LinguaGraph — Prüfung mehrsprachiger KI: Messung sprachübergreifender Wertedivergenz in LLMs

Linguistic Divergence Score (LDS) • LLM-as-Subject • 55-Messungs-Replikation

BWKI 2026 – Bundeswettbewerb Künstliche Intelligenz

Autor: Teilnehmer/in (Eigenarbeit)

Kurzfassung: LinguaGraph misst, ob mehrsprachige KI-Systeme wertbeladene Konzepte (Gerechtigkeit, Freiheit, Verantwortung, Heimat, Erfolg) sprachübergreifend unterschiedlich strukturieren. Ein Within-Subject-Experiment (LLM-as-Subject) und eine Replikation über 55 Messungen (50 eindeutige Modelle) zeigen: Alle ZH-DE-Paare signifikant, die Kulturrichtung (DE Autonomie/Regeln vs. ZH Raum/Anspruch) übersteigt ein Zufalls-Nullmodell. Der Bericht ordnet die Befunde ehrlich ein (Grenzen: ~87 % chinesische Anbieter, sieben westliche Messungen, 8 nicht-signifikante englisch-haltige Paare, Alignierungs-Artefakte der Domänen-Kontrolle). Vollständige Unterstützungs-Offenlegung im Anhang.

LinguaGraph — BWKI 2026: Abstract and Introduction

Language: German (BWKI submission language) Status: Final (v0.13.3)

Abstract

Sprichst du eine andere Sprache, denkst du eine andere Welt?

Diese Arbeit stellt die Frage, ob Sprache nicht nur die Kommunikation, sondern die strukturelle Organisation von Wissen beeinflusst. Dazu wurde LinguaGraph entwickelt — ein System, das mithilfe von Large Language Models (LLMs) kognitive Graphen aus Texten extrahiert und sprachübergreifend vergleicht.

Die Methodik basiert auf zwei komplementären Pipelines:

LinguaGraph Pipeline: Extraktion von Konzepten aus Probandentexten (ZH/DE/EN), Konstruktion kognitiver Graphen und Berechnung des Linguistic Divergence Score (LDS) — eines neuartigen metrischen Maßes für strukturelle Divergenz zwischen Sprachen.

CognitiveSpace Pipeline: Automatisierte Extraktion eines mathematischen Wissensgraphen aus 68 Lehrbüchern (39 ZH, 18 EN, 11 DE) mit 556 Konzepten und 525 Relationen über vier Bildungsstufen (Grundschule bis Universität). Die sprachübergreifende Alignierung erzielt 219 dreisprachig abgedeckte Konzeptgruppen (39 %) bei null strukturellen Konflikten.

Der CognitiveSpace-Graph wird als 3D-Kugelschalenvisualisierung dargestellt, die die Wissensexpansion von der Kernmathematik zur Hochschulmathematik sichtbar macht — mit interaktiven Sprachfiltern (ZH/EN/DE).

Anwendungsszenario — Prüfung mehrsprachiger KI-Systeme: Da LLMs überwiegend mit englischen Daten trainiert werden, ist bislang unklar, ob ihre Wertkonzepte (Gerechtigkeit, Freiheit, Verantwortung, Heimat, Erfolg) sprachübergreifend konsistent sind — ein blinder Fleck gängiger KI-Evaluation, die nur Aufgabenerfüllung misst. LinguaGraph macht genau dieses Modellverhalten prüfbar: Das LLM-as-Subject-Experiment (Within-Subject) quantifiziert, ob und wo ein mehrsprachiges Modell wertbeladene Konzepte sprachabhängig strukturiert, und benennt die konkret divergierenden Konzept-Bestandteile. Der Output ist ein interpretierbarer Divergenzbericht pro Modell und Sprachpaar.

Zentrale Beiträge: - Linguistic Divergence Score (LDS) als neuartige Metrik für sprachübergreifende Strukturanalyse — inklusive LDS-K (Wissen) , LDS-C (Kognition) und ΔLDS = LDS-C LDS-K - Null Model Suite: Falsifikation der Annahme, dass LDS-K sprachgetriebene Divergenz misst — tatsächlich dominieren Gradverteilungsstrukturen, und Lehrbuchwissen konvergiert sprachübergreifend - Kernbeitrag verschiebt sich zu ΔLDS (menschliche Kognition minus Lehrbuchstruktur), der den sprachspezifischen Anteil isoliert — die N=15-Analyse zeigt ΔLDS ≈ 0 unter Between-Subject-Bedingungen und spezifiziert, wann ΔLDS > 0 nachweisbar wäre - Erster systematischer Vergleich mathematischer Wissensstrukturen über ZH/EN/DE hinweg (556 Konzepte, 4 Nullmodelle, 19-Modell-Benchmark) - CognitiveSpace: skalierbare 3D-Visualisierung mit 556 Konzepten aus 68 Lehrbüchern - Vollständige Pipeline: Textextraktion → Graphkonstruktion → Alignierung → Analyse → Visualisierung - Humanvalidierung: Erweiterte, QC-geprüfte LDS-C-Analyse an N=15 Probanden (6 DE • 6 ZH • 3 EN) über drei Ebenen (Konzept, Kategorie, Relation) — ehrliches negatives Ergebnis unter Between-Subject: LDS-C (0.93 – 0.96) ist von Teilnehmervariabilität nicht unterscheidbar (Split-Half-Boden 0.92 – 0.96; Label-Permutation 0.94); die früheren N=8-Befunde (0.70 – 0.75) werden nicht repliziert - LLM-as-Subject (Within-Subject): Dasselbe LLM antwortet in ZH/DE/EN auf dieselben 5 Themen → Sprachsignal ist nachweisbar (LDS-C 0.93 – 0.96 Boden 0.85 – 0.87); LMM identifiziert den Sprachcode als dominanten Organisator (same_lang +0.038, p<0.001; same_frame +0.001, p=0.90); ZH-DE trägt eine strukturelle Ebene jenseits der Assoziationsstatistik; eine modellübergreifende Replikation (50 eindeutige Modelle, überwiegend chinesische Anbieter plus sieben westliche Messungen) zeigt: alle 55 ZH-DE-Sprachpaare signifikant (p<0.05), aber 8/165 Sprachpaaren (alle englisch-bezogen) nicht; die Kulturrichtung (DE Autonomie/Regeln

vs. ZH Raum/Anspruch) übersteigt ein frequenz-angepasstes Zufalls-Nullmodell bei allen Schwellen ($p<0.001$), am stärksten bei hoher Übereinstimmung (218 Konzepte mit ≥10 Stimmen vs. 147 zufällig erwartet) - Anwendungsfall — KI-Validierung: Divergenzbericht pro Modell und Sprachpaar (LDS-C + konkret divergierende Konzeptbestandteile + Domänen-Asymmetrie) — für Entwickler (Vorabprüfung mehrsprachiger Modelle), Regulierer (Transparenz gemäß EU AI Act) und Forscher (kulturelle Werte in KI)

