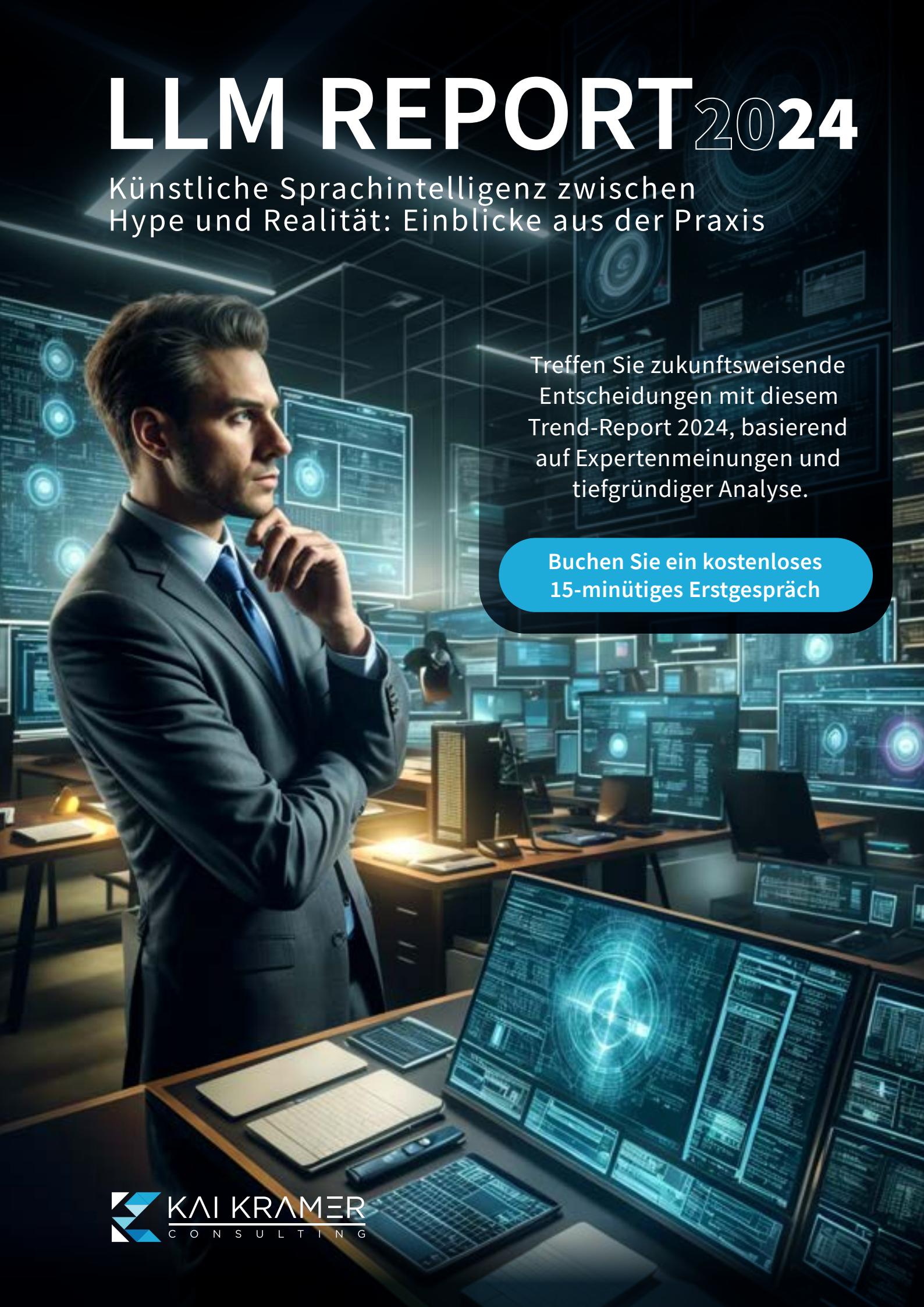
LLM REPORT 2024

Künstliche Sprachintelligenz zwischen Hype und Realität: Einblicke aus der Praxis



Treffen Sie zukunftsweisende Entscheidungen mit diesem Trend-Report 2024, basierend auf Expertenmeinungen und tiefgründiger Analyse.

Buchen Sie ein kostenloses 15-minütiges Erstgespräch

KAI KRAMER
CONSULTING

KAI KRAMER
CONSULTING

LLM REPORT 2024

INHALTSVERZEICHNIS

EINLEITUNG

„Revolutionäre Sprachmodelle für die Masse“

Der vorliegende Trend-Report bietet eine umfassende Übersicht über die Entwicklungen im Bereich künstlicher Sprachintelligenz und Large Language Models (LLMs) des Jahres 2023. Als KI-Berater mit dem Schwerpunkt auf diesen Technologien habe ich das vergangene Jahr genutzt, um die Fortschritte und Veränderungen, die durch die Einführung von OpenAI’s ChatGPT Ende 2022 entstanden sind, zu analysieren.

