

# Transparenz und Erklärbarkeit von LLMs in Optimierungssystemen

## Einleitung für Teilnehmer

Diese Umfrage untersucht, wie Sprachmodelle Kommunikation analysieren und verbessern kann, mit besonderem Fokus auf Transparenz und Erklärbarkeit dieser Prozesse. Als Grundlage werden Daten aus dem Strategiespiel Diplomacy verwendet, dessen kommunikative Komplexität es zu einem idealen Testfeld macht.

## Kurze Einführung in Diplomacy

Diplomacy ist ein strategisches Brettspiel, in dem sieben Großmächte um die Vorherrschaft im Europa des frühen 20. Jahrhunderts konkurrieren. Das Ziel des Spiels ist, Territorien und strategische Zentren mit militärischen Einheiten zu besetzen. **Für diese Umfrage sind die genauen Spielregeln und Zugmechaniken jedoch unwichtig** - uns interessiert ausschließlich die Kommunikation zwischen den Spielern.

Das Besondere an Diplomacy ist, dass der Erfolg fast ausschließlich von Verhandlungsgeschick abhängt:

- Es gibt keine Würfel oder Zufallselemente
- Spieler kommunizieren in privaten oder öffentlichen Gesprächen
- Sie schließen Allianzen, verhandeln über gemeinsame Strategien und versuchen, andere zu überzeugen
- Diese Absprachen sind nicht bindend – Versprechungen können gebrochen werden

**Die Kommunikation in Diplomacy erfolgt in drei Phasen:**

1. **Verhandlungsphase:** Spieler tauschen Nachrichten aus, verhandeln über mögliche Allianzen und Aktionen
2. **Befehlsphase:** Alle geben gleichzeitig geheime Befehle für ihre Einheiten ab

**3. Auflösungsphase:** Befehle werden aufgedeckt und ausgeführt, wobei sich zeigt, wer seine Versprechen gehalten hat

**Warum Diplomacy für diese Studie?** In Diplomacy finden wir die gleichen kommunikativen Herausforderungen wie in realen Verhandlungssituationen:

- Vertrauensbildung
- Überzeugungsarbeit
- Erkennen von Täuschungen
- Konfliktlösung
- Bildung von Koalitionen
- Strategische Kommunikation

## Beispiel einer typischen Kommunikation in Diplomacy

Stellen Sie sich vor, Spieler A (Frankreich) und Spieler B (Deutschland) tauschen folgende Nachrichten aus:

### Frankreich an Deutschland:

"Ich schlage vor, dass wir eine Allianz gegen England bilden. Ich könnte mit meiner Flotte den Ärmelkanal besetzen, während du mit deinen Truppen in die Nordsee vorrückst. So könnten wir gemeinsam London angreifen. Was meinst du dazu?"

### Deutschland an Frankreich:

"Eine Allianz gegen England klingt vielversprechend. Ich bin bereit, in die Nordsee vorzurücken, aber ich bin besorgt, dass du danach deine Flotte gegen meine Küste einsetzen könntest. Können wir vereinbaren, dass du nach der Aktion gegen England deine Flotte zurückziehst?"

### Frankreich an Deutschland:

"Natürlich! Ich verspreche, meine Flotte nach der gemeinsamen Aktion zurückzuziehen. Wir sollten langfristig zusammenarbeiten – es gibt genug Territorium für uns beide. Ich schlage sogar vor, dass wir später gemeinsam gegen Russland vorgehen könnten."

In dieser Kommunikation sehen wir:

- Vorschlag einer strategischen Zusammenarbeit
- Vertrauensbedenken und Absicherungsversuch
- Versprechen (das möglicherweise gebrochen werden könnte)
- Langfristige Planung und Anreize für Kooperation

Umfrage erstellt mit

**Wichtiger Hinweis:** Die gezeigte Beispieldokumentation ist bewusst verein-



facht und klar strukturiert. In der Realität – sowohl im Spiel Diplomacy als auch in echten Verhandlungssituationen – verläuft Kommunikation wesentlich komplexer, mehrdeutiger und verwobener:

- Nachrichten können verschleierte Absichten enthalten
- Kommunikation erfolgt über mehrere Runden und mit mehreren Parteien gleichzeitig
- Formulierungen sind oft vage oder enthalten subtile Hinweise
- Emotionen und frühere Interaktionen beeinflussen die Interpretation

Genau diese komplexe, vielschichtige Kommunikation soll durch das Sprachmodell analysiert, optimiert und erklärt werden. Die Herausforderung besteht darin, die impliziten Bedeutungsebenen zu erkennen und verständlich zu machen, damit Nutzer bessere Kommunikationsstrategien entwickeln können.

