Risikomanagement mit KI

Von Prompting zu Context Engineering

AGENDA – 135 MINUTEN

Ankommen & Einordnung 15 min
Begrüßung, Stimmungsbild, Grundlagen

GLANZ30 & Naives Prompting 20 min
Live-Demo: Warum einfach nicht reicht

Perspektiven-Übung 40 min
Gruppenarbeit mit KI + Gallery Walk

Forschung & Systemdenken 30 min

- Benchmark-Ergebnisse
- Iterationsproblem (Demo)
- Live-Demo: project-problems.app

Reflexion & Abschluss 30 min
Tischdiskussion, Q&A, Feedback

WORKSHOP-ZIELE

✓ Am Ende des Workshops können Sie:

- ✓ Grenzen von KI im Risikomanagement einschätzen
- ✓ Perspektivenwechsel bei Risikoanalyse gezielt nutzen
- ✓ Unterschied zwischen transaktionaler und systemischer KI-Nutzung verstehen
- ✓ Verschiedene Reifegrade von KI-Ansätzen bewerten

MATERIALIEN ZUM WORKSHOP

Scannen für Workshop-Materialien

Projektdaten, Beispiel-Prompts, Dokumentation

`share.wb-fernstudium.de`

Block 1

Ankommen & Einordnung

WAS IST RISIKOMANAGEMENT?

Risiko = Zustand potenzieller Non-Compliance

Harte Non-Compliance:

- Gesetze, Vorschriften, Pläne
- Qualität, Zeit, Ressourcen
- Formal definiert, dokumentiert

Weiche Non-Compliance:

- Stakeholder-Erwartungen
- Mentale Modelle, Annahmen
- Oft implizit, nicht dokumentiert

5 Phasen:

1. Identifikation
2. Bewertung
3. Maßnahmenplanung
4. Monitoring
5. Dokumentation

Standards: IPMA ICB 4.0, ISO 31000, PMBOK 7

URSACHEN FÜR RISIKEN – UND WAS KI KANN

⚠ Typische Ursachen

- Fehlende Informationen
- Konfligierende Informationen
- Falsche Annahmen
- Falsche Prioritäten
- Fehlende Entscheidungen

Was kann KI?

Gut bei:

- ✔ Informationsverarbeitung
- ✔ Mustererkennung
- ✔ Schnelles Brainstorming

Schwach bei:

- ✘ Autoritätsbewertung
- ✘ Vertrauensfragen
- ✘ Kontextpersistenz

Block 2

GLANZ30 – Live-Demo

Das Unternehmen:

- **Fahrlässig GmbH**
- Risikoprävention
- 30 Mitarbeiter
- 30-jähriges Jubiläum

Das Projekt:

- Gala-Veranstaltung
- Viele Stakeholder
- Komplexe Anforderungen

Die Stakeholder:

Karl Kopflos	CEO – Reputation
Siggi Sorglos	PM – Lieferung
Gäste	Erlebnis
Caterer	Zahlung, Image

Material:
Projektdokumente, E-Mails, Konzepte, Budgets via QR-Code

LIVE-DEMO: DREI TOOLS, EIN PROMPT

</> Gleicher Prompt an drei Tools:

```
Hier sind die Projektdokumente für GLANZ30.
Welche Risiken gibt es? Erstelle eine Risikomatrix.
```

ChatGPT	**NotebookLM**	**Claude Code**
Web-Chat, Copy&Paste	Dokument-Upload	Dateizugriff lokal

? Vergleichen Sie:

- Welches **Format** liefert jedes Tool? (Tabelle, Fließtext, TSV?)
- Wie **tief** geht die Analyse? Quellenangaben?
- Kann man das Ergebnis **weiterverarbeiten**? (Excel, Projekt-Tool?)

DREI PROBLEME MIT NAIVEM PROMPTING

1. Keine Perspektive

KI weiß nicht, AUS WESSEN SICHT sie analysiert

→ Generische Risiken statt stakeholder-spezifische

2. Kein Format

Jedes Mal andere Struktur, nicht in Excel kopierbar

→ Keine Weiterverarbeitung möglich

3. Keine Integration

Ergebnis lebt nur im Chat, nicht im Projekt-Tool

→ Checkbox-RM statt echtes RM

→ Gleich: **Problem 1 selbst lösen – Problem 2+3 diskutieren**

Block 3

Perspektiven-Übung

WARUM PERSPEKTIVEN WICHTIG SIND

Risiko ist subjektiv –
jeder Stakeholder hat andere Ziele und Risiken

Stakeholder	Hauptziel	Typisches Risiko
CEO Karl Kopflos	Reputation wahren	Reputationsschaden, Skandal
PM Siggi Sorglos	Termin halten	Lieferverzug, Ressourcenmangel
Gäste	Gutes Erlebnis	Schlechte Organisation

GRUPPENARBEIT: PROMPT-OPTIMIERUNG (20 MIN)

3 Gruppen – 3 Perspektiven

Gruppen:

1. **CEO** – Reputation, Kosten
2. **PM** – Zeitplan, Ressourcen
3. **Gäste** – Erlebnis, Qualität

20 Min Arbeit
Prompt-Vorlage im Handout

Aufgabe:

- Projektdaten laden (QR-Code)
- Perspektive mit KI einnehmen
- Top-5-Risiken identifizieren
- **Wie könnte man den Prompt optimieren?**
- Top 3 + Prompt-Erkenntnisse auf Flipchart

Wir nehmen Perspektiven ein, **weil** Stakeholder eine Rolle spielen – und wir auf ihre Kompetenzen zählen.

BEISPIEL: CEO-PERSPEKTIVE

</> Prompt-Vorlage

```
Du bist Karl Kopflos, CEO der Fahrlässig GmbH.
Dir ist wichtig:
- Reputation des Unternehmens
- Kosten im Rahmen
- Keine negative Presse
Analysiere die Projektinformationen und
identifiziere die Top 5 Risiken aus DEINER Sicht.
Bewerte nach Wahrscheinlichkeit und
Auswirkung auf DEINE Ziele.
```

Passen Sie den Prompt für Ihre Gruppe an!

🖼 Ergebnisse teilen

Ablauf:

1. Flipcharts aufstellen
2. Jede Gruppe präsentiert Top 3 (**je 2 Min**)
3. Gemeinsame Diskussion:
   - Was **bedeutet** es, eine Perspektive einzunehmen?
   - Welche **impliziten Erwartungen** stecken darin?
     Kompetenz, Motivation, Handlungsbereitschaft, Ressourcenverfügbarkeit …
   - Welche Informationen fehlen, die **nur Menschen** kennen?
   - Wo **überlappen** oder **konfligieren** Perspektiven?

PERSPEKTIVEN SIND EIN HEBEL – ABER NICHT DER EINZIGE

✔ Was wir erreicht haben:

- Differenzierte Risikolisten
- Stakeholder-spezifische Bewertung
- Tiefere Analyse als naiv

✖ Was noch fehlt:

- Konsistentes Ausgabeformat
- Integration in Projekt-Tools
- Persistenz über Zeit
- Mitigations-Tracking

Praxis-Realität

Custom GPT: Konsistentes TSV-Format erzwingbar, aber keine Datei-Persistenz

NotebookLM: Gute Analyse, keine Export-/Updatefähigkeit

Claude Code: Dateien schreiben möglich, keine Kollaboration per se

Copilot/Plugins: Für Unternehmen mit Lizenz am sinnvollsten für Analyse, aber keine KI-spez. Risikoanalyse

Kernproblem:
Es mangelt an Integration – nicht an KI-Fähigkeit

Block 4

Forschung & Systemdenken

FORSCHUNG: RISIKOTYPEN IM BENCHMARK

IES-Framework: Information → Events → States

Risikotyp	Was die KI tun muss	Schwere
Benannt	Explizit im Text erkennen	Einfach
Komplex	≥3 Quellen kombinieren	Schwer
Umfeld	Externes Wissen einbeziehen	Schwer
Mitigiert	**Zustandsänderung tracken**	**Sehr schwer**

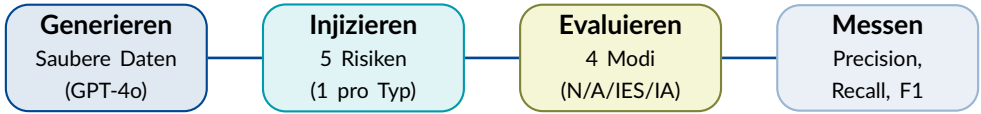
Mitigierte Risiken = ebenso schwer wie Deduplikation. Beide erfordern: gleiche Info in verschiedenen Zuständen wiedererkennen

Benchmark-Studien

Wild (2025), IPMA RC
Wild (2026), BAM Conference

Synthetische Daten, versteckte Risiken, naiv vs. agentisch vs. IES-Pipeline

BENCHMARK-DESIGN: SYNTHETISCHE PROJEKTDATEN



Parameter	Wert
Informationsvolumen	500 Einheiten
Risiken pro Projekt	5 (eines pro Typ)
Modelle	GPT-4o, Mistral Med.
Modi	Naiv, Agentisch, IES, IES-Ag.

