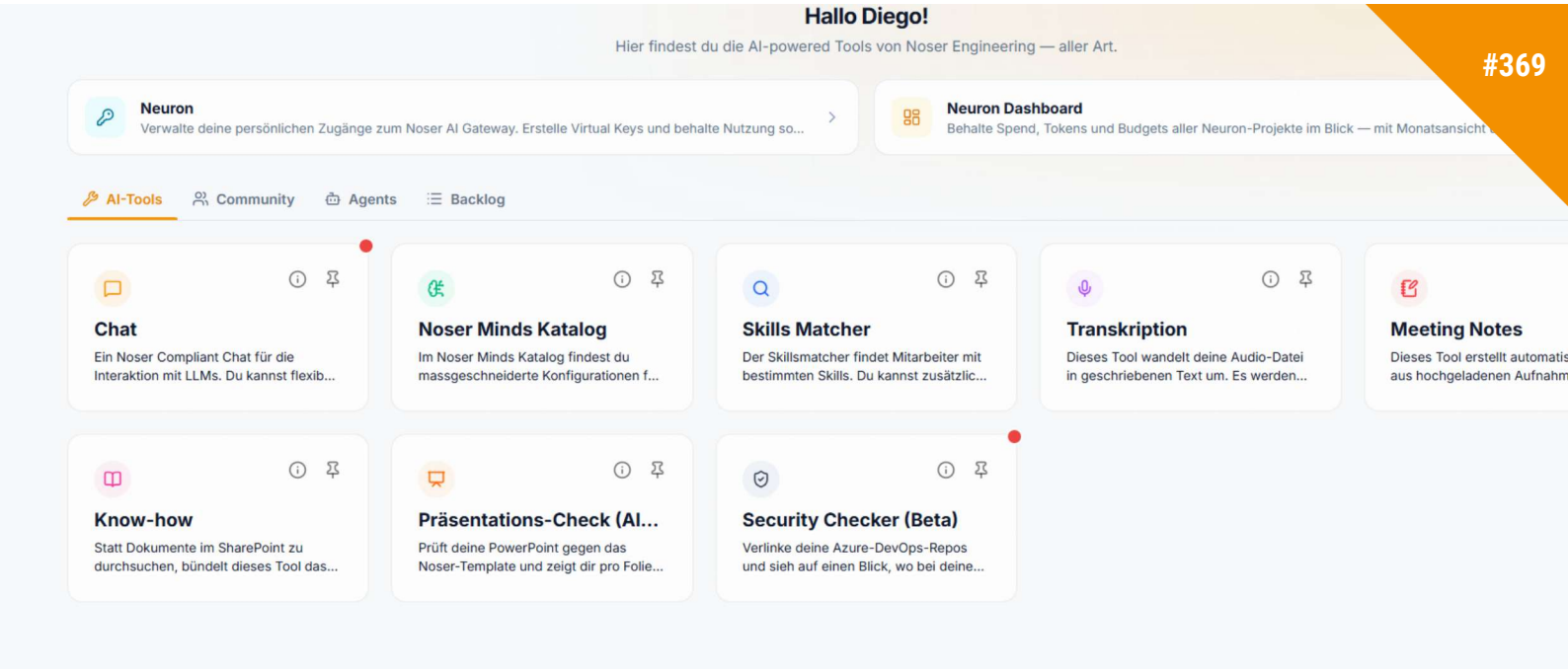
NOSER ENGINEERING



Noser AI Toolbox



NOSER ENGINEERING

Bei AI verändert sich technologisch extrem viel. Deshalb wollten wir keine Lösung bauen, die an ein einzelnes Modell oder einen einzelnen Anbieter gebunden ist. Entstanden ist eine Plattform, die wir laufend erweitern können und bei der Security, Berechtigungen und Kosten von Anfang an Teil der Architektur sind. Zu sehen, wie intensiv sie heute genutzt wird, macht mich schon stolz.

Tobias Ritscher, AI-Engineer, Noser Engineering

Das Projekt in Kürze

Generative AI soll Mitarbeitende im Arbeitsalltag unterstützen, ohne dass Unternehmen dabei die Kontrolle über vertrauliche Daten, Berechtigungen und Kosten verlieren. Noser Engineering entwickelte dafür mit der Noser AI Toolbox (NAIT) eine eigene, mandantenfähige Plattform. Sie verbindet leistungsfähige Sprachmodelle mit internen Daten und Systemen und macht unterschiedliche AI-Werkzeuge zentral zugänglich.

