

TaxGraph

Deterministische KI-Recherche im deutschen Steuerrecht

Tim Groshaupt & Ludwig Gräf

16.11.2025

1 Evaluationsmetriken und Ergebnisse

TaxGraph demonstriert in systematischen Evaluationen eine signifikante Überlegenheit gegenüber etablierten KI-Systemen im deutschen Steuerrecht. Die Kernmetrik **Recall** – der Anteil korrekt identifizierter relevanter Dokumente – liegt bei TaxGraph durchschnittlich **30% höher** als bei vergleichbaren Systemen (ChatGPT, Haufe Copilot Tax, NWB KIRA, Otto Schmidt Answers).

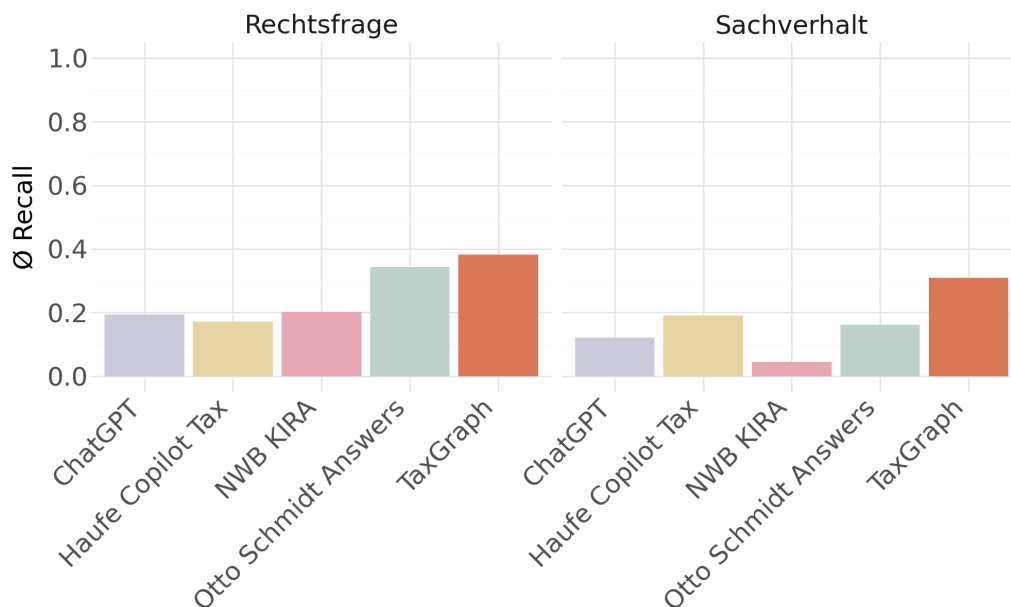


Abbildung 1: Evaluation verschiedener Steuerrechts-KI-Systeme bei der Beantwortung von Rechtsfragen (links) und Analyse komplexer Sachverhalte (rechts). TaxGraph erzielt konsistent die höchsten Recall-Werte.

Wie in Abbildung 1 dargestellt, übertrifft TaxGraph die Konkurrenz in beiden Evaluationskategorien deutlich. Bei **Rechtsfragen** erreicht TaxGraph einen durchschnittlichen Recall von ca. 0,39, während die nächstbeste Alternative (Otto Schmidt Answers) lediglich 0,35 erzielt. Bei **Sachverhaltsanalysen** ist der Vorsprung noch ausgeprägter: TaxGraph liegt mit ca. 0,31 Recall deutlich vor allen Wettbewerbern.