Design-Entscheidungen:

- Alle Daten synthetisch (kontrolliert)
- Saubere Basislinie = explizit risikofrei
- Matching durch separaten LLM-Aufruf
- Alle Schritte reproduzierbar (Python)

KEY FINDINGS: NAIV VS. AGENTISCH

- ✔ ≈**100% Recall** – beide Ansätze finden genannte Risiken
- ✘ **Naiv: 50–150 Extra-Risiken** die nirgends in den Daten stehen
- ✔ **Agentisch: 10× präziser**, teilweise null Extra-Risiken

Experten bestätigen Extra-Risiken als real (κ=1.0, Ernst 2026). Aber: 50–150 Risiken → **kognitive Überlastung**. Gut fürs Brainstorming, schlecht im laufenden Projekt

⚠ Alignment-Problem

GPT-3.5 bewertet **eigene** Antworten: Median 5.0

GPT-3.5 bewertet **PMBOK-RAG**-Antworten: Median 3.5

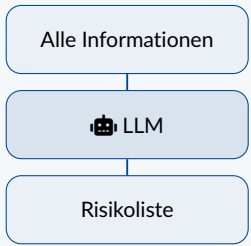
→ **KI bevorzugt eigene Outputs gegenüber wissensbasierten Antworten!**

Wild (2024), Gastvorlesung Univ. Southampton

? Was bedeutet das für Ihre Praxis?

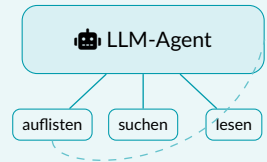
BENCHMARK-ANSÄTZE: NAIV VS. AGENTISCH

Naiv (Single Call)



- Ein Aufruf, alle Daten auf einmal
- Analogie: **Brainstorming-Sitzung**
- Kompression bei großen Datenmengen

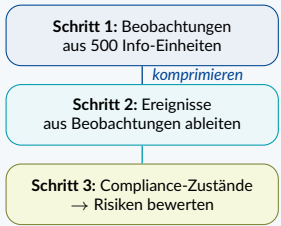
Agentisch (Iterativ)



- Bis zu 5 Reasoning-Schritte
- Analogie: **Laufendes Monitoring**
- Iterative Exploration der Daten

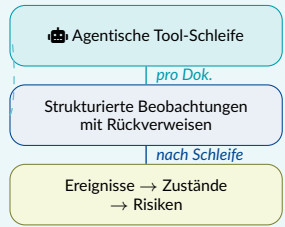
IES-ANSÄTZE: PIPELINE VS. AGENTISCH-HYBRID

IES-Pipeline (3-Pass)



- Operationalisiert IES direkt
- Perfekte Präzision, reduzierter Recall
- Kompressions-Engpass Schritt 1→2

IES-Agentisch (Hybrid)



- Keine verlustbehaftete Kompression
- Beobachtungen steuern Exploration
- Bis zu 10 Reasoning-Schritte

ERGEBNISSE: 500 INFORMATIONSEINHEITEN, 5 RISIKEN

Modell	Modus	Recall	Präzision	F1	Extra-R.	Laufzeit
GPT-4o	Naiv	1.0	1.0	1.0	0	121 s
GPT-4o	Agentisch	1.0	1.0	1.0	0	42 s
GPT-4o	IES-Pipeline	0.6	1.0	0.75	1	375 s
GPT-4o	IES-Agentisch	0.4	1.0	0.57	1	60 s
Mistral Med.	Naiv	1.0	1.0	1.0	113	63 s
Mistral Med.	Agentisch	1.0	1.0	1.0	5	63 s
Mistral Med.	IES-Pipeline	0.8	1.0	0.89	5	692 s
Mistral Med.	IES-Agentisch	1.0	1.0	1.0	10	29 s

Mistral IES-Agentisch: **perfekter Recall**, **Präzision 1.0**, **10× weniger Fehlalarme**, **schnellste Laufzeit** (29 s).
IES-Pipeline am langsamsten (375–692 s) durch Batch-Verarbeitung. **IES-Agentisch = Pareto-optimal**.

DISKUSSION: STAKEHOLDER ALS WISSENSGRAPH?

Das IES-Modell braucht mehr als Projektdaten:

- Informationen über **Stakeholder selbst**
- Kompetenzen, Motivationen, Handlungsmuster
- Beziehungen und Abhängigkeiten
- Implizites Wissen („Kaffeeküche”)

? Leitfragen

1. Wie bilden wir Stakeholder und ihre Eigenschaften **sinnvoll** ab?
2. Geht das **zu weit**? Welche Daten sind ethisch vertretbar zu sammeln?
3. Wie steht es um den **Datenschutz** bei der KI-Verarbeitung?
4. Würden Stakeholder **zustimmen**, so modelliert zu werden?

Kurzdiskussion (5 Min)
An Ihren Tischen

LIVE-DEMO: PROJECT-PROBLEMS.APP

app.project-problems.app

So sieht systemisch aus:

- Multi-Agenten-Workflow
- Automatische Perspektiven
- Semantische Deduplikation
- Persistentes Tracking

Public Beta
Kostenlos ausprobieren

TOOLS IN DER PRAXIS: WAS GEHT, WAS NICHT?

Tool	Format	Persistenz	Kollaboration
ChatGPT/Claude	✗	✗	✗
Custom GPT	✓ TSV	✗	✗
NotebookLM	✗	✗ read-only	✗
Claude Code	✓ Dateien	✓	✗ technisch
Copilot/Plugin	✓	✓	∼
Vollsystem	✓	✓	✓

Das Kernproblem ist Integration, nicht KI-Fähigkeit. Selbst bei voller Tool-Integration: Die Kaffeeküchen-Gespräche werden nicht automatisch zum Projektinhalt.

24% PM-Nutzung, 58% berichten von Integrationsproblemen *(GPM-Studie 2024)*

Block 5

Reflexion & Abschluss

TISCHDISKUSSION: WAS BRAUCHT EIN GUTES SYSTEM?

Diskussion an den Tischen (10 Min)

? Leitfragen

1. Wie kommen die **Kaffeeküchen-Gespräche** in die Risikoanalyse?
2. Wann ist ein Risiko **mitigiert** – und wer entscheidet das?
3. Wie verhindert man **Checkbox-Risikomanagement** mit KI?
4. Was müsste Ihr **PM-Tool** können, damit KI-RM funktioniert?