Die Arbeit demonstriert, dass LLM-gestützte Graphanalyse ein vielversprechendes Werkzeug zur Untersuchung sprachlicher Einflüsse auf die Wissensorganisation darstellt — mit Implikationen für die bilinguale Bildung und die KI-Forschung. Eine Null Model Suite falsifiziert die Annahme sprachgetriebener Lehrbuchdivergenz und etabliert ΔLDS als Kernmetrik. Die erweiterte Humanvalidierung (N=15) falsifiziert ΔLDS > 0 unter Between-Subject-Bedingungen; ein LLM-as-Subject-Experiment (Within-Subject) weist das Sprachsignal dagegen klar nach und klassifiziert den Human-Negativebefund als Design-Artefakt — mit dem Sprachcode als dominanter Organisationsebene und dem kulturellen Rahmen als sekundärem, code-internem Modulator.

---

1. Einleitung

1.1 Motivation

Als zweisprachiger Schüler — aufgewachsen mit Chinesisch als Muttersprache, unterrichtet auf Deutsch und wissenschaftlich geprägt durch Englisch — ist mir immer wieder aufgefallen, dass dieselben Konzepte in verschiedenen Sprachen anders "gefühlte" Bedeutungskerne haben. Das chinesische Wort 「成功」(Chénggōng) betont Leistung durch Anstrengung und familiäre Erwartungen. Das deutsche "Erfolg" ist stärker karriere- und kompetenzorientiert. Das englische "Success" assoziiert Chancen und individuelle Wahlfreiheit.

Diese subjektive Beobachtung wirft eine tiefere Frage auf: Unterscheiden sich Sprachen nicht nur in Wörtern, sondern in der Art, wie sie Wissen organisieren?

Die linguistische Relativitätstheorie (Sapir-Whorf-Hypothese) postuliert genau das: Sprache beeinflusst das Denken. In den letzten zwei Jahrzehnten wurde dies empirisch für Farbwahrnehmung [51], Raumkonzepte [52] und Zeitwahrnehmung [53] belegt. Doch für abstrakte, komplexe Wissensdomänen wie Mathematik blieb diese Frage weitgehend unerforscht.

Hier setzt LinguaGraph an.

Die Relevanz dieser Frage reicht heute über die Sprachwissenschaft hinaus: Moderne KI-Systeme werden in Dutzenden Sprachen eingesetzt, aber überwiegend mit englischen Daten trainiert. Wenn ein Modell „Gerechtigkeit" in einem Kreditentscheidungs-System sprachabhängig anders strukturiert, erhalten Nutzer je nach Sprache inkonsistente Behandlung. Die Prüfung dieser sprachübergreifenden Konsistenz ist ein offenes Problem der KI-Validierung — gängige Evaluation misst Aufgabenerfüllung, nicht die Konzeptstruktur des Modells. Genau dafür stellt LinguaGraph ein quantitatives Werkzeug bereit.

1.2 Forschungsfrage

Die übergeordnete Forschungsfrage lautet:

Organisieren verschiedene Sprachen Wissen auf systematisch unterschiedliche Weise — und kann Künstliche Intelligenz diese Unterschiede messbar machen?

Daraus leiten sich drei Teilfragen ab:

1. Existiert ein messbarer "Linguistic Drift" zwischen kognitiven Graphen aus ZH-, EN- und DE-Texten?
2. Sind LLM-extrahierte Wissensgraphen ein valides Instrument, um sprachübergreifende Strukturunterschiede zu erfassen?
3. Ist der Linguistic Divergence Score (LDS) stabil und interpretierbar über verschiedene Themen und Sprachen hinweg?

1.3 Beiträge

Diese Arbeit leistet folgende Beiträge:

Linguistic Divergence Score (LDS) — Eine neuartige graphentheoretische Metrik, die die strukturelle Divergenz zwischen sprachspezifischen Wissensgraphen quantifiziert. LDS = 1 — GraphSimilarity, wobei Ähnlichkeit über gemeinsame Konzepte und Relationen gemessen wird.

Erster systematischer Vergleich mathematischer Wissensstrukturen über ZH/EN/DE hinweg — basierend auf 68 Lehrbüchern, 556 extrahierten Konzepten und 525 Relationen.

CognitiveSpace — Eine skalierbare 3D-Visualisierung, die Wissensstrukturen als konzentrische Kugelschalen darstellt. Vier Bildungsstufen (Grundschule bis Universität) sind farblich codiert und interaktiv filterbar nach Sprache.

End-to-End-Pipeline — Ein vollständiges System von der Lehrbuchtextextraktion über die sprachübergreifende Konzeptalignierung bis zur Graphanalyse und 3D-Visualisierung. Die Pipeline ist reproduzierbar und auf beliebige Wissensdomänen übertragbar.

KI-Validierungsinstrument — Anwendung des Frameworks als Audit-Werkzeug für mehrsprachige KI: Quantifizierung, ob und wo ein Modell wertbeladene Konzepte sprachabhängig strukturiert, mit interpretierbarem Divergenzbericht (LDS-C, divergierende Konzeptbestandteile, Domänen-Asymmetrie) für Entwickler, Regulierer und Forscher.