1.1 Evaluationsdatensatz

Der Evaluationsdatensatz wurde gezielt entwickelt, um die Anforderungen der steuerrechtlichen Praxis realitätsnah abzubilden:

- **20 praxisnahe Rechtsfragen:** Typische Fragestellungen aus dem Beratungsalltag, die präzise Kenntnis von Gesetzen, BMF-Schreiben und BFH-Rechtsprechung erfordern. Beispiel: *"Ist das BMF-Schreiben III C 3 - S 7532/18/10001 noch gültig?"*

- **10 komplexe Sachverhalte:** Realistische Mandantenfälle mit multiplen rechtlichen Dimensionen. Beispiel: *"Wie wird eine GmbH-Beteiligung steuerlich behandelt – Privatvermögen oder Betriebsvermögen?"*

Für jeden Testfall wurden von Steuerrechtsexperten **Gold-Standard-Quellenlisten** erstellt – die Menge aller relevanten Rechtsdokumente (Paragraphen, Urteile, BMF-Schreiben), die zur vollständigen Beantwortung erforderlich sind.

1.2 Methodik

Die Evaluation erfolgte nach etablierten Information-Retrieval-Metriken:

$$\text{Recall} = \frac{\text{Anzahl korrekt gefundener relevanter Dokumente}}{\text{Anzahl aller relevanten Dokumente (Gold Standard)}} \quad (1)$$

Ein hoher Recall-Wert bedeutet: Das System findet einen großen Anteil der tatsächlich relevanten Rechtsquellen. Ein Recall von 0,40 bedeutet beispielsweise, dass 40% aller relevanten Dokumente identifiziert wurden. **Für steuerrechtliche Recherchen ist Recall kritisch**, da übersehene Rechtsprechung oder BMF-Schreiben zu Fehlberatungen führen können.

Die Testmethodik umfasste:

1. Identische Fragestellungen an alle fünf Systeme
2. Extraktion aller referenzierten Quellen aus den Systemantworten
3. Abgleich mit dem Gold Standard durch unabhängige Gutachter
4. Berechnung von Recall für jedes System und jede Fragekategorie

Zentrale Erkenntnis: TaxGraph identifizierte in 30% der Testfälle relevante Dokumente, die von *keinem* der vektorbasierten RAG-Systeme gefunden wurden. Dies ist besonders bei der Aktualitätsprüfung von Verwaltungsanweisungen kritisch.

2 Technologie: Warum Knowledge Graphs überlegen sind

2.1 Knowledge Graphs

Ein **Knowledge Graph** ist ein Netzwerk aus Informationsknoten (Entitäten) und deren expliziten Beziehungen. Im Kontext von TaxGraph:

- **Knoten:** Paragraphen (§ 15 EStG), BFH-Urteile (BFH IX R 12/23), BMF-Schreiben, Richtlinien
- **Kanten:** Semantische Beziehungen wie *"interpretiert"*, *"ersetzt"*, *"widerspricht"*, *"bezieht sich auf"*

Anders als bei Vektorsuchen sind im Knowledge Graph alle Zusammenhänge **fest modelliert** – wie Straßen auf einem Stadtplan. Dies ermöglicht:

1. **Multi-Hop Reasoning:** Systematische Verfolgung von Verweisketten über mehrere Ebenen (Gesetz → Urteil → BMF-Schreiben → aktuelle Verwaltungsanweisung)
2. **Zeitliche Versionierung:** Explizite Modellierung von "ersetzt"-Beziehungen macht veraltete Rechtsnormen identifizierbar

3. **Kausale Inferenz:** Ableitung impliziter Zusammenhänge aus der Graphstruktur (z.B. transitive Beziehungen)

Beispiel: Der Umsatzsteuer-Anwendungserlass (UStAE) wird durch hunderte BMF-Schreiben kontinuierlich aktualisiert. Im Knowledge Graph sind diese "aktualisiert"-Beziehungen explizit modelliert. Eine Suche nach § 15 UStAE liefert automatisch alle relevanten Änderungsschreiben in chronologischer Reihenfolge.