Jeder Tisch sammelt **2–3 Kernpunkte** für das Plenum

DIE WICHTIGSTEN ERKENNTNISSE

Take-Aways

1. **KI findet Risiken – aber vertraut ihr nicht blind**
   KI bevorzugt eigene Antworten gegenüber wissensbasierten (Alignment-Problem)
2. **Perspektiven helfen – Integration fehlt**
   Das Hauptproblem ist nicht die KI-Fähigkeit, sondern die Einbettung in Workflows
3. **Mitigation erkennen ist so schwer wie Deduplizieren**
   Zustandsänderungen tracken erfordert Persistenz und Architektur
4. **Agentische Ansätze schlagen naive dramatisch**
   10$\times$ weniger Halluzinationen, aber: Architektur + Modell müssen zusammenpassen
5. **Checkbox-RM ist die größte Gefahr**
   KI-generierte Risikoliste abhaken $\neq$ Risikomanagement

MATERIALIEN ZUM MITNEHMEN

Alle Materialien

`share.wb-fernstudium.de`

Enthalten:

- PDF-Zusammenfassung
- Beispiel-Prompts (9 Vorlagen)
- Projektdaten GLANZ30
- Benchmark-Paper (IPMA RC 2025)

Weiterführend:

- project-problems.app (Beta)
- GPM-Studie 2024

IHR FEEDBACK

💬 Start · Stop · Continue

START

Was sollten wir nächstes Mal hinzufügen?

STOP

Was war zu viel oder nicht hilfreich?

CONTINUE

Was sollen wir unbedingt beibehalten?

Vielen Dank für Ihre Teilnahme!

Prof. Dr. Helge Wild

h.wild@gpm-ipma.de

linkedin.com/in/prof-dr-helge-frank-wild-8aab6551

project-problems.app

Beispiel-Prompts für Risikomanagement mit KI

GPM Focus – KI in Projekten anwenden
Prof. Dr. Helge Wild

24. März 2026

Hinweis zur Nutzung

Diese Prompts sind Ausgangspunkte. Passen Sie sie an Ihr Projekt an. Je präziser der Kontext, desto besser die Ergebnisse. **Achtung:** KI-generierte Risikolisten ersetzen kein Risikomanagement – sie unterstützen es.

1. Naiver Ansatz – Erste Risikoidentifikation

Prompt 1: Einfache Risikoanalyse

```
Analysiere die folgenden Projektinformationen und
identifiziere potenzielle Risiken:

[HIER PROJEKTINFORMATIONEN EINFUEGEN]

Erstelle eine Risikomatrix mit folgenden Spalten:
- Risikobeschreibung
- Eintrittswahrscheinlichkeit (niedrig/mittel/hoch)
- Auswirkung (niedrig/mittel/hoch)
- Kategorie (technisch/organisatorisch/extern)
- Prioritaet (niedrig/mittel/hoch/kritisch)
```

Grenzen dieses Ansatzes

- Generische Ergebnisse („Budget könnte überschritten werden")
- Keine Perspektivendifferenzierung – KI weiß nicht, FÜR WEN sie analysiert
- Jedes Mal anderes Format – nicht in Excel kopierbar
- 50–150 „Extra-Risiken" die nicht aus den Daten stammen (Benchmark-Ergebnis)

2. Perspektivisches Prompting

2.1 CEO/Sponsor-Perspektive

Prompt 2: CEO-Sicht

```
Du bist [NAME], [POSITION] bei [UNTERNEHMEN].

Deine Hauptziele fuer dieses Projekt sind:
- [ZIEL 1, z.B. Reputation schuetzen]
- [ZIEL 2, z.B. Budget einhalten]
- [ZIEL 3, z.B. strategische Ausrichtung]

Dir ist besonders wichtig:
- Keine negativen Auswirkungen auf das Unternehmensimage
- Einhaltung aller Compliance-Vorgaben (hart UND weich)
- Rechtzeitige Information bei kritischen Entwicklungen

Analysiere die folgenden Projektinformationen aus
DEINER Perspektive:

[PROJEKTINFORMATIONEN]

Identifiziere die Top 5 Risiken, die DEINE Ziele
gefaehrden koennten. Bewerte sie nach:
- Eintrittswahrscheinlichkeit
- Auswirkung auf DEINE spezifischen Ziele
- Dringlichkeit der Massnahmen

Priorisiere die Risiken nach ihrer Bedeutung fuer DICH.
```

2.2 Projektmanager-Perspektive

Prompt 3: PM-Sicht

```
Du bist [NAME], verantwortlicher Projektmanager fuer
[PROJEKTNAME].

Deine Hauptverantwortung (harte Compliance):
- Termintreue und Meilenstein-Erreichung
- Budget einhalten
- Qualitaetssicherung

Deine zusaetzlichen Ziele (weiche Compliance):
- Gute Teamstimmung erhalten
- Stakeholder zufriedenstellen
- Konflikte frueh erkennen

Rahmenbedingungen:
- Deadline: [DATUM]
- Team: [TEAMGROESSE]
- Budget: [BUDGETRAHMEN]

Analysiere die Projektinformationen aus operativer Sicht:

[PROJEKTINFORMATIONEN]

Identifiziere Risiken fuer deine Projektdurchfuehrung.
Schlage fuer jedes Risiko Massnahmen vor.
```

2.3 Externe Stakeholder-Perspektive

Prompt 4: Externer Stakeholder

```
Du bist [NAME/ROLLE], ein externer Stakeholder
fuer [PROJEKTNAME].

Deine Erwartungen an das Projekt:
- [ERWARTUNG 1]
- [ERWARTUNG 2]

Deine Abhaengigkeiten:
- [Z.B. "Mein Geschaeft haengt vom Projekterfolg ab"]

Analysiere die Projektinformationen aus DEINER Sicht:

[PROJEKTINFORMATIONEN]

Identifiziere Risiken, die:
1. Deine Erwartungen enttaeuschen koennten
2. Negative Auswirkungen auf dich haben
3. Konflikte zwischen deinen und anderen Zielen erzeugen
```

3. Format-Kontrolle: TSV für Excel-Kompatibilität

Warum TSV?

Tab-separierte Werte (TSV) können direkt in Excel/Google Sheets eingefügt werden. Fügen Sie diesen Block an **jeden** Prompt an, um ein konsistentes, weiterverarbeitbares Format zu erhalten.

Prompt 5: Format-Suffix (an jeden Prompt anhängen)

```
WICHTIG - Ausgabeformat:
Erstelle die Risikomatrix als TSV (Tab-separiert).
Erste Zeile = Header. Exakt diese Spalten:

ID  Risiko  Perspektive  Wahrscheinlichkeit  Auswirkung
Prioritaet  Kategorie  Massnahme  Status

Werte fuer Wahrscheinlichkeit/Auswirkung: niedrig/mittel/hoch
Werte fuer Prioritaet: niedrig/mittel/hoch/kritisch
Werte fuer Status: offen/in_bearbeitung/mitigiert/akzeptiert

Keine Markdown-Tabellen, keine Aufzaehlungen.
Nur reines TSV, das ich direkt in Excel einfuegen kann.
```

4. Iteratives Risiko-Update

Prompt 6: Risikomatrix aktualisieren

```
Hier ist die aktuelle Risikomatrix (TSV-Format):

[AKTUELLE RISIKOMATRIX EINFUEGEN]

Neue Projektinformationen:

[NEUE INFORMATIONEN]

Aufgaben:
1. Pruefe, welche bestehenden Risiken sich veraendert haben
2. Identifiziere neue Risiken aus den neuen Informationen
3. Pruefe, ob Risiken durch die neuen Infos MITIGIERT wurden
   - Aendere Status auf "mitigiert" mit Begruendung
4. Erstelle eine aktualisierte Risikomatrix im EXAKT
   GLEICHEN TSV-Format

Markiere Veraenderungen:
- [NEU] fuer neue Risiken
- [GEAENDERT] fuer veraenderte Bewertungen
- [MITIGIERT] fuer behobene Risiken

KEINE Duplikate erzeugen. Wenn ein neues Risiko einem
bestehenden aehnelt, aktualisiere das bestehende.
```

Bekanntes Problem: Mitigation

KI erkennt mitigierte Risiken schlecht. Die gleiche Information in verschiedenen Zuständen wiederzuerkennen („war ein Risiko, ist jetzt gelöst") ist laut Benchmark **genauso schwer wie Deduplikation**. Prüfen Sie Mitigations-Bewertungen der KI immer manuell!