1.4 Gliederung

Die Arbeit ist wie folgt aufgebaut. Kapitel 2 gibt einen Überblick über verwandte Arbeiten aus Linguistik, KI-Forschung und Wissensgraphen. Kapitel 3 beschreibt die Methodik beider Pipelines. Kapitel 4 präsentiert die Ergebnisse — vom CognitiveSpace-Wissensgraphen über die LDS-Analyse bis zur Humanvalidierung (F1 – F12). Kapitel 5 weist das Sprachsignal im LLM-as-Subject-Within-Subject-Design nach, Kapitel 6 validiert fächerübergreifend (Physik, Chemie) und exploriert die Sprachproduktion (LPA). Kapitel 8 diskutiert Ergebnisse, Limitationen und die methodologische Einordnung; Kapitel 9 fasst zusammen und formuliert die Audit-Anwendung.

2. Verwandte Arbeiten

Dieser Abschnitt verortet LinguaGraph in vier Forschungssträngen: pädagogische Wissensgraphen, Curriculumanalyse und -vergleich, Concept Mapping und kognitive Struktur sowie sprachübergreifende Wissensintegration.

2.1 Pädagogische Wissensgraphen

Die Konstruktion von Wissensgraphen aus Bildungsressourcen hat in den letzten Jahren bedeutende Fortschritte erfahren. Yao et al. (2019) schlugen Joint Embedding Learning für pädagogische Wissensgraphen vor und zeigten, dass Konzeptbeziehungen automatisch aus Curriculumsdokumenten abgeleitet werden können [1]. Zhao, Sun und Xu (2022) entwickelten EDUKG, einen heterogenen K-12-Bildungswissensgraphen, der mehrere Fächer umfasst, und demonstrierten die Machbarkeit einer groß angelegten Wissensorganisation auf Curriculumniveau [2].

Am relevantesten für LinguaGraph ist die Arbeit von Ain, Chatti und Qussa (2025), die eine optimierte Pipeline zur automatischen Konstruktion pädagogischer Wissensgraphen entwickelten [3]. Ihr Ansatz verwendet Large Language Models zur Konzeptextraktion, ähnlich unserer MIMO-Prompt-Methodik. Eine Folgestudie verglich Top-Down- und Bottom-Up-Konstruktionsansätze und stellte fest, dass die Bottom-Up-Extraktion aus Lehrbuchinhalten eine vollständigere Konzeptabdeckung liefert [4]. LinguaGraph übernimmt einen Bottom-Up-Ansatz, erweitert ihn jedoch um sprachübergreifende Ausrichtung — eine Dimension, die in bestehenden EKG-Pipelines fehlt.

Alatrash, Chatti und Wibowo (2025) befassten sich speziell mit der Inferenz von Voraussetzungsbeziehungen in pädagogischen Wissensgraphen und schlugen einen multikriteriellen Ansatz vor, der textuelle, strukturelle und taxonomische Signale berücksichtigt [5]. Ihre Methode fließt unmittelbar in die HDS-Metrik unseres Frameworks ein, die die Tiefe von Voraussetzungsketten als Proxy für die Wissenshierarchie misst.

2.2 Curriculumanalyse und -vergleich

Der länderübergreifende Curriculumsvergleich ist ein Eckpfeiler der internationalen Bildungsforschung. Das TIMSS-Rahmenwerk (Trends in International Mathematics and Science Study) bietet eine systematische Methodik zum Vergleich von Curricula über Länder hinweg, einschließlich der Dimensionen Curriculumintention, -implementierung und -ergebnis [6]. Die OECD-PISA-Studien analysieren in ähnlicher Weise, wie Curriculumsstrukturen Lernergebnisse beeinflussen [7].

Spezifische Vergleiche zwischen deutschen und chinesischen Mathematikcurricula wurden von Liang und Heckmann (2013) durchgeführt, die Lehrbuchaufgaben analysierten und systematische Unterschiede in Aufgabenkomplexität und Darstellungsstil feststellten [8]. Fan, Zhu und Miao (2013) erweiterten dies auf einen breiteren länderübergreifenden Vergleich von Lehrbuchaufgaben [9]. Während diese Studien sich auf Aufgaben- oder Inhaltsebene konzentrieren, führt LinguaGraph einen Graphenvergleich ein — den Language Drift Score (LDS) — der strukturelle Unterschiede erfasst, die eine oberflächliche Inhaltsanalyse übersieht.

Der kürzlich veröffentlichte Kernlehrplan NRW (2019, 2023) für das deutsche Gymnasium im Fach Mathematik bietet eine formale kompetenzbasierte Curriculumsstruktur [10]. Das chinesische Pendant, der Yiwu Jiaoyu Shuxue Kecheng Biaozhun (2022) und der Putong Gaozhong Shuxue Kecheng Biaozhun (2017), definieren in ähnlicher Weise Lernprogressionen und Inhaltsstandards [11]. LinguaGraph ist nach unserem Kenntnisstand das erste System, das diese Curriculumsstandards in strukturierte Wissensgraphen für den direkten sprachübergreifenden Vergleich überführt.

Zur institutionellen Einordnung stützen wir uns auf öffentlich dokumentierte Governance-Unterschiede (als Kontext, nicht als getestete Ursache): Das deutsche System ist föderal organisiert — primäre Zuständigkeit der Länder, koordiniert über die KMK, Bundesrolle via BMBF [28]; das chinesische System ist hochzentralisiert unter dem MoE mit nationaler Lehrbuchzulassung und Prüfungskopplung (Gaokao) [29]. Die TIMSS-2023-Enzyklopädie dokumentiert länderspezifische Curriculumspolitiken systematisch [30]; die TIMSS-2023-Insights-Reihe untersucht explizit, inwieweit curriculare Vorgaben im Unterricht umgesetzt werden und wie dies mit Leistung zusammenhängt [31]; OECD Education at a Glance 2025 liefert vergleichbare Systemkennzahlen (u. a. Deutschland: 4,4 % des BIP für Bildung) [32]. Diese Quellen stützen die Governance-Hypothese (F10) als institutionellen Kontext; die Unterrichts-Implementierungskette bleibt ungetestet.