Ausgangslage
Mitarbeitende wollten generative AI einsetzen, griffen dafür jedoch häufig auf frei verfügbare Chat-Dienste zurück. Damit fehlte die Kontrolle darüber, wo Daten verarbeitet werden. Getrennte Abonnements erschwerten zugleich den Überblick über Nutzung und Kosten. Interne Wissensquellen und Systeme wie SharePoint, Azure DevOps oder Jira waren nicht eingebunden.

Ziel
Gesucht war eine zentrale Plattform, die leistungsfähige Sprachmodelle mit internen Daten und Systemen verbindet: sicher, transparent in den Kosten und jederzeit erweiterbar.

Lösung
Noser Engineering entwickelte eine modulare Plattform aus zwei zusammenwirkenden Systemen: Die AI Toolbox stellt den Mitarbeitenden die Werkzeuge und Benutzeroberflächen zur Verfügung. Neuron übernimmt als zentraler LLM-Gateway die Modellanbindung sowie die Steuerung von Freigaben, Budgets und Kosten.

Ergebnis
Die Plattform bündelt den betrieblichen AI-Einsatz in einer kontrollierten Umgebung. Mitarbeitende können unterschiedliche Modelle und Werkzeuge nutzen, während Datenzugriffe an ihre bestehenden Berechtigungen gebunden bleiben. Das Unternehmen erhält Transparenz über Verbrauch und Kosten und kann Modelle differenziert pro Firma freigeben.

Ausgangslage: AI-Nutzung ohne gemeinsame Leitplanken

Generative AI eröffnet neue Möglichkeiten für Recherche, Wissensarbeit, Softwareentwicklung und die Bearbeitung wiederkehrender Aufgaben. In der betrieblichen Praxis entsteht jedoch schnell eine fragmentierte Tool-Landschaft: Mitarbeitende nutzen unterschiedliche Dienste, jedes Angebot besitzt ein eigenes Abonnement, und sensible Informationen werden teilweise ausserhalb der kontrollierten Unternehmensumgebung verarbeitet.

Ein weiterer Engpass liegt beim Unternehmenswissen. Frei verfügbare Chat-Dienste kennen weder interne Dokumente noch Projektdaten oder etablierte Systeme. Relevantes Wissen liegt jedoch gerade in Anwendungen wie SharePoint, Azure DevOps, Jira oder Confluence. Ohne eine kontrollierte Verbindung bleiben viele AI-Anwendungen auf allgemeines Wissen beschränkt.

Für Noser Engineering ergab sich daraus eine konkrete Aufgabe: AI sollte breit nutzbar werden, ohne Abstriche bei Datenschutz, Governance und Kostenkontrolle. Statt weitere Einzellösungen einzuführen, sollte eine gemeinsame technische Grundlage entstehen, die sich schrittweise ausbauen lässt.

Ziel: Ein kontrollierter Zugang zu Modellen, Daten und Werkzeugen

Im Zentrum stand eine Plattform, über die Mitarbeitende auf Sprachmodelle verschiedener Anbieter zugreifen können. Die Modelle sollten sich austauschen und gezielt freigeben lassen. So bleibt das Unternehmen flexibel, wenn neue Anbieter, Modellgenerationen oder spezialisierte Anwendungen hinzukommen.

Die Plattform sollte gleichzeitig klare organisatorische und technische Leitplanken setzen:
- datenschutzkonforme Verarbeitung in der eigenen Azure-Umgebung
- transparente Budgets pro Person, Projekt und Firma
- kontrollierter Zugang zu internen Daten und Systemen
- konsequente Verwendung der Berechtigungen der angemeldeten Person
- strikte Trennung von Daten, Modellen und Budgets zwischen den Firmen
- schnelle Integration zusätzlicher Modelle und AI-Werkzeuge

Herausforderung: Ein dynamischer AI-Markt trifft auf Unternehmensanforderungen

Die Plattform musste Modelle verschiedener Anbieter über einen einheitlichen Zugang verfügbar machen. Gleichzeitig sollten Modellfreigaben und Budgets in Echtzeit durchgesetzt werden. Diese Anforderungen betreffen nicht nur die Benutzeroberfläche, sondern die gesamte Architektur: Identitäten, Berechtigungen, Datenzugriffe, Modellkommunikation, Verbrauchserfassung und Mandantentrennung müssen zuverlässig zusammenspielen.

Besonders anspruchsvoll war der Zugriff auf Unternehmenssysteme. Das Modell darf nur Informationen verwenden, auf welche die angemeldete Person ohnehin zugreifen kann. Bestehende Zugriffsrechte mussten deshalb auch bei AI-gestützten Abfragen erhalten bleiben. Für mehrere Firmen galt zugleich eine strikte Trennung ihrer Daten, Budgets und verfügbaren Modelle.