5. Deduplikation & Konsolidierung

Prompt 7: Mehrere Perspektiven zusammenführen

```
Ich habe Risikoanalysen aus verschiedenen Perspektiven:

ANALYSE 1 (CEO-Perspektive):
[RISIKOLISTE 1 ALS TSV]

ANALYSE 2 (PM-Perspektive):
[RISIKOLISTE 2 ALS TSV]

ANALYSE 3 (Stakeholder-Perspektive):
[RISIKOLISTE 3 ALS TSV]

Aufgaben:
1. Identifiziere semantisch aehnliche Risiken ueber Listen
2. Konsolidiere zu einem gemeinsamen Risiko
3. Notiere bei jedem konsolidierten Risiko:
   - Welche Perspektiven betroffen sind
   - Unterschiede in der Bewertung je Perspektive
   - Potenzielle Zielkonflikte
4. Behalte perspektivenspezifische Risiken separat
5. Erstelle konsolidierte Master-Matrix als TSV

Ausgabe als TSV mit Spalten:
ID  Risiko  Betroffene_Perspektiven  Wahrscheinlichkeit
Auswirkung  Prioritaet  Zielkonflikte  Owner_Rolle
```

6. Custom GPT: Konsistenter Risiko-Assistent

Custom GPT / Agent

Dieser System-Prompt erzwingt konsistente Ausgaben über mehrere Sitzungen hinweg. Einsetzbar als Custom GPT (OpenAI), Agent (Mistral), oder Project (Claude).

Prompt 8: System-Prompt für Custom GPT

```
# Rolle
Du bist ein Risikomanagement-Assistent fuer Projekte.

# Standards
IPMA ICB 4.0, ISO 31000, PMBOK 7

# Risikodefinition
Risiko = Zustand potenzieller Non-Compliance
- Harte NC: Gesetze, Plaene, Qualitaet, Zeit, Budget
- Weiche NC: Stakeholder-Erwartungen, Annahmen

# Ausgabeformat - IMMER TSV
Jede Risikomatrix als Tab-separierte Werte:
ID  Risiko  Perspektive  Wahrscheinlichkeit  Auswirkung
Prioritaet  Kategorie  Massnahme  Status

# Regeln
1. NUR Risiken aus den gegebenen Projektinformationen
2. Fuer jedes Risiko: Quelle angeben
3. Keine generischen "koennte"-Risiken
4. Bei Updates: EXAKT gleiches Format beibehalten
5. Duplikate erkennen und zusammenfuehren
6. Mitigierte Risiken explizit als solche markieren
7. Bei Unklarheiten: NACHFRAGEN statt annehmen

# Grenzen (transparent kommunizieren)
- Ich validiere keine Autoritaeten
- Ich treffe keine Priorisierungsentscheidungen
- Ich bevorzuge eigene Antworten gegenueber wissensbasierten
  (Alignment-Problem: eigene Median 5.0 vs. RAG Median 3.5)
- Menschliche Validierung ist IMMER notwendig
```

7. Prompt-Engineering: Best Practices

6 Regeln für bessere Prompts

1. Kontext geben – Projektgröße, Branche, Rahmenbedingungen, Besonderheiten
2. Rolle definieren – „Du bist [Rolle] mit [Verantwortung]" + Ziele + Einschränkungen
3. Format fixieren – TSV, gleiche Spalten, gleiche Werte. „Exakt dieses Format verwenden"
4. Iterativ arbeiten – Erste Version als Basis, schrittweise verfeinern, bisherige Version mitgeben
5. Quellen einfordern – „Begründe jedes Risiko mit Bezug zur Quellinformation"
6. Nicht blind vertrauen – KI bevorzugt eigene Outputs gegenüber wissensbasierten. Immer manuell validieren

8. Häufige Probleme & Lösungen

⚠ Problem 1: Zu generische Risiken

Lösung: Prompt ergänzen mit:

```
Identifiziere nur Risiken, die sich aus den konkreten
Projektinformationen ableiten lassen. Nenne fuer jedes
Risiko die Quellinformation, aus der es abgeleitet wurde.
Keine allgemeinen Projektrisiken ohne konkreten Bezug.
```

⚠ Problem 2: Format ändert sich bei Updates

Lösung: TSV-Format fixieren + bisherige Matrix mitgeben:

```
WICHTIG: Verwende EXAKT das gleiche TSV-Format wie die
bisherige Matrix. Gleiche Spalten, gleiche Reihenfolge,
gleiche Kategoriewerte. Nur Inhalte aktualisieren.
```

⚠ Problem 3: Duplikate bei mehreren Analysen

Lösung: Expliziten Deduplikations-Schritt einbauen (Prompt 7). Nicht darauf vertrauen, dass die KI das automatisch erkennt.

⚠ Problem 4: KI markiert Risiken fälschlich als mitigiert

Lösung: Mitigation-Status immer manuell prüfen. Die KI erkennt Zustandsänderungen schlecht – „war ein Risiko, ist jetzt gelöst" ist eine der schwierigsten Aufgaben im Benchmark.

9. Forschungsergebnisse auf einen Blick

Benchmark: Wild (2025/2026)

Vier Benchmark-Modi im Vergleich:

Modus	Funktionsweise	Stärken/Schwächen
Naiv (Single Call)	Alle Daten auf einmal, ein LLM-Aufruf	50–150 Extra-Risiken, ≈100% Recall, 7–15% Präzision
Agentisch (Iterativ)	Bis zu 5 Tool-Aufrufe, schrittweise Exploration	10× präziser, teils null Extra-Risiken
IES-Pipeline (3-Pass)	Information → Events → States in 3 Durchläufen	Perfekte Präzision, aber Kompressions-Engpass
IES-Agentisch (Hybrid)	Agentische Schleife + IES-Struktur	Keine Kompression nötig, bis zu 10 Schritte

Extra-Risiken (Ernst 2026): Experten bestätigen sie als real (κ=1.0). **Aber:** 50–150 Extra-Risiken = kognitive Überlastung. Nützlich als Brainstorming am Projektstart, problematisch im laufenden Projekt.

Alignment-Problem (Wild 2024, Gastvorlesung Univ. Southampton): GPT-3.5 bewertet eigene Antworten mit Median 5.0, aber PMBOK-RAG-basierte Antworten mit Median 3.5. → KI bevorzugt eigene Outputs gegenüber wissensbasierten Antworten. Menschliche Validierung immer nötig.

Materialien: `share.wb-fernstudium.de`

Tool ausprobieren: `app.project-problems.app`

Kontakt: h.wild@gpm-ipma.de

INTEGRATING LLMS INTO PROJECT RISK MANAGEMENT: A BENCHMARK PROPOSAL

Helge Frank WILD, Wilhelm Büchner Hochschule, 64297 Darmstadt, Germany, helge.wild@wb-fernstudium.de

Abstract: This paper proposes a benchmark for assessing AI usage in project risk management. I demonstrate the gap between computer scientists' approaches to evaluating and what the field of project management research has yet come up with. I characterize project risk management and therefrom deduce a benchmark model, encompassing data, tasks and metrics. I propose data structures and synthetic data generation approaches. Preliminary results indicate that LLMs can achieve high accuracy but struggle with precision on given tasks. More research on benchmark approaches is needed to shape the future of AI-supported risk management on assured quality levels.

Keywords: Risk Management; Artificial Intelligence; Benchmark

1. Introduction

Risk management is a cornerstone of effective project management, often determining the success or failure of a project. In practice, project management offices or project managers themselves act in the role of project risk managers and are responsible for identifying, evaluating, monitoring, and controlling risks. Project risk managers rely on information contained within the project and continuously compare new pieces of information against previously set objectives to determine whether a recent development might lead to detrimental effects on the project's outcome if uncontrolled.

Project risk managers often use lists that contain information on potential or substantial risks as well as issues or problems, depending on whether their effects lie in the future or have already been incurred. A method of choice is often a risk matrix or risk register that requires permanent and effort-intensive updating. As positive effects are often only seen in hindsight, continuously managing risks is often de-prioritized against other operational tasks in day-to-day project work.