2.3 Concept Mapping und Wissensorganisation

Die theoretische Grundlage der Wissensstrukturanalyse geht auf Ausubels Assimilationstheorie des sinnvollen Lernens (1963) zurück, die argumentiert, dass Wissen hierarchisch und nicht als isolierte Fakten organisiert ist [12]. Novak und Cañas (2008) operationalisierten diese Theorie durch Concept Mapping und zeigten, dass Wissensstrukturen als propositionale Netzwerke externalisiert werden können [13]. Ihr Rahmenwerk liegt den CDS- und HDS-Metriken in LinguaGraph zugrunde: Der Concept Density Score erfasst die Vernetztheit von Konzepten innerhalb einer Wissensdomäne, während der Hierarchy Depth Score die Tiefe der Voraussetzungsketten misst.

Jüngste Fortschritte haben die automatische Generierung von Concept Maps ermöglicht. Schwach überwachte Ansätze mittels Graphtranslation (2021) [14] und generative LLM-basierte Methoden (2025) [15] haben groß angelegtes Concept Mapping praktikabel gemacht. LinguaGraph unterscheidet sich von diesen Ansätzen durch seine mehrsprachige Ausrichtung: Anstatt Concept Maps für eine einzelne Sprache zu generieren, konstruieren wir parallele Wissensgraphen für Chinesisch, Deutsch und Englisch und ermöglichen so einen sprachübergreifenden Strukturvergleich.

2.4 Sprachübergreifende Wissensgraph-Ausrichtung

Die sprachübergreifende Wissensgraph-Ausrichtung zielt darauf ab, äquivalente Entitäten und Relationen sprachübergreifend zu identifizieren. Frühe Arbeiten von McGillivray et al. (2016) schlugen multilinguale Wissensgraph-Embeddings für die sprachübergreifende Ausrichtung vor [16]. Nachfolgende Forschung entwickelte Methoden zur Entitätsausrichtung mittels adversarialem Training [17], Subgraph-Netzwerken [18] und Co-Training mit Entitätsbeschreibungen [19].

Diese ausrichtungsfokussierten Ansätze unterscheiden sich grundlegend von LinguaGraphs Zielsetzung. Die Ausrichtungsforschung fragt: Wie lassen sich äquivalente Konzepte sprachübergreifend abgleichen? LinguaGraph fragt: Welche strukturellen Unterschiede verbleiben bei gegebenen abgeglichenen Konzepten zwischen sprachspezifischen Wissensorganisationen? Die LDS-Metrik erfasst diese verbleibenden strukturellen Unterschiede — die nach der Konzeptausrichtung fortbestehende Divergenz — die von der bestehenden sprachübergreifenden KG-Forschung nicht systematisch quantifiziert wurde.

Die systematische Übersichtsarbeit von Chen et al. (2026) zu cross-lingualem Transfer für Knowledge-Graph-Akquisition bestätigt, dass sprachbias-bedingte Alignierungsschwierigkeiten und geringe Transfereffizienz fortbestehende Kernherausforderungen sind — trotz mehrsprachiger Embeddings und LLMs [25]. Dies verortet unsere P2-Recheck-Einschränkung (Alignierungs-Labels als partielle Artefaktquelle, s. §8.14.5) als bekanntes Domänenproblem, nicht als Einzelfall. Han et al. (unter Begutachtung) schlagen komplementär eine Transfer-Lokalisierungs-Ebene vor: Universelles Wissen soll konvergieren, kulturell situiertes Wissen soll divergieren — hohe Transferleistung um den Preis kultureller Auslöschung („cultural erasure") ist die unerwünschte Quadrant [26]. Diese Unterscheidung rahmt unseren Zentralbefund (institutionelle Konvergenz vs. kulturelle Divergenz) als erwartbares Muster statt als Anomalie. Dass Mehrsprachigkeit nicht Multikulturalität impliziert, zeigen Wang et al. (2025) empirisch über WVS-Opinionverteilungen in vier Sprachen [27].

2.5 Multilinguale Wertausrichtung von LLMs

Ein wachsender Forschungsstrang untersucht, ob Large Language Models (LLMs) menschliche Werte sprachübergreifend konsistent vertreten. WorldValuesBench (Zhao et al., 2024) etablierte eine groß angelegte Benchmark zur Vorhersage kultureller Wertantworten aus dem World Values Survey [20]. Xu et al. (2024) zeigten, dass Wertkonzepte in LLMs über 16 Sprachen hinweg als lineare Richtungen im Repräsentationsraum darstellbar sind [21]. Agarwal et al. (2024) wiesen nach, dass die moralische Bewertung von LLMs von der Prompt-Sprache abhängt [22]. Farid et al. (2025) deckten über die Übersetzung zweier Moral-Benchmarks in fünf Sprachen systematische kreuzlinguistische Fehlausrichtungen auf [23]; Lee et al. (2026) schlugen mit MET ein theoriebasiertes, kulturbewusstes mehrsprachiges Moral-Entscheidungsbenchmark vor [24].

Gemeinsam messen diese Arbeiten die Auswahl von Werten über Sprachen hinweg — durch Vorhersage, Klassifikation oder Rating. LinguaGraph ergänzt diese Perspektive um ein strukturelles Maß: Während bestehende multilinguale Wert- und Moral-Benchmarks (WorldValuesBench, 2024; One Model, Many Morals, 2025; MET, 2026) die Auswahl von Werten über Sprachen hinweg messen, erfasst LinguaGraph mit dem LDS die strukturelle Organisation von Wertkonzepten (Knoten + Kanten) — ein komplementäres, bisher ungemessenes Maß. Das LLM-as-Subject-Design (Within-Subject, §5) isoliert dabei die Sprache als einzige Variable; die Replikation über 55 Messungen (§5.10) zeigt, dass die strukturelle Divergenz von Wertkonzepten über Sprachpaare hinweg breit reproduzierbar und kulturell gerichtet ist (DE Autonomie/Regeln vs. ZH Raum/Anspruch). Keine der bestehenden Arbeiten misst diese strukturelle Divergenz systematisch.