Die Veröffentlichung der fortschrittlichen Sprachmodelle von OpenAI, Meta und Co. markiert einen signifikanten Meilenstein in der KI-Entwicklung. Dieser besitzt das Potenzial, bestehende IT-Strukturen zu transformieren und neue Lösungsansätze zu bieten. Trotz des vorherrschenden Hypes ist es essenziell, einen klaren und objektiven Blick auf die tatsächlichen Auswirkungen und zukünftigen Potenziale dieser Technologien zu werfen.

Um eine fundierte Basis für meinen Report zu schaffen, habe ich über 50 Interviews und Gespräche mit Fachexperten, Beratern und Nutzern von LLMs geführt. Diese Gespräche dienten mir dazu, ein breites Spektrum an Perspektiven zu sammeln. Die daraus resultierenden Erkenntnisse habe ich hier einfließen lassen.

Mein Report zielt darauf ab, ein klares Verständnis für die aktuellen Trends zu schaffen, wichtige Diskussionen und Meinungen innerhalb der Branche widerzuspiegeln und einen Ausblick auf die mögliche zukünftige Entwicklung zu bieten. Die daraus gewonnenen Erkenntnisse dienen als Grundlage für strategische Entscheidungen zur Anwendung von künstlicher Sprachintelligenz.

DIE TRENDS 2023

Das Jahr 2023 stellte sich als ein entscheidendes Jahr für die Entwicklung und Anwendung von Large Language Models (LLMs) und künstlicher Sprachintelligenz heraus. Die Fortschritte und Veränderungen habe ich analysiert und beleuchte die Schlüsseltrends, die sich als besonders prägend erwiesen haben.

RETRIEVAL AUGMENTED GENERATION (RAG)

Von ChatGPT zu RAG

Als Mittel der Wahl für die Verbindung von internem Unternehmenswissen mit künstlicher Sprachintelligenz haben sich RAG-Modelle bewährt. Sie ermöglichen es, das in Datenpools verborgene Wissen effektiv für die Generierung von Antworten zu nutzen und Halluzinationen effektiv zu vermeiden. In diesem Zusammenhang haben Vektor-Suchverfahren eine große Popularität erlangt.

MARKTKONZENTRATION UND ANBIETERABHÄNGIGKEIT

One Model to Rule them All? – OpenAI als Goldstandard

OpenAI hat sich mit seinen LLMs, insbesondere im multilingualen Bereich, als Qualitätsstandard etabliert. Die hohe Qualität und die Fähigkeit, kontextreiche und grammatikalisch korrekte Texte in verschiedenen Sprachen zu generieren, haben OpenAI eine dominierende Marktstellung eingebracht.

Die Abhängigkeit von einem einzigen Anbieter wie OpenAI wird als Risiko wahrgenommen. Es besteht die Notwendigkeit für andere Anbieter, ihre Technologien zu verbessern, um mehr Vielfalt im Markt zu schaffen und eine Lock-In-Situation zu vermeiden.

HARDWAREBEDARF UND DIE ROLLE VON NVIDIA
ELON MUSK – TESLA EARNINGS CALL Q2-23:

“We actually take NVIDIA hardware as fast as NVIDIA will deliver it to us.”

In der Betrachtung der Trends für 2023 darf der Aspekt des Hardwarebedarfs, insbesondere die Rolle von NVIDIA mit ihren GPUs, nicht unerwähnt bleiben.
Die Auslieferungen der NVIDIA H100 im Jahr 2023 waren von vielen Fragezeichen geprägt. Große Player wie Microsoft und Meta haben beachtliche Mengen dieser Chips bezogen, wobei Meta’s Bedarf besonders bemerkenswert war und Spekulationen über ihre internen KI-Projekte aufwirft.

Interessanterweise liegen Google und Amazon mit ihren Bestellungen von nur 50K Einheiten verhältnismäßig niedrig, was darauf hindeuten könnte, dass sie in hohem Maße auf eigene Hardware setzen. Tesla’s vergleichsweise geringe Bestellmenge von 15K lässt vermuten, dass sie entweder sehr effizient trainieren oder alternative Chips nutzen.

Die Liste offenbart ebenfalls, dass einige erwartete Namen wie z.B. Apple nicht auftauchen, und unterstreicht die immense Nachfrage nach GPUs, die weit über die Liefermöglichkeiten hinausgeht.