LLM-based systems could significantly reduce the effort required for managing risks in projects. These systems automate the continuous monitoring and evaluation of new information against project objectives, identifying potential risks and potentially suggesting mitigation strategies. By leveraging advanced reasoning and context-aware capabilities, they could enable project managers to focus on strategic decision-making while maintaining comprehensive risk coverage. As mentioned by Geraldi et al. (2024): "Research on AI can, for example, investigate how AI can be used for creating risk registers and improving project scheduling or decision-making. These are relevant topics as stated in the recent PMI report on AI (Geraldi et al., 2024):

Corresponding author: Helge Frank WILD, helge.wild@wb-fernstudium.de;
Copyright © 2025, The Authors. Published by IPMA.
This is an open access article under the CC BY-NC 4.0 license
(https://creativecommons.org/licenses/by-nc/4.0/)
https://doi.org/10.56889/ivwj9650
PAGE 261-273

'This is important evidence supporting the need to experiment more to understand the full potential of this technology in project management.' (p. 8)."

In the last three years, no article in the major International Journal of Project Management has explicitly discussed the interface of risk management and artificial intelligence. In major Project Management Journal, recently only Müller et al. (2024) have authoritatively discussed artificial intelligence and project management in general, in a PMJ thoughtlet. They have discussed various opportunities and challenges associated with research on AI.

These calls have not been followed extensively, however.

Currently, it remains unclear how effectively AI can be integrated into project management tasks to maximize its potential benefits. In a recent study conducted with a colleague, we surveyed 176 project managers to assess the application of AI across various project management tasks and its perceived impact on project success (published in non-peer reviewed journal; peer-reviewed issue forthcoming). Interestingly, the results revealed a notable trend: the more extensively AI was employed in risk management, the greater its perceived benefits. This positive correlation was particularly strong, second only to the use of AI in enhancing project communication. These findings suggest that risk management and communication may represent key areas where AI adoption can deliver significant value, reinforcing the need for focused research and tailored AI solutions in these domains.

LLM-based risk management remains an unsolved problem for several reasons, rooted in both the inherent complexity of the task and the limitations of current approaches to leveraging Large Language Models (LLMs). While advancements in LLM capabilities, such as multi-step reasoning, retrieval-augmented generation (RAG), and benchmark-driven improvements, have shown promise in addressing structured and well-defined tasks, the dynamic and context-specific nature of project risk management presents unique challenges that remain unaddressed. Current benchmarks used to evaluate LLMs, such as those focusing on mathematical reasoning or truthfulness, are illuminating in their respective domains but not directly translatable to project risk analysis. For example, benchmarks like TruthfulQA (Lin et al., 2022) aim to identify societal misconceptions in generated responses—a task that, while valuable, is fundamentally different from uncovering and addressing erroneous assumptions or knowledge gaps among project stakeholders. In project management, risks often emerge from misalignments or gaps in understanding between stakeholders, which are compounded by incomplete or evolving information. This context-dependent complexity cannot be fully replicated by static benchmarks. Some further problems with currently available benchmarks exist, which will be discussed below.

In summary, LLM-based risk management remains unsolved because it demands capabilities that go beyond current benchmarks and methodologies. Effective risk management in projects requires dynamic, context-sensitive reasoning; the ability to revise prior conclusions in light of new information; and a nuanced understanding of human factors and stakeholder relationships. Despite the challenges, it is evident that LLMs have advanced rapidly and continue to evolve at an impressive pace. Base model performance steadily improves, and the effectiveness of these models can be significantly influenced by end users through skillful prompting or the application of carefully designed system prompts. Beyond base models, new developments such as Large Action Models (LAMs), which extend LLMs with actionable capabilities, and

advanced models incorporating internalized reasoning techniques—such as Chain-of-Thought (CoT) prompting seen in the OpenAI o1 model series—are pushing the boundaries of what LLMs can achieve. Additionally, wrapper solutions that enhance functionality and the integration of tool use have further expanded the range of tasks LLMs can perform, particularly in areas where they have traditionally struggled, such as precise calculations or accessing real-time data.

These advancements are blurring the line between deep neural network-based AI models and fully functional business applications that rely on these models as their core technology. This shift significantly broadens the potential applications of LLM-based solutions, making them more accessible and adaptable to real-world business needs. This evolution also complicates the decision-making process for businesses. Organizations face increasing difficulty in determining the most suitable solutions for their specific challenges and in evaluating the business cases for investing in rapidly advancing LLM technologies. As the distinction between AI models and their practical applications diminishes, the need for informed, strategic decisions about technology adoption becomes all the more critical.

What we need in this situation is a testbed for risk management in projects. It should be positioning itself towards the mentioned perspectives on risk management scope and processes. It should also be largely technology/approach-independent while stil focusing on LLM/AI-based approaches.

The primary contribution of this paper is the design of a benchmark framework for evaluating and advancing LLM-based approaches to project risk management. This benchmark framework addresses critical gaps in current research by providing a structured methodology to simulate, test, and measure the effectiveness of LLMs in project environments. Unlike existing benchmarks, which often rely on static datasets or narrowly scoped tasks, the proposed framework emphasizes the complexities of real-world project scenarios, including continuous integration of new information, context-sensitive reasoning, and the reconciliation of conflicting stakeholder objectives. By focusing on the challenges and opportunities specific to LLM-based solutions, whether as part of a multi-agent, RAG-based system or as standalone large-context model solution, this benchmark framework lays the foundation for developing and refining methods that are better equipped to address the nuanced and evolving demands of project risk management.

2. Project risk management

Project risk management is a critical aspect of project management, involving identifying risks, assessing their impact and likelihood, and planning their elimination, reduction, or mitigation (Levene & Lewis, 2015). Some definitions emphasize risk as an uncertain event that can have positive or negative effects on project objectives (impact risks), while others consider risks that are bound to materialize, but their extent of deviation is their more important aspect. Williams et al. (2006) distinguish predictable, known and unknown risks: predictable risks are known to be stochastic events, while organizations do or do not know that they will face known, respectively unknown risks. Some schools of though consider risks as objective, i.e. as a measurable and quantifiable fact. Others consider risks as subjective constructions influenced by perceptions and interpretations (Zhang, 2011). Furthermore, risks can be categorized in a multitude of categories, stemming from the perspective or regime that they can best be

explained out of or root in: organizational, processual, business, design, financial, reputational, compliance, …

Various international standards such as PMI, PRINCE2, IPMA, and ISO 31000 provide frameworks for project risk management, each with its own methodologies and processes (Rehácek, 2017). Project risk management is often perceived as adding value to projects, though this value is subjective and depends on stakeholder perceptions and the specific context of the project (Willumsen et al., 2019).

While there is no single, universally accepted definition of project risk management, it is generally understood as a process of managing uncertainties that could affect project outcomes and objectives. Different schools of thought and standards offer varied perspectives and methodologies, reflecting the complexity and context-specific nature of risk management in projects. Just like outputs of LLMs, risk management is far from exact. Unlike LLMs, it is not part of benchmarking yet.

For the benchmark envisioned in this paper, focus will be on risk identification and assessment: this involves recognizing potential risks that could affect the project or its objectives. It is crucial to identify risks early in the project life cycle to mitigate their impact effectively. Once risks are identified, they need to be analyzed to understand their potential impact and likelihood. This process helps prioritize risks based on their severity and probability of occurrence (Al-Ajmi & Makinde, 2018; Fang et al., 2017; Levene & Lewis, 2015; Perry, 1986; Thamhain, 2013).

While this benchmarks focus will not be on risk response planning, it considers risk monitoring and control, i.e. continuous observation of risks throughout the project to detect their ongoing existence or resoltuion. (Thamhain, 2013).

3. Benchmarks for LLMs

There is an entire universe for LLM benchmarking online, as well as LLM leaderboards in online repositories like Huggingface or competition platforms lie Kaggle.

