- openHPI: ChatGPT: Was bedeutet generative KI für unsere Gesellschaft? -

LLM als Puzzleteil

Johannes Hötter
Christian Warmuth

Was passieren kann, wenn man bei ChatGPT nicht aufpasst..

Name	Firma	Titel	Email	Mobilnr
Mickey Mouse	Disney	CEO	mickey.mouse@disney.com	+1 (555) 123-4567
Harry Potter	Hogwarts School	Student	harry.potter@hogwarts.edu	+44 1234 567890
Sherlock Holmes	221B Baker St.	Private Detective	sherlock.holmes@221B.com	+44 207 123 4567
Darth Vader	Galactic Empire	Sith Lord	darth.vader@galacticempire.com	+1 (555) 555-5555

HPI Hasso Plattner Institut
Digital Engineering · Universität Potsdam

KI Service Zentrum HPI
by Hasso-Plattner-Institut

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

LLM als Puzzleteil

Johannes Hötter,
Christian Warmuth

openHPI

Was passieren kann, wenn man bei ChatGPT nicht aufpasst..

Idee: personalisierte Nachricht, um positiv aufzufallen - mit ChatGPT automatisiert erzeugt.

Z.B:

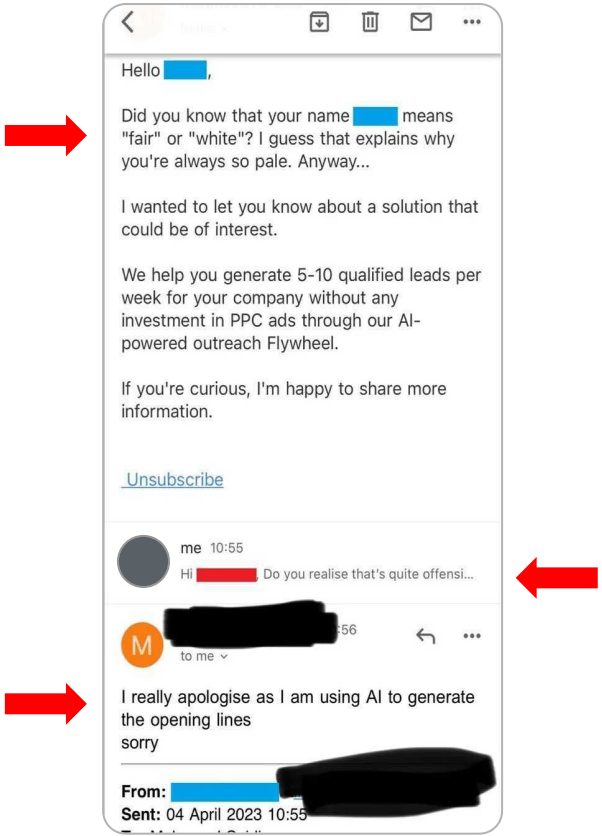
“Sehr geehrter Herr Mouse,

es ist mir eine große Ehre, Ihnen eine Nachricht zu senden! Als einer der bekanntesten und beliebtesten Charaktere in der Geschichte der Animationsfilme haben Sie unzählige Kinder und Erwachsene auf der ganzen Welt begeistert und inspiriert.”

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

Was passieren kann, wenn man bei ChatGPT nicht aufpasst..

Und dann passiert das..



Hasso Plattner Institut
Digital Engineering · Universität Potsdam

KI Service Zentrum
by Hasso-Plattner-Institut

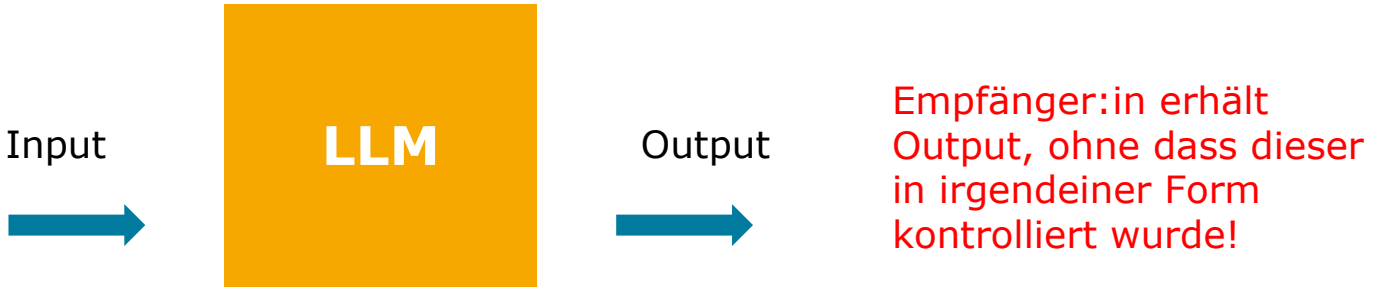
GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