QUALITÄT DURCH SPEZIALISIERUNG KLEINERER MODELLE

Größe ist nicht alles

Ein weiterer Trend geht zu kleineren, offenen Modellen, die durch Finetuning und Spezialisierung auf eingeschränkte Anwendungsfälle ebenfalls eine hohe Qualität erreichen können. Diese Entwicklung wird durch den Wunsch nach datenschutzkonformen und kosteneffizienten Lösungen angetrieben, da diese Modelle mit vertretbarem Aufwand auch im eigenen Unternehmen betrieben werden können.

AUS DER PRAXIS

VEKTORDATENBANKEN

Wiedergeburt einer alten Technologie

Im Jahr 2023 wurde deutlich, dass der einfache Einsatz von Vektor-Suchtechnologien kein Allheilmittel für die Herausforderungen im Bereich der künstlichen Sprachintelligenz ist. Meine Analyse zeigte, dass eine strategische Kombination von altbewährten und neuen Technologien notwendig ist, um den komplexen Anforderungen gerecht zu werden.

RAG — RETRIEVAL ARGUMENTEN GENERATION

RAG und mehr

RAG-Fusion und Komposition verschiedener Suchtechnologien: Die Kombination von RAG mit verschiedenen Suchtechnologien sowie „HyDE“, einem Ansatz zur Verbesserung von Anfragen, bildet die Basis für einen neuen, effektiveren Ansatz in der Sprachintelligenz. Knowledge Graphen spielen ebenfalls eine zentrale Rolle, indem sie das gesamte Unternehmenswissen strukturieren und für LLMs zugänglich machen.

Finetuning und Anpassung: Das Finetuning und die Anpassung von LLMs an spezifische Anwendungsfälle werden zunehmend wichtiger, um den Output zu verbessern und die Modelle an individuelle Geschäftsbedürfnisse anzupassen.

LOKALE MODELLE UND DATENSCHUTZ

Wird uns Apple überraschen?

Der Bedarf an lokalen Modellen, getrieben von Datenschutzbedenken und Kosteneffizienz, wird weiterhin steigen. Ich erwarte, dass Unternehmen wie Apple in diesem Bereich wichtige Beiträge leisten werden, insbesondere in der Grundlagenforschung.

MEINE PROGNOSE 2024

Wir werden im Jahr 2024 signifikante Entwicklungen in verschiedenen Bereichen der künstlichen Sprachintelligenz erleben.

RETRIEVAL AUGMENTED GENERATION

RAG als essenzielles Werkzeug etabliert

RAG wird sich als essenzielles Werkzeug etablieren, um Unternehmenswissen effektiv mit künstlicher Sprachintelligenz zu verknüpfen. Einfache RAG-Anwendungen sind bereits jetzt durch die direkte Integration mit OpenAI Agenten oder Azure möglich. Die wirkliche Innovation sehe ich jedoch in der Verbindung von symbolischem Wissen und Knowledge Graphen mit RAG, was zu einer noch tieferen und reichhaltigeren Wissensverarbeitung führen wird.

> „Wir werden eine neue Ebene der Informationsverarbeitung und -analyse erreichen!“

Die Integration von LLMs mit strukturierten und externen Datenquellen wird neue Anwendungsfälle ermöglichen, die weit über das bisherige Spektrum hinausgehen. Durch die Kombination von LLMs mit externem Wissen werden wir eine neue Ebene der Informationsverarbeitung und -analyse erreichen.

HARDWAREBEDARF

Hardware-Ressourcen weiterhin ein kritischer Faktor für die Zukunft der künstlichen Sprachintelligenz

Ein weiterer Durchbruch wird im Bereich der lokalen LLMs stattfinden. Die Implementierung von LLMs, die lokal auf Geräten laufen, wird Datenschutzbedenken adressieren und die Kontrolle über sensible Daten zurück in die Hände der Nutzer legen.
Die gegenwärtige Knappheit an Hochleistungs-GPUs kann paradoxerweise auch positive Effekte haben, da sie den Fokus weg von der ständigen Erweiterung der Modelle hin zur Reifung bestehender Verfahren und der Integration in praktische Anwendungen lenkt. Diese Zwangspause in der Hardwarebeschaffung wird den Unternehmen Zeit geben, die bisherigen Fortschritte zu konsolidieren und die Technologie effektiver in ihre Systeme zu integrieren.

Die Einschätzung von Elon Musk, dass für das Training eines hypothetischen GPT-5-Modells 30k-50k H100-GPUs benötigt würden, und eine ähnliche Anzahl für den Betrieb, zeigt, dass die Hardware-Ressourcen ein begrenzender Faktor sind. Dies gilt nicht nur für OpenAI, sondern auch für andere große Technologieunternehmen und Cloud-Anbieter.