The Huggingface Open LLM Leaderboard e.g. evaluates language models using a diverse set of benchmarks tailored to different tasks and domains. IFEval measures a model's ability to follow explicit formatting instructions, focusing on strict compliance rather than content quality (Zhou et al., 2023). BBH (Big Bench Hard, Srivastava et al., 2023) assesses performance on 23 highly challenging tasks, including arithmetic, reasoning, and language understanding, providing insights into human-aligned capabilities. MATH focuses on high-school competition-level problems, emphasizing precise solutions to complex algebra, geometry, and calculus tasks. GPQA presents PhD-level questions in fields like biology, chemistry, and physics, validated for difficulty and factual accuracy to ensure robustness (Rein et al., 2023). MuSR tests reasoning over long texts, requiring integration of logic and context for scenarios like team allocation and murder mysteries, where most models struggle (Gao et al., 2023).

Other LLM-specific benchmarks include GLUE (General Language Understanding Evaluation) assesses language understanding through tasks like sentiment analysis, paraphrase

detection, and natural language inference (Wang et al., 2018). Its successor, SuperGLUE, evaluates more complex tasks, including textual entailment, coreference resolution, and question answering (Wang et al., 2020), MMLU (Massive Multitask Language Understanding) measures knowledge breadth and problem-solving skills across 57 domains, spanning humanities, STEM, and social sciences (Hendrycks et al., 2021).

Benchmarks like HumanEval and MBPP (Mostly Basic Python Problems) test coding abilities by evaluating functional correctness in code generation (Chen et al., 2021). TruthfulQA focuses on generating accurate and truthful answers, addressing model hallucinations (Lin et al., 2022). Winogrande evaluates commonsense reasoning using a large set of crowdsourced problems (Sakaguchi et al., 2019).

In multilingual and cross-lingual contexts, XTREME measures performance across tasks such as document retrieval, translation, and sentiment classification (Hu et al., 2020). For question answering and reasoning, the ARC (AI2 Reasoning Challenge) assesses models' ability to handle complex reasoning problems (Clark et al., 2018).

These benchmarks are vital for advancing LLM development, providing quantitative performance metrics, guiding fine-tuning strategies, and helping organizations select models suited to their needs.

Existing benchmarks provide a robust framework for evaluating large language models (LLMs) across diverse tasks, yet they exhibit notable gaps when looking for specific applications like project risk management. Benchmarks such as IFEval and MMLU emphasize instruction-following and domain-specific knowledge, while others like GLUE, SuperGLUE assess general language understanding and reasoning. However, these benchmarks fall short in evaluating a model's (or system's) ability to handle dynamic, real-world scenarios, integrate heterogeneous data sources, including subjectivity or project stakeholder and their subjective mental models and individual goals and objectives. To address these shortcomings, an LLM benchmark for project risk management would need to simulate the complexities of real-world tasks, assess adaptability to evolving conditions, and prioritize interpretability and actionable outputs.

Moreover, the dynamic nature of project environments creates an additional layer of difficulty. Unlike static datasets, where the corpus remains fixed and well-defined, project risk management requires constant integration of new information. Stakeholder objectives evolve, external factors change, and new dependencies arise. Existing LLM approaches, which often excel at processing vast datasets or generating insights within static contexts, are not inherently equipped to operate in scenarios where they must continuously synthesize new information and adjust their analyses accordingly.

Another challenge lies in the nature of the reasoning required for effective risk management. While multi-step reasoning offers a structured framework for tackling complex problems, it is often constrained by the model's ability to maintain logical consistency and coherence over long reasoning chains. In dynamic project contexts, where risks can depend on subtle, implicit relationships between disparate data points, maintaining such consistency becomes critical. The addition of new data further exacerbates this challenge, as previous conclusions may need

to be invalidated or revised in light of new evidence—a process current LLM architectures do not inherently handle well.

Furthermore, project risks often stem from subjective interpretations, such as stakeholders' differing priorities or hidden assumptions about acceptable outcomes. These human factors are not easily encoded into benchmarks or modeled by current LLM capabilities. For instance, identifying that a project stakeholder's approval of "delivery on the 25th" contradicts another stakeholder's objective if there is a perceived and relevant different between end-of-day delivery and end-of-business-day delivery. Analyzing this requires not only nuanced reasoning but also an understanding of interpersonal dynamics and the context of the communication.

While advances like RAG enable LLMs to retrieve relevant information dynamically, they often lack mechanisms for prioritizing and reconciling conflicting data or assessing the relative importance of different stakeholder needs. Similarly, while benchmarks that simulate dynamic reasoning—such as those requiring real-time synthesis of information—are emerging, they remain in early stages and do not yet address the full scope of challenges posed by dynamic, multi-stakeholder project environments.

4. Toward a Benchmark for LLM-Based Project Risk Management

The development of a benchmark tailored to LLM-based project risk management represents a critical step toward advancing the field. This benchmark must address the complexities of real-world project environments, including the varied and dynamic nature of risks, the diversity of data sources, and the challenges of evaluation. On the other hand, its structure and task need to be rather open and simple in order to separate issues of information presentation from issues of risk detection and analysis.

A well-designed benchmark would not only enable rigorous testing of systems but also provide insights into their applicability and effectiveness for different project scenarios. In the following, a benchmark approach is proposed that encompasses data structures and data for such a benchmark, tasks and metrics for benchmark evaluation.

The concept is deliberately streamlined to ensure scalability across projects of all sizes—from small initiatives to large-scale endeavors. While the approach supports all project types, complex project data may require simple, straightforward transformations to align with the framework. Ultimately, this benchmark concept is designed for easy adoption, experimentation, and further enhancement.

4.1 Data for the Benchmark

A cornerstone of this benchmark is a comprehensive and diverse dataset designed to emulate real-world project environments. The dataset must incorporate a variety of data sources, including emails, notes, messages, contacts, terms and conditions, structured and unstructured tables, and complex documents requiring tool-based processing. It must also include metadata for each piece of information, capturing attributes such as the time of receipt, source, context, and relationships within the project. The complexity of data must range from straightforward mentions to intricate patterns requiring advanced analytical capabilities.

The dataset would encompass both formatted and non-formatted content, including text, images, structured data, and hard-to-structure elements. Scenarios would range from small datasets, suitable for in-context handling within prompts, to large datasets that necessitate iterative or multi-step analysis, potentially involving multi-agent systems.

An example dataset structure would look like this:

```
DataSet[1..x]_directory/
├── project_objective.txt
├── project_information.csv
├── project_risks.csv
└── resources_subdirectory/
    ├── resources_metadata.csv
    └── [Any additional files]
```

Fig. 1 example folder and file structure for one data set / one project

Project_objective.txt should contain the project title and any textual description of the project objective as agreed upon by the initiating stakeholders. Project_information.csv would hold information relevant to the chronological progression of the project. It could reference documents within the resources_subdirectory folder. Project_risks.csv would contain risks that are known to be present in the data set. The resources_subdirectory can contain additional files containing project-relevant documents. The corresponding resources_metadata.csv should contain identifiers of these files, and additional metadata of unspecified format, e.g. ownerships, timestamps, version or status information etc.

Regarding the other csv files, the following structure is proposed to allow for simple storage in a tabular way.

Table 1 project_information.csv example structure

ID	date-time-description	sender_identifier	receiver_identifier	content
com_1	1st day of the project	CEO	project manager	goal is to set up a bike shed by day 7
com_2	1st day of the project	project manager	CEO	affirmation of the goal.
com_3	2nd day of the project	builder	project manager	able to build the bike shed. Will need 6 days for this.
info_piece_1	3rd day of the project	project manager assistant	none	CEO and vendor met in private – vendor agreed on project specific delivery dates and guarantees no delays wrt/ CEOs timeline. Information obtained in elevator, personal conversation.

The contents here are intentionally kept simple and abstract, as if they were already summarize before data ingestion. They could, however, hold any form of encoded data as well, like entire emails or documents.

This way, a simple data set models is defined that would serve the purpose of being able to store and access all project-relevant data during as-if chronologically progressing through the

project, with the ability to dive into potentially all project-relevant data, files and information. The definition of the data storage is intentionally not precise (cf. dates in the project_information.csv), as project risk information typically is not either.

