

ANHANG G

Tool- und Anbieter-Vergleich

Zusatzmaterial zum Buch „PREDICTABLE“ von Thorsten Pätzold —
Amazing Agents.

ANHANG G

Tool- und Anbieter-Vergleich

Dieser Anhang ist eine kuratierte Übersicht über die Anbieter und Tools, mit denen wir bei Amazing Agents in Mittelstands-Projekten regelmäßig arbeiten. Es ist keine vollständige Marktübersicht — und keine Bezahl-Empfehlung. Wir erhalten von keinem der genannten Anbieter Provision.

Die Auswahl folgt drei Kriterien: produktive Reife (kein experimentelles Tool aufgenommen), DSGVO-Tauglichkeit (oder zumindest belastbare Wege dorthin), und Mittelstands-Geeignetheit (kein Enterprise-only-Tool mit unzugänglichen Pricing-Modellen). Was Sie auf den nächsten Seiten finden, ist eine ehrliche Einordnung: Wo glänzt der Anbieter, wo nicht, was kostet das, wann sollten Sie es einsetzen, wann nicht.

„Tools sind austauschbar. Architektur und Disziplin nicht.“

G.1 Modell-Anbieter

Die Modell-Anbieter sind das Foundation-Layer Ihres Setups. Hier zahlen Sie pro Token — die Kosten sind transparent, die Performance unterscheidet sich nicht mehr massiv zwischen den Top-Anbietern. Die wichtigeren Unterscheidungs-Kriterien sind heute: Datenresidenz, AVV-Qualität, Tool-Calling-Reife, Reasoning-Fähigkeiten.

MODELL-ANBIETER

Anthropic (Claude)

Stärken	Sehr starkes Reasoning, exzellentes Tool-Calling, lange Kontext-Fenster (200k Tokens), klare Sicherheits-Architektur
Schwächen	Höhere Token-Kosten als manche Alternativen, US-Hosting (Standard), kein Bild-Generation
DSGVO	AVV verfügbar, EU-Hosting in Vorbereitung, Daten werden nicht für Training verwendet (Opt-out per Default)
Pricing-Klasse	Premium (ca. 3 EUR pro 1M Input-Tokens, 15 EUR pro 1M Output-Tokens für Sonnet-Klasse)
Wann passend	Anspruchsvolle Reasoning-Aufgaben, lange Dokumenten-Analyse, regulierte Bereiche mit hohen Anforderungen an Output-Qualität
Wann nicht	Sehr kostensensitive Massenanwendungen, Bedarf an Bild-Generierung, lokale Deployment-Anforderungen

MODELL-ANBIETER

OpenAI (GPT)

Stärken	Größtes Ökosystem, viele Tools und SDKs, sehr breite Modell-Palette inkl. Bild- und Audio-Generation, starke Reasoning-Modelle (o-Familie)
Schwächen	Datenschutz war historisch problematisch (mittlerweile besser), Modell-Versionierung manchmal verwirrend
DSGVO	Enterprise-Plan mit AVV und Daten-Residenz-Optionen, Standard-API mit AVV verfügbar, Daten-Training Opt-out möglich
Pricing-Klasse	Mittel bis Premium, je nach Modell stark variabel (von 0,15 bis 60 EUR pro 1M Output-Tokens)
Wann passend	Multi-Modal-Anwendungen, breites Tooling-Ökosystem nötig, agile Entwicklungs-Teams
Wann nicht	Wenn Sie Bedenken haben gegen US-Anbieter, sehr regulierte Branchen mit strenger Datenresidenz-Pflicht

MODELL-ANBIETER

Mistral AI

Stärken	EU-basiert (Frankreich), Modelle in EU gehostet möglich, Open-Source-Optionen, gute Performance-Kosten-Ratio
Schwächen	Modelle leicht hinter US-Spitze in Reasoning, Tool-Calling weniger ausgereift, kleineres Ökosystem
DSGVO	EU-Hosting Standard, sehr gute Position aus deutscher / europäischer Compliance-Sicht
Pricing-Klasse	Mittel (oft 30–50% günstiger als Premium-US-Anbieter)
Wann passend	Compliance-getriebene Setups, EU-Datenresidenz Pflicht, kostensensitive Anwendungen
Wann nicht	Anspruchsvolle Reasoning-Aufgaben, wenn Sie das beste verfügbare Modell brauchen