Dass englisch-haltige Paare schwächere Signale tragen (§5.10: alle 8 nicht-signifikanten Tests betreffen ZH-EN/DE-EN), ist mit der Englisch-Zentriertheit aktueller Modelle konsistent: MEXA evaluiert gezielt englisch-zentrierte LLMs über cross-linguale Alignierung [43]; im ENZH-Transfer überwiegt negativer Transfer vom Englischen aufs Chinesische [44]; cross-linguale faktische Inkonsistenz ist ein dokumentierter Mechanismus multilingualer Modelle [45]. Englisch wirkt als Trainings-Hub mit generalisierteren, weniger differenzierten Konzeptrepräsentationen — schwächere EN-Paardivergenz ist damit theoretisch erwartbar, kein Gegenbefund.

Zur Designfrage (Between- vs. Within-Subject, §4 vs. §5): Within-Subject-Designs erreichen bei gleicher Effektstärke mit deutlich kleineren Stichproben diagnostische Power als Between-Subject-Designs [40][41] — die menschliche N=15-Null unter Between-Subject ist daher auch power-analytisch erwartbar (s. §8.10), kein Beleg für Abwesenheit.

Als Hintergrund — nicht als Evidenz dieser Studie — ist dokumentiert, dass KI-generierte Texte messbare linguistische Marker tragen: Fu und Yang (2025) vergleichen maschinell identifizierte vs. menschlich inferierte Prädiktoren [33]; RoBERTa-basierte Detektion erreicht 96,1 % Genauigkeit mit interpretierbaren stilistischen Unterschieden (LIME/SHAP) [34]; eine Großstudie über 27 LLMs × 10 Domänen systematisiert 284 interpretierbare Merkmale und deren domänenübergreifende Generalisierung [35]; ein Survey synthetisiert die verstreuten Befunde zu AIGT-vs-HWT-Merkmalen [36]. Diese Literatur stützt die allgemeine Beobachtung, dass KI-Texte systematische Stilmerkmale aufweisen — sie ersetzt keine eigene Messung im Chinesischen und wird hier nicht als Befund beansprucht.

2.6 Forschungslücke

Während jeder dieser Forschungsstränge unabhängig Aspekte der Wissensstrukturanalyse untersucht hat, integriert keine bestehende Arbeit:

1. Automatische Wissensgraphkonstruktion aus mehrsprachigen Lehrbuchinhalten
2. Systematischer Vergleich von Wissensstrukturen über Sprachen hinweg
3. Quantitative Metriken (LDS, CDS, HDS) für den Strukturvergleich
4. Curriculumsbezogene Analyse, die Lehrbuchwissensgraphen mit offiziellen Curriculumsstandards vergleicht
5. Strukturelle (statt auswahlbasierte) Messung der Wertdivergenz mehrsprachiger LLMs — die Trennung von Konzeptauswahl und Konzeptorganisation über Sprachen hinweg

LinguaGraph schließt diese Lücke durch eine integrierte Pipeline von der Lehrbuchextraktion bis zur sprachübergreifenden Strukturanalyse, mit validierten Metriken, die Unterschiede in der Wissensorganisation auf mehreren Ebenen erfassen — über Sprachen (LDS), Bildungsstufen (CDS) und hierarchische Tiefe (HDS) hinweg.

Literaturverzeichnis

[1] Yao, S., Wang, R., & Sun, S. (2019). Joint Embedding Learning of Educational Knowledge Graphs. arXiv:1911.08776.

[2] Zhao, B., Sun, J., & Xu, B. (2022). EDUKG: a Heterogeneous Sustainable K-12 Educational Knowledge Graph. arXiv:2210.12228.

[3] Ain, Q. U., Chatti, M. A., & Qussa, J. (2025). An Optimized Pipeline for Automatic Educational Knowledge Graph Construction. arXiv:2509.05392.

[4] Ain, Q. U., Chatti, M. A., & Shakhshir, A. (2025). Top-Down vs. Bottom-Up Approaches for Automatic EKG Construction. arXiv:2505.10069.

[5] Alatrash, R., Chatti, M. A., & Wibowo, N. (2025). Inferring Prerequisite Knowledge Concepts in EKGs. arXiv:2509.05393.

[6] IEA. (1995 – 2023). Trends in International Mathematics and Science Study (TIMSS). International Association for the Evaluation of Educational Achievement.

[7] OECD. (Various). PISA Curriculum Analysis and Education at a Glance. OECD Publishing.

[8] Liang, S., & Heckmann, K. (2013). A comparison of German and Chinese mathematics textbooks. ZDM, 45(5), 743 – 756.

[9] Fan, L., Zhu, Y., & Miao, Z. (2013). A comparative study on mathematics textbook problems across countries. ICMT.

[10] MSB NRW. (2019; Neufassung 2023). Kernlehrplan Mathematik für das Gymnasium (Sek I). Ministerium für Schule und Bildung NRW.

[11] MoE China. (2017; 2022). Putong Gaozhong Shuxue Kecheng Biaozhun (2017) / Yiwu Jiaoyu Shuxue Kecheng Biaozhun (2022) [Senior High / Compulsory Education Mathematics Curriculum Standards]. Ministry of Education, PRC.

[12] Ausubel, D. P. (1963). The Psychology of Meaningful Verbal Learning. Grune & Stratton.

[13] Novak, J. D., & Cañas, A. J. (2008). The Theory Underlying Concept Maps. IHMC.

[14] Lu, J., Dong, X., & Yang, C. (2023). Weakly Supervised Concept Map Generation through Task-Guided Graph Translation (GT-D2G). IEEE Transactions on Knowledge and Data Engineering, 35(10), 10871 – 10883. https://doi.org/10.1109/tkde.2023.3252588.

[15] Zhai, X. (2025). Generative Large Language Models for Knowledge Representation: A Systematic Review of Concept Map Generation. arXiv:2509.14554.

[16] Chen, M., Tian, Y., Yang, M., & Zaniolo, C. (2017). Multilingual Knowledge Graph Embeddings for Cross-lingual Knowledge Alignment (MTransE). Proc. IJCAI 2017, 1511 – 1517. https://doi.org/10.24963/ijcai.2017/209.