EU AI ACT

„Amerika innoviert, Europa reguliert“

Schließlich wird der EU AI Act eine entscheidende Rolle spielen. Die genaue Ausgestaltung dieses Regelwerks im Jahr 2024 wird sich zeigen, aber es ist davon auszugehen, dass es erhebliche Auswirkungen auf die Entwicklung und Anwendung von KI-Technologien in Europa haben wird. Es bleibt spannend zu beobachten, wie Unternehmen auf diese Vorschriften reagieren und wie sie Innovationen vorantreiben, während sie gleichzeitig die Compliance sicherstellen.

„Insgesamt stehen wir an der Schwelle zu einem Jahr, in dem die Kombination von fortschrittlichen neuen und etablierten Technologien neue Wege eröffnen wird. Diese Wege werden nicht nur die Möglichkeiten von LLMs und künstlicher Sprachintelligenz erweitern, sondern auch die Art und Weise, wie Unternehmen und Einzelpersonen von diesen Technologien profitieren können.“

FAZIT

Nach dem Hype

Der erste Hype ist vorbei, und jetzt ist es an der Zeit, mit qualitativ hochwertigen Anwendungsfällen zu überzeugen. Ebenso ist eine realistische Kommunikation über die Fähigkeiten und Grenzen der Technologie von Beratern notwendig. Der Fokus wird darauf liegen, reale Geschäftsprobleme effektiv zu lösen und die Erwartungen an die Technologie realistisch zu halten. Die Qualität und Wirksamkeit von LLMs muss durch solide Anwendungsfälle unter Beweis gestellt werden.

WOLLEN SIE MEHR ERFAHREN?

Über Kai Kramer

Ich bin Kai Kramer, diplomierter KI-Experte der ersten Stunde. Meine Begeisterung für Computer begann 1984 mit meiner ersten Programmierung. Mein Wissen und meine Leidenschaft nutze ich seit meiner Diplomarbeit am Deutschen Forschungszentrum für KI zur Entwicklung von Anwendungen in den Bereichen KI, Sprachanalyse und Big Data. Mein Spezialgebiet sind maßgeschneiderte KI-Lösungen für Steuertexte, technische Dokumente und Sprachassistenten.

2023 gründete ich meine Beratungsfirma, um künstliche Sprachintelligenz in Unternehmen zu verankern. Zudem teile ich mein Wissen in verständlichen Vorträgen und Gastvorlesungen an Hochschulen wie die htw saar.

Mit KITT, meinem KI-Think-Tank, revolutioniere ich Ihr Unternehmen. Von der Ideenfindung bis zur Implementierung - in nur 4 Wochen realisieren wir effektive Sprachintelligenz-Anwendungen.

WOLLEN SIE KÜNSTLICHER SPRACHINTELLIGENZ EFFIZIENT FÜR IHR UNTERNEHMEN NUTZEN?

Wir finden eine maßgeschneiderte Lösung für Ihr Unternehmen.

Buchen Sie ein kostenloses 15-minütiges Erstgespräch

AI-TRAININGS

ChatGPT, Cognitive Services, LLMs, Embeddings, LLama und Co. Ich gebe Ihnen einen Überblick, was es für AI-Verfahren gibt, wie sie anzuwenden sind und was rechtlich zu beachten ist.

- **Inhouse-Schulung**
- **Hands-On für Entwickler**
- **Workshop für Umsetzungsideen**

AI-WORKSHOPS

Wenn Sie nicht wissen, was heute alles mit AI möglich ist, keine Ideen haben, wie sie Ihre Software durch AI aufwerten können. Wir erarbeiten gemeinsam einen Use Case für Ihre Software.

- **Gemeinsame Ausarbeitung eines AI Anwendungsfall**
- **Auswahl der Daten und der Technologie**
- **Umsetzungsstrategie und Planung der Umsetzung**

AI-INTEGRATION

Sie wollen Ihre Software mit AI aufwerten, Ihren Anwendern neue Möglichkeiten zur Verfügung stellen. Ich helfe Ihnen.

- **Umsetzung einer AI Integration und Datenaufarbeitung**
- **Gemeinsam mit ihrem Dev Team Integration in Ihre Anwendung**
- **Schulung und Coaching ihrer Mitarbeiter**

Kai Kramer Consulting GmbH
Kreuzweg 3
67724 Gundersweiler

Vertreten durch:

Kai Oliver Kramer

Kontakt

Telefon: 06361 2559283
Telefax: 06361 2559284
E-Mail: kai@kkc.ai

Redaktionell verantwortlich

Kai Oliver Kramer

Layout & Design

©Daniel Masullo, Berlin
www.masullo.de, Branding, Corporate Publishing

Fotografien

© M&P, 2023

Titel & Image Prompting

©Daniel Masullo, Berlin
www.masullo.de, Branding, Corporate Publishing

Version 2024-02-KKC-LLMR24-1.0
©Kai Kramer Consulting 2024