MODELL-ANBIETER

Microsoft Azure OpenAI

Stärken	Identische Modelle wie OpenAI, aber mit MS-Compliance-Wrapper, EU-Region wählbar, integriert mit Microsoft 365 / Entra ID, hohes Vertrauen bei IT-Abteilungen
Schwächen	Höhere Preise als OpenAI direkt, Modell-Updates kommen verzögert
DSGVO	Sehr gute Position, Standard-AVV, EU-Datenresidenz wählbar, Enterprise-Compliance-Suite
Pricing-Klasse	Premium (ca. 10–20% Aufschlag gegenüber OpenAI direkt)
Wann passend	Microsoft-zentrierte IT-Landschaft, hohe Compliance-Anforderungen, Konzern-Strukturen
Wann nicht	Sehr kostensensitive Setups, agile Startup-Umgebungen ohne MS-Bindung

MODELL-ANBIETER

Google Gemini

Stärken	Sehr lange Kontext-Fenster (bis 2M Tokens), starke Multi-Modal-Fähigkeiten, gute Integration mit Google Workspace
Schwächen	Tool-Calling-Reife teils hinter Konkurrenz, Modell-Verhalten manchmal unkonventionell
DSGVO	Vertex AI mit EU-Region verfügbar, AVV verfügbar, Daten-Training Opt-out
Pricing-Klasse	Mittel
Wann passend	Lange Dokumente / Audio-Transkripte, Google Workspace-Umfeld, Multi-Modal-Anwendungen
Wann nicht	Wenn Sie maximale Tool-Calling-Reife brauchen oder ein zurückhaltenderes Modell-Verhalten bevorzugen

MODELL-ANBIETER

Lokale Open-Source-Modelle

Stärken	Volle Datenkontrolle, keine Token-Kosten (nur Hardware), unabhängig von Anbieter-Strategien, Llama / Mixtral / DeepSeek inzwischen sehr leistungsfähig
Schwächen	Hardware-Investition (50–500k EUR), eigenes Operations-Knowhow nötig, Modelle leicht hinter Top-Cloud-Anbietern
DSGVO	Beste mögliche Position — Daten verlassen das Haus nie
Pricing-Klasse	Hohe Fixkosten, niedrige Grenzkosten
Wann passend	Hochsensitive Daten (Anwaltskanzleien, Gesundheit, Behörden), sehr hohe Volumina, Souveränitäts-Anforderungen
Wann nicht	Niedrige bis mittlere Volumina, fehlende IT-Tiefe, agile Iteration mit häufigen Modell-Wechseln

G.2 Orchestrierungs-Plattformen

Diese Tools verbinden Modelle, Tools und Workflows zu produktiven Agenten. Sie unterscheiden sich stark in Zielgruppe (No-Code vs. Developer), Flexibilität und Operations-Reife.

ORCHESTRIERUNG

n8n

Stärken	Self-Hostable (DSGVO-Traum), sehr breite Integrations-Bibliothek, visuelles Workflow-Design, kombinierbar mit eigenem Code, faires Pricing
Schwächen	Steile Lernkurve, manchmal performance-kritisch bei komplexen Workflows, weniger AI-spezifisch als Wettbewerber
DSGVO	Top — Self-Hosting möglich, alternativ Cloud mit EU-Region und AVV
Pricing-Klasse	Niedrig (Self-Host komplett kostenlos für Self-Hosting, Cloud ab ~20 EUR / Monat)
Wann passend	Mittelständler mit IT-Affinität, Datenschutz-getriebene Setups, hochintegrierte Workflows mit Bestandssystemen
Wann nicht	Reine Business-User ohne Technik-Background, sehr AI-zentrische Use Cases

ORCHESTRIERUNG

Make (ehemals Integromat)

Stärken	Sehr Business-User-freundlich, große Template-Bibliothek, hohe Stabilität, einfaches Pricing-Modell
Schwächen	Cloud-only, US-basiert (mit EU-Optionen), bei sehr hohen Volumina teuer
DSGVO	Cloud mit EU-Datenresidenz möglich, AVV verfügbar, US-Mutterhaus aber im Backend
Pricing-Klasse	Mittel (10 bis 300 EUR / Monat je nach Volumen)
Wann passend	Business-User-Teams ohne tiefe Technik-Ressourcen, mittlere Volumina, schnelle Iteration
Wann nicht	Sehr hohe Volumina (Pricing skaliert), strenge Self-Host-Anforderungen

