

Schwärzungs-Wächter: Technische Beschreibung und Messprotokoll

FabStack GmbH · Stand: 12.08.2026 · Für Datenschutzbeauftragte und IT-Verantwortliche

1. Zweck und Einordnung

Der Wächter pseudonymisiert Eingaben der Beschäftigten, bevor eine Anfrage den Schulserver in Richtung eines Sprachmodells verlässt. Er anonymisiert nicht: Die Verarbeitung bleibt personenbezogen im Sinne der DSGVO und gehört in das Verarbeitungsverzeichnis der Schule (Textbaustein liegt der Mappe bei).

2. Funktionsweise (drei Schichten)

1. **Namensliste der Schule:** Abgleich gegen den Export der Schulverwaltung (z. B. EduPage; Spalten Klasse, Nachname, Vorname). Erkennt vollen Namen, Vorname, Nachname, die Schreibweise "Nachname, Vorname", Genitivformen ("Lenas", "Jonas") sowie abweichende Groß- und Kleinschreibung; fremde Doppelnamen (z. B. "Kleinschmidt-Meier" bei gelistetem "Kleinschmidt") werden nicht teil-ersetzt. Ersetzung durch stabile Platzhalter je Person ("[Schüler A]"); die Muster-Schicht läuft vor der Namensschicht, damit Namen innerhalb von E-Mail-Adressen die Adresse nicht zerlegen. Die Liste wird bei Änderung automatisch neu geladen (einmal je Schuljahr aktualisiert, zum Schuljahresende gelöscht).
2. **Deterministische Muster:** E-Mail-Adressen, Telefonnummern (Mobil und Festnetz, +49/0049/0), Geburtsdaten (TT.MM.JJJJ, auch einstellig).
3. **Rest-Fänger (Ausbaustufe):** lokale Namenserkennung für unbekannte Namen (Presidio-Schichtenmodell), vor Produktivnahme mit erneuter Recall-Messung.

Ablauf: Eingabe im Browser → Plattform-Server: Wächter ersetzt, Prüfansicht zeigt jede Ersetzung mit Quelle → Modell erhält die pseudonymisierte Fassung → Rücktausch-Zuordnung verbleibt auf dem Plattform-Server. **Plattform-Server heißt:** standardmäßig eine eigene Instanz je Schule in einem deutschen Rechenzentrum, betrieben von FabStack als Auftragsverarbeiter; auf Wunsch derselbe Aufbau auf Hardware des Trägers. Namensliste und Zuordnungstabelle liegen ausschließlich dort und werden nie an Modellanbieter übermittelt (auf Wunsch Abgleich gegen eine gehashte Liste).

Diktierfunktion, gesondert betrachtet: Die Spracherkennung (Whisper) läuft bei einem EU-Dienstleister als Auftragsverarbeiter. Die Tonaufnahme durchläuft den Wächter nicht; gesprochene Namen würden das Haus unpseudonymisiert verlassen. Diktat ist deshalb in der Dienstanleitung auf grüne Aufgaben beschränkt; auf Wunsch wird es schulweit abgeschaltet oder als lokale Spracherkennung auf dem Server des Trägers betrieben. Diese Verarbeitung gehört als eigene Position in das Verarbeitungsverzeichnis.

3. Bekannte Grenzen (bewusst dokumentiert)

Nicht erkannt werden: Spitznamen und Namen außerhalb der Liste, Tippfehler in Namen, identifizierende Umschreibungen über den Kontext ("der Junge aus der 7b mit dem gebrochenen Arm"), Namen Dritter in eingefügten Fremdtexen (z. B. Geschwister in Eltern-E-Mails). Gegenmaßnahmen: Prüfansicht vor dem Absenden, Ampel-Dienstanweisung (Datensparsamkeit), Schulung, Rest-Fänger als Ausbaustufe.

4. Messprotokoll vom 12.08.2026 (Stufe 2.1)

Testset: 24 Fälle, davon 19 mit zu erkennender personenbezogener Angabe (Namen in vier Schreibweisen, Genitiv mit und ohne Apostroph, Versalien und Kleinschreibung, Bindestrich-Name, zwei Personen in einem Satz, Mobil- und Festnetznummer, Geburtsdatum zweistellig und einstellig, E-Mail-Adresse, Kombination Name+E-Mail+Telefon) und 5 Gegenproben, die nicht ersetzt werden dürfen (Klassenbezeichnung ohne Person, Fachtext, Uhrzeit, Jahreszahl, fremder Doppelname).

Ergebnis: 24 von 24 Fällen korrekt; alle 23 enthaltenen personenbezogenen Angaben ersetzt (Recall 100 %), keine Fälschersetzung in den Gegenproben.

Einordnung: Dies ist ein Laborwert auf unserem eigenen Testset und kein Betriebsversprechen. Das Testskript wird der Schule offengelegt; vor Produktivnahme der gelben Stufe wird die Messung mit Testfällen der Schule wiederholt und protokolliert.

5. Speicherung und Protokolle

Das Sprachmodell wird ohne Speicherung der Inhalte betrieben (vertraglich und, wo verfügbar, technisch erzwungen; Nachweise je Steckplatz in der Mappe). Serverprotokolle der Plattform enthalten Zugriffszeiten, keine Chatinhalte, Löschung nach 14 Tagen. Der Administratorzugriff auf fremde Chatverläufe ist in der Auslieferung abgeschaltet; es handelt sich um eine Einstellung, deren Änderung protokolliert wird und laut Musterdienstvereinbarung die vorherige Zustimmung der Personalvertretung voraussetzt.