Table 2 project_risks.csv example structure

ID	title	description	explanation	severity	urgency	references	assumption	type
1	Schedule Risk Due to Tight Timeline	The project might not meet the 7-day deadline due to a tight schedule and little margin for delay.	The builder requires 6 days starting on day 2 which leaves no cushion for unexpected issues. This presents a high risk of missing the fixed deadline, making the schedule very tight.	High	High	project goal, com_1, com_3	No unforeseen delays, builder's timeline is accurate	[see 4.2]

4.2 Synthetic data generation

Synthetic data generation is essential for constructing the dataset, as there are no large scientific data sets on projects, probably due to their typically high business-criticality in organizations (Mishra et al., 2023). In addition, it is established that synthetic data generation can enhance effectiveness of LLM-based solutions (Biswa, 2024), especially in tasks that are typically the domain of expert ratings.

The base scenarios would simulate flawless project progress, establishing a control environment. To introduce challenges, deliberate problems would be injected, such as delays, conflicting requirements, or ambiguities. These risks or risk-related occurrences would cover a spectrum of risk types:

- **Mentioned risks:** Explicit risks, alludes to 'known risks' (see above), e.g. taken from communication between two stakeholders. Basically copy & paste from observed communication or other pieces of information. Even this type of risk may already be highly subjective.
- **Temporal risks:** A risks that is imminent unless action is taken. E.g. a vendor said: "if we have not agreed on a mode of delivery until March 1st, I cannot guarantee timely delivery of the finale product". Such risks can also be mentioned indirectly, e.g. in counter-proposed project plans.
- **Complex risks**: Risks embedded within project data, in which only multi-step analyses would reveal a risk's existence. For example, PM demands delivery within a week, vendor says they will deliver according to terms and conditions. This category can obviously be made as complicated as imaginable. For the benchmark there would need to be a defined maximum number of information pieces that compounds a risk.
- **Context-specific dependencies:** For example, rollout milestones conflicting with significant public holidays like July 4th in the United States. This is obviously context-specific, e.g. depentent on countries in focus of the project.

- **Mitigated risks:** Some risks could be present within the data set, but also be mitigated somewhat later throughout the chronological progression of the project. For example, say a project contract foresaw the issuance of a non-conformance report at the time of reaching a delivery milestone and penalties in case the vendor did not comply. The project manager, however, has trust in the vendor's statement that they would hand in such a report, only later. This would be either a risk for the project and its manager. If the benchmark askked for a specific date and time for risk analysis that lies after the handing-in of the report, this should be considered a non-risk.

The benchmark, by means of project information, will also address informal or implicit information, such as personal information notes or information on stakeholder side conversations. While potentially introducing noise, this data could reveal unstructured insights critical to risk identification. Systems must balance handling informal information with mitigating false positives.

For the benchmark data we first generate 500 rough project descriptions that are the basis for the flawless project narratives. These are from the industries: information technology, construction, healthcare, marketing, finance, education, manufacturing, energy, retail, transportation. The python library ‚fakr' is used to generate these project names. An additional script invoking OpenAI's GPT 3.5-turbo helps generate: a project background with metadata like project start and end date, a background story on characters (not to be included in the data sets), the project objective, project_information.csv table with randomly between 5 and 200 messages or pieces of information with connection to the project. An independent AI judge is used to dismiss messages that may themselves already contain risks for the project. In addition between one and five of the risk scenarios (above) are randomly selected. A Chain-of-Thought approach is used to modify project information, adding additional details or altering existing ones to ensure the dataset includes project risks. The risks and their self-rated severity, urgency (from and end-points perspective) are recorded in the project_risks.csv.

4.3 Benchmark Tasks

Task Mode 1: Total Analysis

In this task, systems are provided with all available data at once. The objective is to produce a comprehensive risk analysis within one or multiple limited timeframe(s), e.g. one hour, which would be defined in the final benchmark tasking. This would very likely not matter for small project data sets, but for larger ones when multiple pieces of information have to be connected. In such instances, systems would need make informed decisions regarding what pieces of the data set to look at in each analytical step, as there are just too many options to evaluate the entire data set every time. Just like chess vs. Go, some data sets would technically be computable, some would not be and make clever heuristics necessary for analytical success.

Inputs would be 1) the data set as described in the section earlier and 2) date and time for which the analysis should be made, in no specific format and 3) potential further context information.

Outputs would include a list of risks, each accompanied by:

- An ID
- A title < 100 characters

- A brief description < 500 characters
- An explanation of the reasoning behind the identified risk, < 500 characters
- Severity indicator, ‚low', ‚medium', ‚high' or 'very high'
- urgency indicator, ‚low', ‚medium', ‚high' or 'very high'
- A reference list of the information used to inform the decision, < 500 characters
- Any assumption made, < 500 characters

This task mode evaluates a system's ability to process a fixed dataset and identify risks effectively, simulating scenarios where a subset of project information is pre-compiled for analysis.

Task Mode 2: Ongoing Analysis

The extended task involves dynamic, real-time analysis as new information becomes available. Systems must identify risks iteratively, using advanced retrievers and rankers to navigate large datasets. This mode incorporates layers of abstraction and requires the system to adapt dynamically, invalidating previously identified risks if subsequent information renders them irrelevant. This task mode better evaluates systems designed for continuous integration, analysis, and synthesis of project data, reflecting real-world conditions. In a first conceptualization of the benchmark, that task would be to ingest all information available and present the latest risk profile that is available. But future conceptualizations could also review the trajectories of risk analysis and evaluations by the systems for a better understanding of the dynamics of risks in projects.

4.4 Evaluation metrics and preliminary results

Metrics must assess whether the AI system successfully identified the predefined risks. We propose precision and accuracy to be the first main indicators of the benchmark. Precision – this would ensure the system identifies only relevant risks that are by definition of the set part of the risks that are supposed to be found. For this metric to be meaningful, the data sets will need to be reviewed by experts or their creation process needs to be curated in a final version of the benchmark. Accuracy would measure the completeness of identified risks with regard to the defined risks in the data set.

Evaluating systems for project risk management presents unique challenges. A primary issue is the "chicken-and-egg" problem: the lack of a perfect, automated assessors means that evaluations often rely on approximations. Automated assessment could compare system outputs against expected outcomes in predefined scenarios, offering a scalable solution, though not without limitations. Expected solutions could be quality-assured by experts to enhance reliability.

It is to be expected that after the first attempts to solve the benchmark task, new risks will be identified that have not been part of the data sets but would convince experts that these could be reasonable risks for the underlying project. These risks would need to be either added to the risks if reasonable, or marked as insignificant in the context information, or deemed as irrelevant in the context by raters.

First tests showed promising results for both LLMs and this proposed benchmark. For one accuracy was quite high, showing that in naïve and simple settings, LLMs (GPT 4o used, all file data simply dumped as context into a prompt, advising on desired output format) in general were able to detect the risks introduced into the datasets. However, precision suffers from AI systems oftentime finding novel risks that were originally not part of the risks. These findings may not be entirely incorrect. However, they show that this benchmark may eventually have to account for subjectivity in risk assessment in different ways.

5. Systems to be Benchmarked

Systems suitable for benchmarking would share several core components. A knowledge base would act as a repository for all collected data and metadata. Algorithms would, at a minimum, ingest the project_information.csv file, potentially transforming it to facilitate more effective insight generation and efficient reasoning. An analyzer component would identify risks or inconsistencies within the retrieved information. Additional components could be introduced to track and model the mental representations assumed to be held by the identified stakeholders. An output component would synthesize analytical insights into the prescribed formats.

In the case of compound risks that depend on multiple pieces of information, the analyzer would need to employ strategies to effectively combine data points without resorting to exhaustive juxtaposition, which would result in exponential growth in the number of analyses. Approaches that leverage agentic reasoning to prioritize and identify relevant parts of the database for further exploration are likely to perform well. Such solutions would be conceptually aligned with the transformer principle, which underpins the architecture of modern GPT-style language models, where attention mechanisms dynamically focus on the most relevant information to achieve efficient and scalable processing.