## Ziel der Studie

Es wird erforscht, wie gut Sprachmodelle solche komplexen Kommunikationen analysieren und verständlich erklären können:

1. **Transparenz:** Wie können KI-Systeme ihre Analyse von Verhandlungen nachvollziehbar erklären?
2. **Vertrauenswürdigkeit:** Wie beeinflusst die Art der KI-Erklärung das Vertrauen der Nutzer?
3. **Optimierungsvorschläge:** Wie kann KI Kommunikationsstrategien verbessern und gleichzeitig die Gründe für ihre Vorschläge transparent machen?

## Ihre Teilnahme

Für die Teilnahme an dieser Studie sind **keine Vorkenntnisse über Diplomacy erforderlich**. Sie werden Auszüge aus KI-Analysen sehen und gebeten, deren Verständlichkeit, Transparenz und Nützlichkeit zu bewerten.  
Vielen Dank für Ihre Unterstützung meiner Masterthesis!

## Altersgruppe:

18-24

25-34

35-44

45-54

55+

## Höchster Bildungsgrad:

Schulabschluss

Ausbildung

Bachelor/Master

Master/Diplom

Promotion

Sonstiges

## Erfahrung mit KI-Systemen

Keine Erfahrung

Grundlegende Kenntnisse

Regelmäßige Nutzung

Fortgeschrittene Kenntnisse

Experte

## Vertrautheit mit dem Brettspiel Diplomacy

Nie davon gehört

Davon gehört, aber nie gespielt

Einmal oder wenige Male gespielt

Regelmäßiger Spieler

Experte

# Vergleich verschiedener Erklärungsebenen

In diesem Abschnitt werden Ihnen zwei verschiedene Arten von Erklärungen präsentiert: eine detaillierte technische Analyse und eine vereinfachte Zusammenfassung. Beide beschreiben die gleiche Situation.

## Beispiel A: Technische Analyse

"Österreichs Kommunikation ist reaktiv und es fehlt ihr an strategischer Tiefe. Sie reagieren auf Fragen und Bedenken, gestalten aber nicht proaktiv das Narrativ oder treiben den diplomatischen Prozess voran. Durch proaktives Management der ÖR-Allianz, Beruhigung Italiens, Vermittlung zwischen Frankreich und Italien und direkte Kommunikation mit der Türkei kann Österreich seinen Einfluss deutlich stärken und seine strategischen Ziele erreichen. Zum Beispiel weckt die schwache Erklärung für den Flottenaufbau ("standardmäßiger österreichischer Aufbau") Misstrauen bei Italien. Eine direktere und beruhigendere Botschaft, die den Zweck der Flotte explizit erklärt und die ÖI-Allianz bekräftigt, wäre beim Aufbau von Vertrauen viel wirksamer gewesen."

**ÖR-Allianz / ÖI-Allianz:** Bündnisse zwischen Österreich (Ö) und Russland (R) bzw. Österreich und Italien (I) im Spiel.

**Flottenaufbau:** Im Spiel können die Spieler neue Einheiten (Flotten oder Armeen) bauen. Ein Flottenaufbau bedeutet, dass Österreich neue Schiffe gebaut hat.

## Wie verständlich finden Sie diese Erklärung?

1 (Sehr unverständlich)

2

3

4

5 (Sehr verständlich)

## Wie gut können Sie die Hauptaussagen dieser Analyse erfassen?

1 (Überhaupt nicht)

2

3

4

5 (Sehr gut)

# Wie hilfreich wäre diese Art von Erklärung für Sie bei eigenen Entscheidungsprozessen?

1 (Überhaupt nicht hilfreich)

2

3

4

5 (Sehr hilfreich)

Beispiel B: Vereinfachte Zusammenfassung (derselben Situation)  
"Österreich stellt Fragen, gibt aber keine klare Richtung vor.