LLM als Puzzleteil

Johannes Hötter,
Christian Warmuth

openHPI

Der Fehler, welcher hier passiert ist:

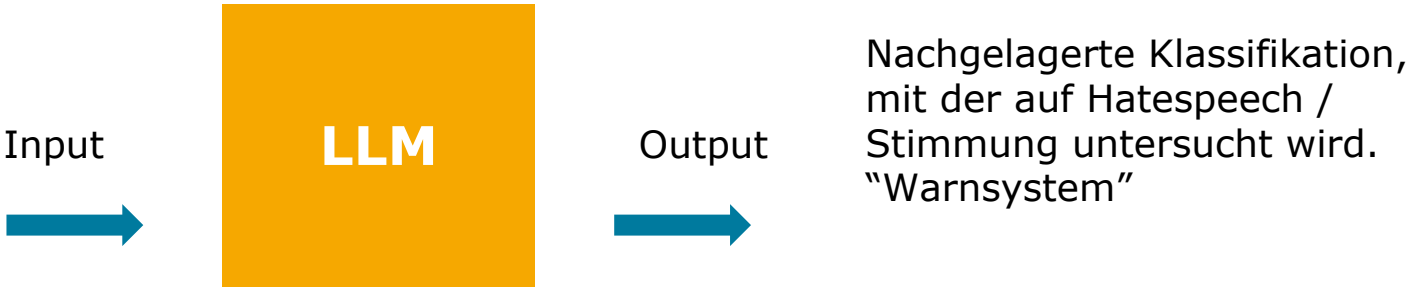


Empfänger:in erhält Output, ohne dass dieser in irgendeiner Form kontrolliert wurde!

Wie kann so etwas verhindert werden?

Entwurf wird von GPT erzeugt, und von einem Menschen kontrolliert.
Schreibt Schreibarbeit, aber erhöht Kontrolle.

Wie kann so etwas verhindert werden?

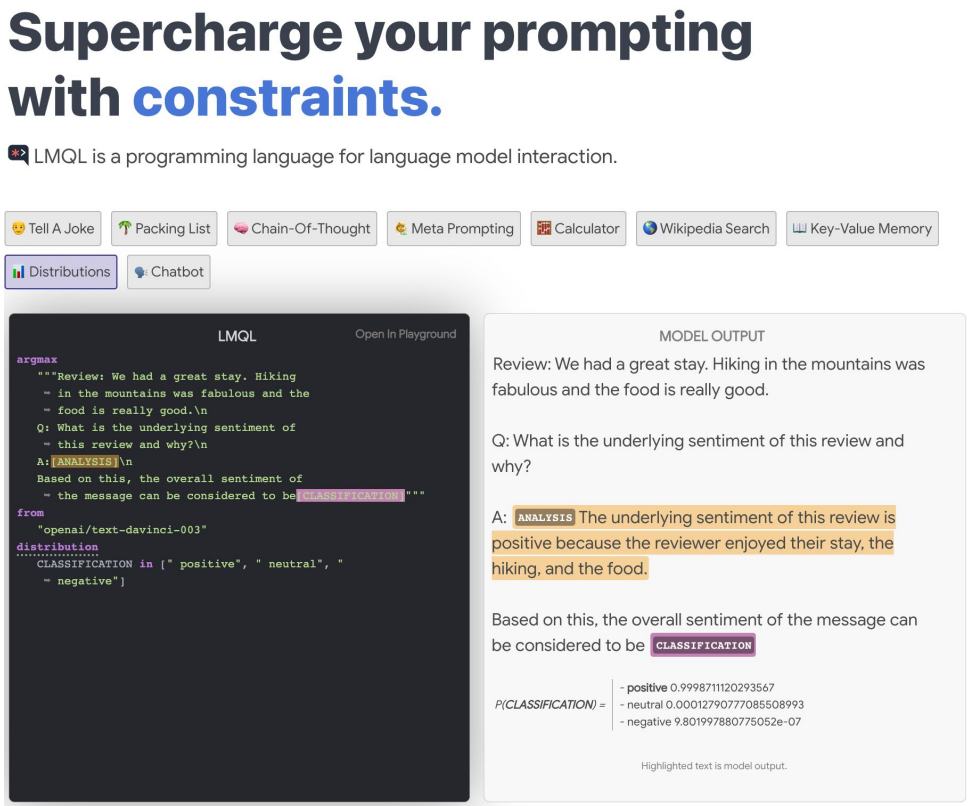


Nachgelagerte Klassifikation, mit der auf Hatespeech / Stimmung untersucht wird. “Warnsystem”

Wie kann so etwas verhindert werden?