Hinzu kam die hohe Veränderungsgeschwindigkeit des AI-Markts. Neue Modelle und Werkzeuge sollten integrierbar sein, ohne die Plattform bei jeder Erweiterung grundlegend umbauen zu müssen. Modularität und Erweiterbarkeit wurden daher zu zentralen Architekturprinzipien.

Vorgehen: Iterative Entwicklung aus dem Arbeitsalltag heraus

Noser Engineering entwickelte die Plattform iterativ und orientierte sich eng an den konkreten Bedürfnissen der Mitarbeitenden. Neue Werkzeuge wurden früh im Arbeitsalltag eingesetzt, überprüft und auf Basis der Erfahrungen weiterentwickelt.

Die technische Basis bildet eine Cloud-native Architektur auf Microsoft Azure. Das Backend wurde mit .NET, C# und .NET Aspire umgesetzt, das Frontend mit ReactJS und TypeScript. PostgreSQL mit pgvector unterstützt die Datenhaltung und Vektorsuche. LiteLLM dient als technologische Grundlage für den LLM-Gateway. Microsoft Entra ID übernimmt Single Sign-on und Berechtigungen. Isolierte Code-Ausführungen erfolgen über Azure Container Apps, die Entwicklungs- und Bereitstellungsprozesse über Azure DevOps.

Die Architektur folgt vier Grundsätzen: Jedes Werkzeug wird als eigenes Modul geführt, Security wird von Beginn an berücksichtigt, Mandanten bleiben voneinander getrennt und die Plattform lässt sich um weitere Anwendungsfälle erweitern.

Die Lösung: AI Toolbox und Neuron

AI Toolbox: AI-Werkzeuge in einer gemeinsamen Arbeitsumgebung

Die AI Toolbox ist die Arbeitsumgebung für die Mitarbeitenden. Ihr Kern ist ein agentischer Chat, der kontrolliert auf Microsoft 365, SharePoint, Azure DevOps, Jira, Confluence und die Websuche zugreifen kann. Jeder Zugriff erfolgt mit den Berechtigungen der angemeldeten Person.

Neuron: Steuerung von Modellen, Budgets und Kosten

Neuron bildet den zentralen LLM-Gateway der Plattform. Das System bindet Modelle verschiedener Anbieter über eine einheitliche Schnittstelle an. Verantwortliche können Modelle pro Firma freigeben und Budgets pro Person, Projekt oder Firma definieren. Mitarbeitende erhalten eigene API-Keys, die sich beispielsweise für AI-gestützte Softwareentwicklung einsetzen lassen. Ein Dashboard macht Verbrauch und Kosten nachvollziehbar. Damit bleibt die Modellnutzung steuerbar, während Mitarbeitende je nach Aufgabe auf unterschiedliche freigegebene Modelle zugreifen können.

Business Impact: Erfolgreicher Release und wachsende Nutzerzahlen

Seit dem Go-live im Januar 2026 hat sich die Plattform als zentrales AI-Werkzeug etabliert. Über Neuron laufen monatlich mehr als 900´000 AI-Anfragen. Den Anwenderinnen und Anwendern stehen dabei über 40 Sprachmodelle zur Verfügung.

Verbrauch und Kosten sind pro Person, Projekt und Firma nachvollziehbar. Gleichzeitig dient die Plattform als praktisches Entwicklungs- und Lernumfeld: Erkenntnisse zu agentischer AI, Retrieval-Augmented Generation und dem Model Context Protocol fliessen direkt in Kundenprojekte ein.

Das Projekt zeigt: Für den produktiven Einsatz generativer AI braucht es neben leistungsfähigen Modellen eine belastbare Architektur, kontrollierte Datenzugriffe, klare Berechtigungen und transparente Kosten.

Über den Kunden: Noser Engineering

Noser Engineering ist ein Schweizer Engineering- und Lösungspartner für anspruchsvolle digitale Systeme mit Kompetenzen in Software, Embedded, Cloud, AI und Security. Mit der eigenen AI-Plattform schafft das Unternehmen die Grundlage, generative AI sicher, kontrolliert und mit konkretem Nutzen im Arbeitsalltag einzusetzen.

Benötigen auch Sie Unterstützung für die Umsetzung Ihrer Vision?



Kontaktieren Sie unsere Expert:innen für ein unverbindliches Beratungsgespräch.

Oliver Buntefuss, Head of Sales, +41 79 827 82 07, oliver.buntefuss@noser.com