[17] Trisedya, B. D., Qi, J., & Zhang, R. (2019). Entity Alignment between Knowledge Graphs Using Attribute Embeddings. Proc. AAAI 2019, 33(01), 297 – 304.

[18] Yu, S., Zhang, S., Zhang, J., Zhou, J., Xuan, Q., Li, B., & Hu, X. (2022). SubGraph Networks based Entity Alignment for Cross-lingual Knowledge Graph. arXiv:2205.03557.

[19] Chen, M., Tian, Y., Chang, K.-W., Skiena, S., & Zaniolo, C. (2018). Co-training Embeddings of Knowledge Graphs and Entity Descriptions for Cross-lingual Entity Alignment. Proc. IJCAI 2018, 3998 – 4004. arXiv:1806.06478.

[20] Zhao, W., Mondal, D., Tandon, N., et al. (2024). WorldValuesBench: A Large-Scale Benchmark Dataset for Multi-Cultural Value Awareness of Language Models. arXiv:2404.16308.

[21] Xu, S., Dong, W., Guo, Z., et al. (2024). Exploring Multilingual Concepts of Human Value in Large Language Models: Is Value Alignment Consistent, Transferable and Controllable across Languages? arXiv:2402.18120.

[22] Agarwal, U., Tanmay, K., Khandelwal, A., et al. (2024). Ethical Reasoning and Moral Value Alignment of LLMs Depend on the Language we Prompt them in. arXiv:2404.18460.

[23] Farid, S., Lin, J., Chen, Z., et al. (2025). One Model, Many Morals: Uncovering Cross-Linguistic Misalignments in Computational Moral Reasoning. arXiv:2509.21443.

[24] Lee, A., Kwon, R., Zhang, Y., et al. (2026). MET: Theory-Grounded and Culture-Aware Multilingual Moral Reasoning. arXiv:2607.11736.

[25] Chen, W.-L., Zhou, K.-Q., Sarkheyli-Hägele, A., et al. (2026). Cross-lingual transfer learning for knowledge graph acquisition: Paradigms, resources and challenges. Expert Systems with Applications, 303, 130434. https://doi.org/10.1016/j.eswa.2025.130434.

[26] Han, H., Agrawal, S., & Briakou, E. (unter Begutachtung). Rethinking Cross-lingual Alignment: Balancing Transfer and Cultural Erasure in Multilingual LLMs. arXiv:2510.26024.

[27] Wang, Y., et al. (2025). Multilingual != Multicultural: Evaluating Gaps Between Multilingual Capabilities and Cultural Alignment in LLMs. arXiv:2502.16534.

[28] KMK / BMBF. (o. J.). Ständige Konferenz der Kultusminister der Länder; Rahmenrolle des BMBF. https://www.kmk.org/en/index.html; Eurydice Germany. (Zugriff Sept. 2026.)

[29] Ministry of Education of the PRC. (o. J.). Nationale Lehrbuchzulassung und Gaokao-Kopplung. http://en.moe.gov.cn/. (Zugriff Sept. 2026.)

[30] IEA. (2023). TIMSS 2023 Encyclopedia: Education Policy and Curriculum in Mathematics and Science. https://timss2023.org/encyclopedia/.

[31] TIMSS & PIRLS International Study Center. (2024). TIMSS 2023 Insights: Curriculum Alignment report (curricular specifications → classroom implementation → achievement). https://timss.bc.edu/latest-news/timss-2023-insights-curriculum-alignment.html. (Zugriff Sept. 2026.)

[32] OECD. (2025). Education at a Glance 2025 (Germany profile: 4.4% of GDP). OECD Publishing. https://www.oecd.org/en/publications/education-at-a-glance-2025_1a3543e2-en/germany_fa91d155-en.html.

[33] Fu, K., & Yang, X. (2025). Linguistic Markers of AI-Generated Text: A Comparative Analysis of Machine-Identified and Human-Inferred Predictors. AMCIS 2025 TREOs. https://aisel.aisnet.org/treos_amcis2025/1.

[34] Masih, A., Afzal, B., et al. (2025). Classifying human vs. AI text with machine learning and explainable transformer models. Scientific Reports, 15, 43310. https://www.nature.com/articles/s41598-025-27377-z.

[35] El Attar, Y., Dönmez, E., Maurer, M., & Falenska, A. (2026). A Systematic Analysis of Linguistic Features in AI-Generated Text Detection Across Domains and Models (27 LLMs × 10 domains, 284 features). arXiv:2606.04177.

[36] Teron, L., & Dobrovoljc, K. (2025). Linguistic Characteristics of AI-Generated Text: A Survey. arXiv:2510.05136.

[37] Markus, H. R., & Kitayama, S. (1991). Culture and the self: Implications for cognition, emotion, and motivation. Psychological Review, 98(2), 224–253.

[38] Hofstede, G. (2001). Culture's Consequences: Comparing Values, Behaviors, Institutions and Organizations Across Nations (2nd ed.). Sage. (Original work published 1980.)

[39] Nisbett, R. E. (2003). The Geography of Thought: How Asians and Westerners Think Differently … and Why. Free Press.

[40] Bellemare, C., Bissonnette, L., & Kröger, S. (2014). Statistical power of within- and between-subjects designs in economic experiments. IZA Discussion Paper No. 8583. https://docs.iza.org/dp8583.pdf.

[41] Seltman, H. J. (2018). Experimental Design and Analysis, Chapter 14: Within-Subjects Designs. Carnegie Mellon University. https://www.stat.cmu.edu/~hseltman/309/Book/chapter14.pdf.

[42] Sukiennik, N., Gao, C., Xu, F., & Li, Y. (2025). An Evaluation of Cultural Value Alignment in LLM. arXiv:2504.08863.

[43] Kargaran, A. H., Modarressi, A., Nikeghbal, N., Diesner, J., Yvon, F., & Schuetze, H. (2025). MEXA: Multilingual Evaluation of English-Centric LLMs via Cross-Lingual Alignment. Findings of ACL 2025, 27001–27023. https://aclanthology.org/2025.findings-acl.1385/.