ORCHESTRIERUNG

LangChain / LangGraph

Stärken	Maximale Flexibilität, Open-Source, sehr aktive Community, Standardbausteine für Agents, tief in Python integriert
Schwächen	Steile Lernkurve, Entwickler-Tool (kein Klick-Setup), Updates manchmal Breaking Changes, höherer Wartungsaufwand
DSGVO	Hängt vom Setup ab — Sie hosten selbst, daher liegt die DSGVO-Verantwortung bei Ihnen
Pricing-Klasse	Open-Source kostenlos, LangSmith für Observability ab ~50 EUR / Monat
Wann passend	Entwickler-Teams, komplexe Agent-Architekturen, maximale Kontrolle gewollt
Wann nicht	Reine Business-User, schnelle Marktexperimente, wenig Engineering-Ressourcen

ORCHESTRIERUNG

Zapier

Stärken	Größte Integrations-Bibliothek, sehr stabil, einfacher Einstieg, breit etablierter Marktstandard
Schwächen	Pro-Aktion-Pricing wird bei Volumen schnell teuer, KI-Features hinter Konkurrenz, US-zentriert
DSGVO	AVV verfügbar, aber EU-Datenresidenz nicht der Standard
Pricing-Klasse	Mittel bis hoch bei Volumen
Wann passend	Sehr breite Integrations-Anforderungen, schnelle einfache Workflows zwischen Standard-SaaS-Tools
Wann nicht	AI-zentrische Setups, sehr hohe Volumina, strenge Compliance

ORCHESTRIERUNG

Microsoft Copilot Studio

Stärken	Sehr enge MS-365-Integration, gute Compliance-Position für MS-Kunden, gutes Tooling für Enterprise-Setups
Schwächen	Stark MS-zentriert, Lock-in, Pricing-Modell teils intransparent
DSGVO	Sehr gut für MS-Bestandskunden, Enterprise-Compliance-Wrapper
Pricing-Klasse	Mittel bis Premium (in MS-Verträgen oft inkludiert)
Wann passend	MS-365-zentrierte Organisationen, Compliance-getriebene Konzerne, IT-Abteilungen mit MS-Affinität
Wann nicht	Heterogene Tool-Landschaften, Open-Source-Präferenz

G.3 Memory- und Vector-Datenbanken

Memory-Layer brauchen spezialisierte Speicher. Hier die wichtigsten Optionen für Mittelstands-Setups.

VECTOR-DATENBANK

pgvector (PostgreSQL-Erweiterung)

Stärken	Open-Source, in PostgreSQL integriert (das Sie wahrscheinlich eh haben), gute Performance bis Millionen Vektoren, kein neuer Stack nötig
Schwächen	Bei sehr großen Skalen (>10M Vektoren) Performance-Limits, kein dedizierter Cloud-Service
DSGVO	Sie hosten selbst, daher volle Kontrolle
Pricing-Klasse	Niedrig (Self-Host kostenlos, Managed über Anbieter wie Neon ab ~20 EUR / Monat)
Wann passend	Mittelständler mit PostgreSQL-Stack, niedrige bis mittlere Volumina, einfache RAG-Setups
Wann nicht	Sehr große Skalen (>10M Vektoren) oder hochfrequente Realtime-Workloads

VECTOR-DATENBANK

Qdrant

Stärken	Open-Source, sehr performant, gut skalierbar, Self-Hosting straightforward, klare API
Schwächen	Eigenes System, das gewartet werden will, kleinerer Markt als Pinecone
DSGVO	Self-Host = volle Kontrolle, Cloud mit EU-Region verfügbar
Pricing-Klasse	Niedrig (Self-Host kostenlos, Cloud ab ~25 EUR / Monat)
Wann passend	Höhere Volumina, dedizierter Vector-Stack gewünscht, Self-Hosting bevorzugt
Wann nicht	Wenn Sie schon einen Postgres-Stack haben und keine sehr hohen Volumina brauchen

VECTOR-DATENBANK

Pinecone

Stärken	Marktführer im Cloud-Vector-DB-Bereich, sehr stabil, einfaches API, gute Performance bei Skalierung
Schwächen	Cloud-only (kein Self-Host), US-basiert, bei niedrigen Volumina teuer
DSGVO	AVV verfügbar, EU-Region möglich, US-Mutterhaus
Pricing-Klasse	Mittel (Free Tier, dann ab ~70 EUR / Monat)
Wann passend	Wenn Sie Cloud-only akzeptieren und einen einfachen, stabilen Service wollen
Wann nicht	Strenge Self-Host-Anforderungen, sehr kostensensitive Setups

G.4 Observability und Monitoring

Wer einen Agenten produktiv betreibt und kein Observability-Tool hat, fliegt blind. Diese Tools machen Agent-Verhalten messbar und debugbar.