Sorgfältiges Benchmarking und Verproben zahlreicher Prompts, z.B. um Situation zu schildern, in dem der Eisbrecher erzeugt werden soll.

Wie kann so etwas verhindert werden?



Quelle: https://lmql.ai/

LLM als Puzzleteil

Johannes Hötter,
Christian Warmuth

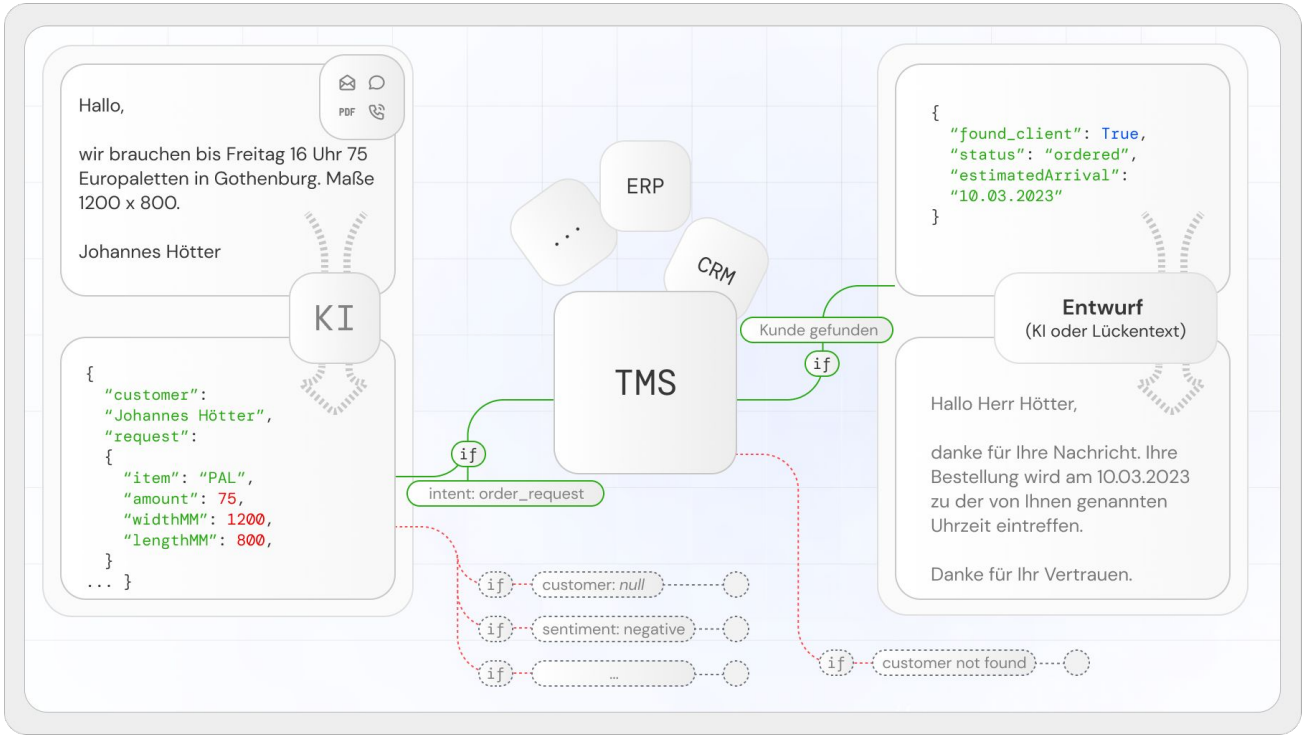
openHPI

Allgemein: Möglichkeiten des Preprocessings

Reichere Input an,
z.B. mit Produktinfos,
Suchergebnissen,
oder IT-System-Infos
(siehe Anwendungen
diese Woche)

HPI Hasso Plattner Institut
Digital Engineering · Universität Potsdam
KI Service Zentrum by Hasso-Plattner-Institut HPI
GEFÖRDERT VOM Bundesministerium für Bildung und Forschung

GPT ist ein Puzzleteil, meist nicht der ganze Prozess



LLM als Puzzleteil

Johannes Hötter,
Christian Warmuth

openHPI

- openHPI: ChatGPT: Was bedeutet generative KI für unsere Gesellschaft? -

LLM als Puzzleteil

Johannes Hötter
Christian Warmuth