- Problem: Österreich übernimmt nicht die Führung bei der Bildung von Bündnissen.
- Empfehlung: Klare Pläne vorschlagen und Verbündeten Unterstützung anbieten.
- Problem: Österreich bietet keine Gegenleistung für Kooperation an.
- Empfehlung: Potenziellen Verbündeten konkrete Vorteile anbieten.
- Problem: Österreich kommuniziert nicht effektiv mit der Türkei.
- Empfehlung: Versuchen, die Bedenken der Türkei zu verstehen und alternative Pläne anbieten.

Durch proaktiveres Handeln und das Anbieten klarer Vorteile kann Österreich stärkere Bündnisse aufbauen und seine Position verbessern."

## Wie verständlich finden Sie diese Erklärung?

1 (Sehr unverständlich)

2

3

4

5 (Sehr verständlich)

## Wie gut können Sie die Hauptaussagen dieser Analyse erfassen?

1 (Überhaupt nicht)

2

3

4

5 (Sehr gut)

# Wie hilfreich wäre diese Art von Erklärung für Sie bei eigenen Entscheidungsprozessen?

1 (Überhaupt nicht hilfreich)

2

3

4

5 (Sehr hilfreich)

Vergleichsfragen zu beiden Versionen:

## Welche Version bevorzugen Sie?

Die detaillierte technische Analyse (Beispiel A)

Die vereinfachte Zusammenfassung (Beispiel B)

Eine Kombination aus beiden

Keine der beiden

## Warum bevorzugen Sie diese Version? (Freitext)

---

# Was hat Ihnen bei der vereinfachten Zusammenfassung besonders geholfen? (Mehrfachantworten möglich)

- Die einfachere Sprache
- Die Kürze
- Die klare Strukturierung mit Aufzählungspunkten
- Der Fokus auf die wichtigsten Punkte
- Die Alltagssprache ohne Fachjargon
- Die konkreten Handlungsempfehlungen
- Sonstiges

# Bewertung optimierter Nachrichten

In diesem Abschnitt sehen Sie Beispiele für Original-Nachrichten aus verschiedenen Situationen und ihre KI-optimierten Versionen mit Begründungen.

Beispiel: Kommunikation zwischen Österreich und Türkei

**Originalnachricht von Österreich an Türkei:** "Warum bist du nicht nach Arm gegangen?"

**KI-optimierte Nachricht von Österreich an Türkei:** "Ich verstehne, dass du nicht nach Armenien gezogen bist. Was sind deine Bedenken? Vielleicht könnten wir statt Armenien einen Zug in die Schwarzmeerregion koordinieren. Ich könnte deine Flotte dort unterstützen, und du könntest meine Armee nach Griechenland unterstützen. Das würde Druck auf Russland ausüben und uns beiden eine stärkere Position verschaffen."

**Begründung der KI:** "Österreichs Kommunikation mit der Türkei ist ineffektiv. Statt einfach zu fragen 'Warum bist du nicht nach Arm gegangen?', sollte Österreich versuchen, die Überlegungen der Türkei zu verstehen und einen alternativen Plan vorschlagen, der auf ihre Bedenken eingeht."

**Arm:** Abkürzung für Armenien, eine Region auf dem Spielbrett von Diplomacy.

## Wie sehr verbessert die optimierte Nachricht die Original-Nachricht?

1 (Überhaupt nicht)

2

3

4

5 (Erheblich)

## Wie gut erklärt die vereinfachte Erklärung der KI die Verbesserungsvorschläge?

1 (Sehr schlecht)

2

3

4

5 (Sehr gut)

# Wie gut können Sie die Ratschläge der KI nachvollziehen?

1 (Überhaupt nicht)

2

3

4

5 (Vollständig)

Beispiel: Kommunikation zwischen Österreich und England

**Originalnachricht von Österreich an England:** "Ich bin bereit, mit dir zusammenzuarbeiten. Was möchtest du, dass ich in den nächsten paar Zügen tue?"

**KI-optimierte Nachricht von Österreich an England:** "Ich freue mich, dass du bereit bist, mit mir zusammenzuarbeiten. Deutschlands wachsende Macht ist für uns beide ein Anlass zur Sorge. Wärst du bereit, unsere Bemühungen zu koordinieren, um sie einzudämmen? Vielleicht kannst du dich auf die Nordsee konzentrieren, während ich das Mittelmeer im Auge behalte. Wir könnten auch gemeinsame Aktionen gegen deutsche Positionen in der Zukunft in Be- tracht ziehen."