[44] Zhou, Y., & Matusevych, Y. (2025). Curse of bilinguality: Evaluating monolingual and bilingual language models on Chinese linguistic benchmarks. Proc. GEM 2025. https://aclanthology.org/2025.gem-1.58/.

[45] Wang, M., Adel, H., Lange, L., Liu, Y., Nie, E., Strötgen, J., & Schuetze, H. (2025). Lost in Multilinguality: Dissecting Cross-lingual Factual Inconsistency in Transformer Language Models. Proc. ACL 2025, 5075–5094. https://aclanthology.org/2025.acl-long.253/.

[46] Fluss, R., Faraggi, D., & Reiser, B. (2005). Estimation of the Youden Index and its Associated Cutoff Point. Biometrical Journal, 47(4), 458–472. https://doi.org/10.1002/bimj.200410135.

[47] Binz, M., & Schulz, E. (2023). Using cognitive psychology to understand GPT-3. PNAS, 120(6), e2218523120. https://doi.org/10.1073/pnas.2218523120.

[48] Hagendorff, T., Fabi, S., & Kosinski, M. (2023). Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in ChatGPT. Nature Machine Intelligence, 5, 1163–1176. https://doi.org/10.1038/s42256-023-00740-0.

[49] Whorf, B. L. (1956). Language, Thought, and Reality. MIT Press.

[50] Lucy, J. A. (1997). Linguistic relativity. Annual Review of Anthropology, 26, 291–312.

[51] Winawer, J., Witthoft, N., Frank, M. C., Wu, L., Wade, A. R., & Boroditsky, L. (2007). Russian blues reveal effects of language on color discrimination. PNAS, 104(19), 7780–7785.

[52] Levinson, S. C. (1996). Language and space. Annual Review of Anthropology, 25, 353–382.

[53] Boroditsky, L. (2001). Does language shape thought? Mandarin and English speakers' conceptions of time. Cognitive Psychology, 43(2), 1–22.

[54] Schmidt, W. H., et al. (2001). Why Schools Matter: A Cross-National Comparison of Curriculum and Learning. Jossey-Bass.

LinguaGraph — Methodology

Language: German Status: Final (v0.13.3)

2. Methodik

2.1 Überblick

LinguaGraph besteht aus zwei komplementären Analyse-Pipelines:

Pipeline A (Kognitive Graphen): Probandentexte → LLM Extraktion → Konzeptgraph → LDS → Sprachenvergleich Pipeline B (Lehrbuch-Wissensgraph): Lehrbuchkorpus → MIMO Extraktion → Alignierung → CognitiveSpace 3D

Pipeline A dient der kognitiven Analyse auf individueller Ebene (Probandenstudie). Pipeline B dient der Validierung der Extraktions- und Alignierungsmethodik im großen Maßstab — 68 Lehrbücher, 556 Konzepte, 525 Relationen.

2.2 Lehrbuchkorpus (Pipeline B)

Die Grundlage des CognitiveSpace-Wissensgraphen bildet ein Korpus von 68 Lehrbüchern aus drei Sprachräumen:

Sprache	Anzahl	Lehrwerke	Stufen
Chinesisch	39	Volksverlag (Renjiao) K-12, Tongji Analysis, Lineare Algebra, Wahrscheinlichkeit	Grundschule bis Universität
Englisch	18	Stewart Calculus, MIT OCW, Khan Academy, IGCSE, IB	K-12 bis Universität
Deutsch	11	Forster Analysis, Fischer LA, Lambacher Schweizer, Papula	Sekundarstufe bis Universität

Die Lehrbücher decken ein breites mathematisches Spektrum ab: Arithmetik, Algebra, Geometrie, Analysis, Lineare Algebra, Differentialgleichungen, Wahrscheinlichkeitstheorie und Statistik.

2.3 Konzeptextraktion (MIMO-Verfahren)

Die Extraktion mathematischer Konzepte und ihrer Relationen erfolgte mittels eines strukturierten LLM-Prompts ("MIMO"-Verfahren):

[System] Extrahiere alle mathematischen Konzepte aus dem folgenden Lehrbuchabschnitt. Für jedes Konzept: - id: eindeutiger Bezeichner - name: kanonischer Name - type: Konzepttyp (definition, theorem, method, example) [Relationen] Für jedes Paar verwandter Konzepte: - source: Quellkonzept - target: Zielkonzept - type: Beziehungstyp (depends_on, part_of, related_to, representation, prerequisite) - evidence: Textbeleg aus dem Lehrbuch

Die Extraktion wurde für jedes der 68 Lehrbücher separat durchgeführt, was 75 JSON-Extraktionsdateien ergab (einige Lehrbücher wurden aufgrund ihres Umfangs in Kapitel aufgeteilt).

2.4 Graphkonstruktion und -fusion

Die Roh-Extraktionen durchlaufen einen mehrstufigen Fusionsprozess:

Step 1 — Merging: Alle 75 Extraktionsdateien werden eingelesen und zu einem einheitlichen Graphen zusammengeführt. Aliase und Synonyme werden anhand einer Konfigurationsdatei (concept_taxonomy.json) normalisiert:

"微分" → "Differential" "导数" → "Ableitung" (Derivative) "极限" → "Grenzwert" (Limit)

Step 2 — Deduplizierung: Konzepte mit identischer ID oder nachgewiesener Synonymie werden fusioniert. Nach der Deduplizierung verbleiben 556 eindeutige Konzepte.

Step 3 — Relationsextraktion: Aus den extrahierten Abhängigkeiten wird ein gerichteter Graph konstruiert. Zusätzlich zu den 525 direkt extrahierten Relationen werden ~3000 transitive Inferenzen zur Sicherstellung der Konnektivität hinzugefügt.

