000

direkte Läufe in ~85 Sessions

ADA im Hauptsystem: Karten geschrieben, Wellen konzipiert, Briefings formuliert, Korrekturen eingearbeitet. Das Denken hinter dem System.

SCHICHT 3 • FACH-REVIEW

~300

manuelle Reviews (10–20%)

Gezielte Gegenprüfung von Hand über Codex und GPT-Modelle, dort wo ein zweites, unabhängiges Urteil den Ausschlag gab.

Grundlage der Schätzung. 88 Projektstage, 20 Releases, 5 Recherche-Wellen, 73 Karten und rund 414 Arbeitsdokumente. Angesetzt: Architektur-Arbeit an etwa 40 aktiven Tagen, rund 2 Sessions je Tag, rund 25 Modell-Läufe je Session. Die orchestrierte Maschine wurde um Läufe anderer Mandate bereinigt. Gezählte Werte (Bestand, Versionen, Arbeitsdokumente, der vollständig protokollierte 20.08.) sind an anderer Stelle als solche ausgewiesen; die Werte dieser Seite sind bewusst konservativ modelliert.

DAS MODELL-ENSEMBLE • ELF MODELLE AUS FÜNF FAMILIEN, JEDES IN EINER ROLLE

MODELLKLASSE	ROLLE IM SYSTEM
● Claude Opus	Trust-Plane: Rechtsauslegung, finales Urteil, Freigabe-Entscheidung
● GPT-5-Klasse (Codex)	Beschaffung, Bulk-Arbeit, deterministische Existenz- und Verweis-Prüfung
● Google Gemini	Vollständigkeitsprüfung („was fehlt?“) — bewusst nie für Recht, Risiko, Security
● Kimi (Moonshot)	Reichweite und Red-Team: findet, was engere Suchen übersehen
● Claude Sonnet	Ausgangs- und Freigabepässe vor der Auslieferung
● GLM, Haiku u. a.	Neben- und Spezialrollen im Ensemble

Die Ökonomie dahinter. Teure Modelle liefern nur das Urteil. Beschaffung und Prüfung laufen über Abo-Modelle ohne Kosten pro Lauf; nur Reichweite und Red-Team verursachen Cent-Beträge. Am vollständig protokollierten Abschlusstag: 2,75 USD externe Modellkosten bei 0 manuell verbrauchten Token.

5 Ein Tag als Beleg

20.08. — vollständig protokolliert, Modell für Modell

STUFE	LÄUFE	BESETZUNG	ERGEBNIS
Erhebung	22	vier Modell-Familien, geprüfter Quellen-Snapshot statt Live-Abruf	neun Kartenentwürfe möglich
Gegencheck	9	Reichweite · Vollständigkeit · Existenz · automatische Verweis-Prüfung	11 Korrekturen , davon eine tragend
Abschluss-Beschaffung	3	Codex, kostenneutral	zwei Lücken geschlossen, ein Bestandsfehler gefunden

43/43

FUNDSTELLEN
EXISTIEREN

34/34

QUERVERWEISE
INTAKT

0

ERFUNDENE
AKTENZEICHEN

0

MANUELL
VERBRAUCHTE TOKEN

2,75^{USD}

EXTERNE KOSTEN DES
TAGES

An diesem Tag meldete der Reichweitenprüfer rund fünfzig Fundstellen. Nach der Gegenprüfung hielten elf stand: vier wurden am Primärtext widerlegt, rund zwanzig waren Lücken des Prüfstands, nicht der Quelle. Hätte man die Rohliste ungeprüft übernommen, wäre ein vernichtendes Urteil über korrekte Arbeit entstanden. Genau dafür sitzt ein Orchestrator zwischen Modell und Ergebnis.