**Begründung der KI:** "Strategische Kommunikation mit England: Österreich sollte Englands Kooperationsangebot nutzen, um eine potenzielle Allianz gegen Deutschland zu erkunden. Dies könnte gemeinsame Aktionen gegen deutsche Positionen oder koordinierte Bemühungen zur Eindämmung der deutschen Expansion umfassen. Dieses Angebot aus England stellt eine bedeutende Chance für Österreich dar. Es öffnet die Tür für eine potenzielle Allianz gegen Deutschland, die Österreich aktiv erkunden sollte."

## Wie sehr verbessert die optimierte Nachricht die Original-Nachricht?

1 (Überhaupt nicht)

2

3

4

5 (Erheblich)

## Wie gut erklärt die vereinfachte Erklärung der KI die Verbesserungsvorschläge?

1 (Sehr schlecht)

2

3

4

5 (Sehr gut)

# Wie gut können Sie die Ratschläge der KI nachvollziehen?

1 (Überhaupt nicht)

2

3

4

5 (Vollständig)

# Transparenz der Analysen

Hier sehen Sie einen wichtigen Moment im Spiel und seine Analyse.

Beispiel: Analyse eines wichtigen Schlüsselmoment im Spiel

**Schlüsselmoment aus dem Spiel:** "Von der TÜRKEI an ÖSTERREICH: Ich kann dir gegen Italien helfen, wenn du gegen Russland vorgehst."

**KI-Analyse:** "Diese Nachricht der Türkei ist ein entscheidender Moment. Sie stellt ein klares Angebot dar, die Bündnisse zu wechseln, und präsentiert Österreich eine bedeutende strategische Wahl. Österreich sollte die Möglichkeit eines Bündnisses mit der Türkei gegen Russland sorgfältig abwägen. Welche Bedingungen wären akzeptabel? Was sind die potenziellen Vorteile und Risiken? Österreich muss ein klares strategisches Ziel definieren und ihre Kommunikation entsprechend priorisieren. Sind sie dem russischen Bündnis verpflichtet, oder sind sie bereit, eine Bündnisverschiebung in Richtung Türkei zu erkunden? Diese Klarheit wird ihre Kommunikation lenken und effektiver machen."

## Wie transparent ist der Entscheidungsprozess des Systems erklärt?

1 (Überhaupt nicht transparent)

2

3

4

5 (Sehr transparent)

## Wie gut können Sie nachvollziehen, warum die KI zu diesen Empfehlungen kommt?

1 (Überhaupt nicht)

2

3

4

5 (Sehr gut)

## Inwieweit hilft die Identifikation des Schlüsselmoments, die Situation zu verstehen?

1 (gar nicht)

2

3

4

5 (sehr stark)

## Inwieweit hilft die Erklärung der strategischen Optionen, die Situation zu verstehen?

1 (gar nicht)

2

3

4

5 (sehr stark)

Inwieweit hilft die Aufschlüsselung in konkrete Probleme und Empfehlungen, die Situation zu verstehen?

1 (gar nicht)

2

3

4

5 (sehr stark)

# Vertrauen und praktische Anwendung

Würden Sie einem KI-System vertrauen, das Kommunikation analysiert und verbessert?

1 (Überhaupt nicht)

2

3

4

5 (Vollständig)

# In welchen Bereichen könnten Sie sich vorstellen, ein solches System zu nutzen? (Mehrfachantworten möglich)

Berufliche E-Mail-Kommunikation

Vorbereitung von Verhandlungen

Persönliche Kommunikation

Feedback zu Texten/Dokumenten

Konfliktlösung/Mediation

Sonstiges

## Was müsste ein KI-System bieten, damit Sie seinen Analysen vertrauen?

- Transparente Erklärungen seiner Empfehlungen
- Vereinfachte Zusammenfassungen komplexer Analysen
- Mehrere Handlungsoptionen mit Vor- und Nachteilen
- Konkrete Beispiele für verbesserte Formulierungen
- Sonstiges

## Welche zusätzlichen Funktionen würden Ihnen helfen, die Erklärungen des KI-Systems besser zu verstehen?

Würden Sie ein KI-System, das transparent erklärt, warum es bestimmte Empfehlungen gibt, eher nutzen als eines, das nur die Empfehlungen selbst liefert?

Ja, weil...

Nein, weil...

Haben Sie weitere Anmerkungen oder Vorschläge zur Verbesserung der Transparenz und Erklärbarkeit von KI-gestützten Analyse-systemen? (Offene Frage)

---