For an algorithm, temporality will play a major role. Especially in the case of temporal risks, ther will need to be an understanding of what happened when and what point in time is assumed to be 'now'. This could open up even more modes of dynamic benchmarking in the future.

Candidate systems may range from single-shot large-context models to complex multi-agent frameworks. The benchmark would evaluate their effectiveness in different tasks and scenarios, enabling researchers and practitioners to understand their strengths and limitations.

6. Outlook and conclusion

The proposed benchmark for LLM-based project risk management represents a critical step toward bridging the gap between AI innovation and the practical needs of project environments. By addressing the complexities of dynamic, real-world data and introducing robust evaluation metrics, this benchmark has the potential to drive significant advancements in both research and practical applications. It aims to provide a rigorous framework for evaluating LLM-based systems, ensuring that their capabilities align with the nuanced demands of modern project risk management.

Müller et al. (2023) demonstrated that the effective dissemination of meta-knowledge can significantly enhance outcomes in complex, high-stakes scenarios. Inspired by this, the benchmark seeks to assess whether LLM-based AI systems can play a similarly transformative

role in project risk management or if their application risks becoming a distraction rather than a solution.

Future iterations of the benchmark should prioritize incorporating real-world data, refining synthetic data generation techniques, and exploring dynamic scenarios where systems must adapt and continuously synthesize new information. By standardizing evaluation practices, the benchmark helps uncover the potential and limitations of LLM-based systems, ensuring their adoption leads to actionable benefits without unnecessary complexity.

Ultimately, this benchmark will guide the development of robust, adaptable, and effective solutions capable of addressing the complex demands of project risk management. It provides a foundation for aligning AI advancements with practical needs, fostering tools that not only identify and mitigate risks but also enhance decision-making processes in an increasingly dynamic and data-driven project landscape.

7. References

Al-Ajmi, H., & Makinde, E. (2018). Risk Management in Construction Projects. *Journal of Advanced Management Science*, 113-116. https://doi.org/10.18178/joams.6.2.113-116

Biswa, R. (2024). Eedi - Mining Misconceptions in Mathematics, 1st Place Detailed Solution. Retrieved 28 Jan 2025 from https://www.kaggle.com/competitions/eedi-mining-misconceptions-in-mathematics/discussion/551688.

Chen, M., Tworek, J., Jun, H., Yuan, Q., de Oliveira Pinto, H. P., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., Ray, A., Puri, R., Krueger, G., Petrov, M., Khlaaf, H., Sastry, G., Mishkin, P., Chan, B., Gray, S., ... Zaremba, W. (2021). Evaluating large language models trained on code. *arXiv*. https://arxiv.org/abs/2107.03374

Clark, P., Cowhey, I., Etzioni, O., Khot, T., Sabharwal, A., Schoenick, C., & Tafjord, O. (2018). *Think you have solved question answering? Try ARC, the AI2 reasoning challenge*. arXiv. https://arxiv.org/abs/1803.05457

Fang, C., Marle, F., & Xie, M. (2017). Applying Importance Measures to Risk Analysis in Engineering Project Using a Risk Network Model. *IEEE Systems Journal*, 11, 1548-1556. https://doi.org/10.1109/JSYST.2016.2536701

Gao, P., Zhang, L., He, Z., Wu, H., & Wang, H. (2023). *Learning multilingual sentence representations with cross-lingual consistency regularization*. arXiv. https://arxiv.org/abs/2306.06919

Geraldi, J., Locatelli, G., Dei, G., Söderlund, J., & Clegg, S. (2024). AI for Management and Organization Research: Examples and Reflections from Project Studies. Project Management Journal, 55(4), 339-351. https://doi.org/10.1177/87569728241266938

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., & Steinhardt, J. (2021). Measuring massive multitask language understanding. *arXiv*. https://arxiv.org/abs/2009.03300

Hu, J., Ruder, S., Siddhant, A., Neubig, G., Firat, O., & Johnson, M. (2020). XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. *arXiv*. https://arxiv.org/abs/2003.11080

Levene, R., & Lewis, M. (2015). Project Risk Management. In Wiley Encyclopedia of Management (eds C.L. Cooper, S. Roden, M. Lewis and N. Slack). https://doi.org/10.1002/9781118785317.weom100206

Lin, S., Hilton, J., & Evans, O. (2022). TruthfulQA: Measuring how models mimic human falsehoods. *arXiv*. https://arxiv.org/abs/2109.07958

Mishra, A., Tripathi, A., & Khazanchi, D. (2023). A Proposal for Research on the Application of AI/ML in ITPM: Intelligent Project Management. *Int. J. Inf. Technol. Proj. Manag.*, 14, 1-9. https://doi.org/10.4018/ijitpm.315290.

Müller, R., Graf, B., Ellwart, T., & Antoni, C. H. (2023). How Software Agents Can Help to Coordinate Emergency Response Teams: Adaptive Team Performance Comparing Manual and Automated Team Communication. *Journal of Business and Psychology*, *38*(5), 1121–1137. https://doi.org/10.1007/s10869-022-09858-4

Müller, R., Locatelli, G., Holzmann, V., Nilsson, M., & Sagay, T. (2024). Artificial Intelligence and Project Management: Empirical Overview, State of the Art, and Guidelines for Future Research. Project Management Journal, 55(1), 9-15. https://doi.org/10.1177/87569728231225198

Perry, J. (1986). Risk management — an approach for project managers. *International Journal of Project Management*, 4, 211-216. https://doi.org/10.1016/0263-7863(86)90005-0

Rehácek, P. (2017). Risk management standards for project management. *International Journal of Advanced and Applied Sciences*, 4, 1-13. https://doi.org/10.21833/IJAAS.2017.06.001

Rein, D., Hou, B. L., Stickland, A. C., Petty, J., Pang, R. Y., Dirani, J., Michael, J., & Bowman, S. R. (2023). *GPQA: A graduate-level Google-proof Q&A benchmark*. arXiv. https://arxiv.org/abs/2311.12022

Sakaguchi, K., Le Bras, R., Bhagavatula, C., & Choi, Y. (2019). WinoGrande: An adversarial Winograd Schema challenge at scale. *arXiv*. https://arxiv.org/abs/1907.10641

Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. A. M., Abid, A., Fisch, A., Brown, A. R., Santoro, A., Gupta, A., Garriga-Alonso, A., Kluska, A., Lewkowycz, A., Agarwal, A., Power, A., Ray, A., Warstadt, A., Kocurek, A. W., Safaya, A., Tazarv, A., … Wu, Z. (2023). *Beyond the imitation game: Quantifying and extrapolating the capabilities of language models*. arXiv. https://arxiv.org/abs/2206.04615

Thamhain, H. (2013). Managing Risks in Complex Projects. *Project Management Journal*, 44, 20 - 35. https://doi.org/10.1002/pmj.21325

Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2020). SuperGLUE: A stickier benchmark for general-purpose language understanding systems. *arXiv*. https://arxiv.org/abs/1905.00537

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. (2018). GLUE: A multi-task benchmark and analysis platform for natural language understanding. *In Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP* (pp. 353–355). Brussels, Belgium: Association for Computational Linguistics.

Williams, R., Bertsch, B., Dale, B., Van der Wiele, T., Van Iwaarden, J., Smith, M., & Visser, R. (2006). Quality and Risk Management: What Are the Key Issues? TQM Magazine, 18, 67-86.

Willumsen, P., Oehmen, J., Stingl, V., & Geraldi, J. (2019). Value creation through project risk management. *International Journal of Project Management*, 37(5), 731–749. https://doi.org/10.1016/j.ijproman.2019.01.007

Zhang, H. (2011). Two Schools of Risk Analysis: A Review of past Research on Project Risk. *Project Management Journal*, 42, 18–5. https://doi.org/10.1002/pmj.20250

Zhou, J., Lu, T., Mishra, S., Brahma, S., Basu, S., Luan, Y., Zhou, D., & Hou, L. (2023). Instruction-following evaluation for large language models. arXiv. https://arxiv.org/abs/2311.07911