OBSERVABILITY

Langfuse

Stärken	Open-Source, Self-Hostable, sehr gute Tracing-Funktionen, breite SDK-Unterstützung, stark wachsende Community
Schwächen	Junges Tool (manche Features noch in Bewegung), erfordert etwas Setup
DSGVO	Self-Host = volle Kontrolle, Cloud mit EU-Region
Pricing-Klasse	Open-Source kostenlos, Cloud ab ~50 EUR / Monat
Wann passend	Engineering-Teams, die Tracing und Eval-Workflows ernst nehmen, DSGVO-getrieben
Wann nicht	Sehr kleine Setups ohne Engineering-Ressourcen

OBSERVABILITY

LangSmith

Stärken	Tiefe LangChain-Integration, gutes Eval-Tooling, kommerziell etabliert
Schwächen	Stark an LangChain gekoppelt, Cloud-only
DSGVO	AVV verfügbar, EU-Region in Bewegung
Pricing-Klasse	Mittel (Free Tier, dann ab ~50 EUR / Monat)
Wann passend	LangChain-zentrische Setups, Teams mit Eval-Disziplin
Wann nicht	Wenn Sie LangChain nicht nutzen oder Self-Host brauchen

OBSERVABILITY

Helicone

Stärken	Sehr einfaches Setup (Proxy-basiert), gutes Cost-Tracking, schnelle Inbetriebnahme
Schwächen	Weniger tief als Langfuse oder LangSmith, weniger Eval-Features
DSGVO	AVV verfügbar, US-basiert
Pricing-Klasse	Niedrig (Free Tier, dann ab ~25 EUR / Monat)
Wann passend	Schneller Einstieg in Cost-Tracking und Basic-Tracing
Wann nicht	Wenn Sie tiefe Eval-Workflows oder Self-Host brauchen

G.5 Anbieter-Wahl-Matrix nach Profil

Wenn Sie sich vor der Anbieter-Wahl verlieren, hier eine pragmatische Faustregel-Matrix für die häufigsten Mittelstands-Profile.

Profil	Modelle	Orchestrierung	Memory
MS-365-Konzern	Azure OpenAI	Copilot Studio	Azure-Stack
DSGVO-getrieben EU	Mistral	n8n (self-host)	pgvector
Schneller Einstieg	OpenAI / Anthropic	Make	pgvector
Hohe Volumina	OpenAI / Mistral-Mix	n8n / LangChain	Qdrant
Hochsensitive Daten	Lokales Open-Source	n8n (on-prem)	pgvector (on-prem)
Engineering-getrieben	Anthropic + Mix	LangGraph	Qdrant + Langfuse

Diese Matrix ist eine Orientierung, kein Dogma. Die meisten Setups, die wir bei Amazing Agents bauen, kombinieren mindestens zwei Modell-Anbieter (typisch Anthropic für Reasoning und ein günstigeres Modell für Routine-Aufgaben). Polyglotte Setups sind nicht komplexer in der Wartung, aber günstiger im Betrieb und robuster gegen Anbieter-Risiken.

Dieses Dokument ist Anhang G zum Buch „PREDICTABLE — Wie KI-Agenten Ihr Business von Spielzeug zu Infrastruktur transformieren“ (Thorsten Pätzold, Amazing Agents, Pirmasens 2026). Verweise auf Kapitel beziehen sich auf das Buch, Verweise auf andere Anhänge auf die übrigen Dokumente des Werkzeugkastens. Alle zehn Anhänge: www.amazing-agents.com/predictable. © 2026 Thorsten Pätzold / Amazing Agents. Alle Rechte vorbehalten.

It's predictable when you do the math.

AMAZING AGENTS

Pirmasens · 2026

www.amazing-agents.com