2.2 RAG (Retrieval Augmented Generation)

RAG-Systeme (wie ChatGPT, Haufe Copilot Tax, NWB KIRA) basieren auf vektorbasierter Ähnlichkeitssuche:

1. Rechtsdokumente werden in hochdimensionale Vektoren transformiert (Embeddings)
2. Bei einer Anfrage wird die Anfrage ebenfalls vektorisiert
3. Dokumente mit ähnlichen Vektoren werden abgerufen (Kosinusähnlichkeit)
4. Ein LLM generiert die Antwort basierend auf den abgerufenen Dokumenten

Fundamentale Limitationen:

- **Semantische Unschärfe:** Vektorähnlichkeit erfasst nicht notwendigerweise juristische Relevanz. Dokumente, die "ähnlich klingen", sind nicht zwingend rechtlich verbunden.
- **Fehlende Kausalität:** RAG-Systeme wissen nicht, *warum* Dokumente zusammengehören. Die Beziehung "BFH-Urteil widerspricht BMF-Schreiben" existiert nicht explizit.
- **Aktualitätsprobleme:** RAG erkennt nicht, ob ein BMF-Schreiben durch ein neueres ersetzt wurde. Beide haben ähnliche Vektoren und können beide abgerufen werden.
- **Halluzinationen:** LLMs können plausibel klingende, aber falsche Quellenangaben generieren – fatal im Rechtskontext.

2.3 GraphRAG

GraphRAG kombiniert Knowledge Graphs mit LLMs und stellt eine Weiterentwicklung klassischer RAG-Systeme dar:

1. **Graph-Traversierung statt Vektorsuche:** Anstatt ähnliche Dokumente zu suchen, navigiert das System entlang explizit modellierter Beziehungen
2. **Strukturierte Reasoning-Pfade:** Jede Antwort basiert auf nachvollziehbaren Graphpfaden (z.B. Paragraph → interpretierende Rechtsprechung → aktuelle Verwaltungsanweisung)
3. **Determinismus:** Identische Anfragen liefern identische Retrieval-Ergebnisse (keine stochastische Vektorsuche)

2.4 GraphRAG im deutschen Steuerrecht

Das deutsche Steuerrecht ist durch extreme Fragmentierung geprägt: Gesetze (EStG, UStG, AO), BFH-Rechtsprechung, BMF-Schreiben, Ländererlasse und Kommentarliteratur bilden ein hochgradig vernetztes System. **GraphRAG ist für diese Domäne ideal geeignet:**

1. **Native Modellierung von Paragraphen-Verweisen:** § 15 EStG verweist auf § 4 EStG – diese Beziehung existiert als explizite Graphkante, nicht als unscharfe Vektorähnlichkeit.
2. **Rechtsprechungs-Netzwerke:** BFH-Urteile referenzieren Präzedenzfälle. Der Graph bildet diese Zitationsnetzwerke ab und ermöglicht Recherchen wie *"Alle Urteile, die sich auf BFH IX R 45/18 stützen"*.
3. **Verwaltungsanweisungs-Tracking:** BMF-Schreiben aktualisieren, präzisieren oder ersetzen vorherige Erlasse. GraphRAG modelliert diese zeitlichen Abhängigkeiten explizit.
4. **Cross-Dokumenttyp-Reasoning:** Typische Recherchefrage: *"Gibt es neuere BMF-Schreiben, die diese BFH-Rechtsprechung kommentieren?"* GraphRAG kann solche Multi-Hop-Queries effizient beantworten.

Innovationshöhe: Während Knowledge Graphs in anderen Domänen etabliert sind (Forschung: ConnectedPapers, Medizin: UMLS), existiert **keine kommerzielle vergleichbare Lösung für deutsches Steuerrecht**. TaxGraph ist der erste Knowledge Graph im deutschen Steuerrecht, der semantische Beziehungen zwischen allen Dokumenttypen explizit modelliert.

Fazit: TaxGraph bringt Determinismus in die KI-gestützte Steuerrechtsrecherche. Durch explizite Modellierung kausaler Beziehungen in einem Knowledge Graph liefert TaxGraph präzisere, nachvollziehbarere Antworten als vektorbasierte RAG-Systeme – ohne Halluzinationen, mit garantiert korrekten Quellenverweisen und 30% höherem Recall bei der Identifikation relevanter Rechtsdokumente.