BITEME AI OS · QUALITÄTSPRINZIP

6 Was das System an Qualität trägt

Der Teil, den man an Zahlen nicht sieht

- **Modellübergreifende Gegenlesung.** Kein Modell prüft seine eigene Arbeit, und kein Modell derselben Familie prüft die des anderen. Selbst-Review zählt nie.
- **Halluzinations-sichere Rollen.** Für Vollständigkeitsfragen kommt ein Modell ohne Rechercheauftrag zum Einsatz: in dieser Rolle kann es nichts erfinden, sondern nur echte Lücken benennen.
- **Sechs Rechercheregeln**, jede aus einem realen Fehler entstanden. Die jüngste: ein Suchrezept wird erst am bekannt vorhandenen Fall verifiziert, bevor überhaupt ein Nichtfund gemeldet werden darf.
- **Eingefrorener Quellen-Snapshot** statt Live-Abruf, weil ein technischer Erfolg beim Abruf eine Auslieferung belegt, aber noch keinen Inhalt.
- **Primärquellen-Gate.** Keine Aussage in den Bestand ohne belegte Primärquelle, geführt über einen maschinellen Karten-Index als Wahrheitsliste.
- **Kontrolliertes Altern.** Jede Karte trägt ein Ablaufdatum; überholtes Wissen wird archiviert statt still weitergetragen.

7 Was wir gelernt haben

Qualität entsteht durch Korrektur, nicht durch Perfektionsanspruch

Ein ehrliches System zeigt, wo es korrigiert hat. Jeder der folgenden Punkte hat das System nicht geschwächt, sondern eine dauerhafte Regel erzeugt. Genau das unterscheidet einen geprüften Wissensspeicher von einem gut befüllten Ordner.

50 → 11 Über-Meldung abgefangen

Von rund fünfzig gemeldeten Fundstellen hielten elf der Nachprüfung stand. Ohne den Orchestrator dazwischen wäre gute Arbeit zu Unrecht verworfen worden. **Lehre:** zwischen Modell und Ergebnis gehört eine prüfende Instanz.

202 / 0 Byte Verschwendete Live-Abrufe

Live-Abrufe von Amtsportalen lieferten technische Erfolge ohne Inhalt und damit Fehlschlüsse. **Lehre:** Quellen einmal einfrieren, mit Prüfsumme sichern, dann daraus arbeiten. Spart Läufe und schließt eine Fehlerquelle.

2× Messfehler selbst korrigiert

Zwei Bestandszahlen wurden während der Erhebung zunächst falsch gemessen (falscher Ordner) und noch im selben Lauf berichtigt. **Lehre:** zuerst den Messaufbau prüfen, dann den Prüfling. Ein Nullwert ist oft ein Werkzeugfehler.

7 / 20 Mut zum Korrektur-Release

Sieben der zwanzig Versionen fügten nichts Neues hinzu, sondern prüften Bestehendes nach. **Lehre:** Nachprüfen ist kein verlorener Aufwand, sondern die Substanz. Es ist der Grund für 69 verifizierte Karten.

Der rote Faden: Jeder gefundene Fehler ist als Regel ins System zurückgeflossen. Sechs Rechercheregeln, ein Material-Gate, ein Primärquellen-Gate und die Pflicht zur modellübergreifenden Gegenlesung sind sämtlich aus konkreten, korrigierten Fehlern entstanden.

8 Was das für Sie bedeutet

Wilhelm ist kein befüllter Ordner, sondern ein Fertigungsverfahren für geprüftes Wissen. Was dieses Verfahren für Wilhelm geleistet hat, leistet es genauso für Ihre Domäne.

Jede Aussage ist belegt

Herkunft, Prüfdatum und Prüfer stehen an jeder Karte. Kein Satz ohne Primärquelle, nichts, das nicht einmal jemand gegen den Originaltext gehalten hat.

Das System bleibt aktuell

Ablaufdaten und Refresh-Wellen halten den Bestand frisch. Überholtes wird archiviert und benannt, nicht still mitgeschleppt.

Hohe Qualität, geringe Kosten

Elf Modelle in klaren Rollen, teure nur fürs Urteil. Das hält den externen Modellaufwand im Cent- bis niedrigen Dollarbereich pro Arbeitstag.

Reproduzierbar und auditierbar

Versionshistorie, Prüfprotokolle und eingefrorene Quellen machen jeden Stand nachvollziehbar. Auch der Weg ist dokumentiert, nicht nur das Ergebnis.

Was einen geprüften Wissensspeicher von einer Materialsammlung unterscheidet, ist nicht die Menge, sondern dass **jede Zahl einmal gegen die Primärquelle gehalten wurde** und jeder gefundene Fehler zu einer Regel geworden ist.

biteme.digital

KI-Serviceprovider · Custom AI · Compliance-Wissenssysteme · Wien · Stand 25.08.2026