2.5 Sprachübergreifende Alignierung

Die Alignierung der Konzepte über ZH/EN/DE hinweg erfolgt über ein gemeinsames ID-Schema mit 30 geteilten Konzept-IDs:

{ "shared_id": "math_calculus_derivative", "zh": { "id": "math_calculus_导数", "name": "导数" }, "en": { "id": "math_calculus_derivative", "name": "Derivative" }, "de": { "id": "math_calculus_ableitung", "name": "Ableitung" } }

Die Alignierung wird durch zwei Strategien erreicht:

1. Explizite Abbildung: Lehrbücher, die dasselbe mathematische Konzept behandeln, werden über Kapitelverweise gemappt (z.B. Stewart Kapitel 2 ↔ Forster Kapitel 4 ↔ Renjiao选修2-2).
2. Semantische Inferenz: Konzepte mit übereinstimmenden Nachbarschaftsbeziehungen im Graphen werden als äquivalent betrachtet.

Ergebnis: 219 Konzeptgruppen (39 % von 556 Konzepten) sind dreisprachig vollständig abgedeckt, 120 sind nur im

Chinesischen vorhanden (21,6 %), der Rest in zwei von drei Sprachen.

2.6 CognitiveSpace 3D-Visualisierung

Die Visualisierung erfolgt als interaktiver 3D-Graph mittels der Bibliothek 3d-force-graph (v1.80.0).

Layout: Konzentrische Kugelschalen nach Bildungsstufe:

Stufe	Konzepte	Radius
Grundschule (小学)	26	r = 0 – 50
Mittelschule (初中)	57	r = 60 – 120
Oberstufe (高中)	200	r = 140 – 230
Universität (大学)	273	r = 250 – 320

Deterministische Positionierung: Die Position jedes Knotens wird per deterministischer Hash-Funktion aus seiner ID berechnet:

def position(knoten_id, min_r, max_r): h = hash(knoten_id) r = min_r + (h % 1001) / 1000 * (max_r - min_r) theta = (h % 997) / 997 * 2π phi = acos(2 * (((h * 13 + 7) % 1000) / 1000) - 1) return (r * sin(phi) * cos(theta), r * sin(phi) * sin(theta), r * cos(phi))

Diese deterministische Positionierung garantiert, dass jedes Konzept bei jedem Laden der Visualisierung an derselben Stelle erscheint — essentiell für die Reproduzierbarkeit in der Forschung.

Farbcodierung: Jeder Bildungsstufe ist eine eigene Farbe zugeordnet: - Grundschule: Grün (#4ade80) - Mittelschule: Cyan (#22d3ee) - Oberstufe: Blau (#60a5fa) - Universität: Violett (#c084fc)

Interaktion: Der Graph unterstützt Sprachfilterung (ZH/EN/DE), drei Ansichtsmodi (Universe, Space-Fill, Compare), Knotendetailansicht mit Lehrbuchquellen und einen BFS-Ripple-Effekt bei Klick.

2.7 Linguistic Divergence Score (LDS)

Der Linguistic Divergence Score quantifiziert die strukturelle Divergenz zwischen zwei sprachspezifischen kognitiven Graphen. Die hier verwendete frozen v3-Formel kombiniert Knoten- und Kantenkomponente:

LDS_v3(A, B) = 1 - mean(J(node_set_A, node_set_B), J(edge_set_A, edge_set_B))

wobei (J(X,Y) = \frac{|X \cap Y|}{|X \cup Y|}) der Jaccard-Koeffizient ist und A, B zwei Sprachversionen desselben Inhalts (Textbuchkapitel, Fragebogenantwort oder Wikipedia-Artikel) bezeichnen. Ein LDS von 0 bedeutet identische Knoten- und Kantenstruktur; ein LDS nahe 1 maximale Divergenz. Die drei Sprachpaare ZH-EN, ZH-DE, DE-EN werden auf drei Analyseebenen berechnet: Konzeptebene (nur Knoten), relationale Ebene (Knoten + Kante, v3) und Kategorienebene (6-Klassen-Codebook).

Varianten: - LDS-K (Knowledge): Strukturelle Divergenz institutionellen Wissens (Lehrbuch-Konzeptgraphen). - LDS-C (Cognitive): Divergenz spontaner kognitiver Ausdrucksweise (menschliche oder LLM-Konzeptgraphen). - ΔLDS = LDS-C LDS-K: der zusätzliche, durch kognitive Ausdrucksweise eingeführte Anteil jenseits der Lehrbuchstruktur.

Null-Modell-Rahmen: Um Struktureffekte von Spracheffekten zu trennen, werden drei Null-Modelle eingesetzt (Details in docs/lds_formal_definition.md §3): (1) Within-Language Split-Half — Aufteilung einer Sprache in zwei Hälften → Bodenniveau der Teilnehmervariabilität; (2) Label-Permutation — Permutation der Sprachlabels → formales p-Wert für das getragene Sprachsignal; (3) Cross-Source Null — Textbuch vs. Wikipedia derselben Sprache → Trennung von Quell- und Spracheffekt. Alle LDS-Berechnungen verwenden die frozen v3-Formel; die Ergebnisse der Human- und LLM-Analyse (§4, §5) stützen sich auf denselben Rahmen.

2.8 LLM-Extraktionsqualität

Zur Validierung der Extraktionsqualität wird ein mehrsprachiger Goldstandard verwendet. Der Datensatz umfasst insgesamt 92 manuell annotierte Antworten (36 ZH, 29 DE, 27 EN), verteilt auf zwei Domänen:

1. Mathematische Konzepte (20 Label): Calculus-Grundbegriffe (Grenzwert, Ableitung, Integral) — zur Validierung der domänenspezifischen Extraktionsqualität.
2. Soziale Konzepte (72 Label): Antworten zu Freiheit, Gerechtigkeit, Erfolg, Verantwortung und Heimat — zur Validierung im Hauptdomän der Studie.

Die Extraktion erfolgt mit qwen-plus (Alibaba Cloud Bailian API). Ergebnisse:

Domne	Sprache	F1	Precision	Recall	n
Mathematik	Chinesisch	0,857	1,000	0,798	7
Mathematik	Deutsch	0,506	0,536	0,512	7
Mathematik	Englisch	0,711	0,722	0,722	6
Sozial	Chinesisch	0,974	1,000	0,950	29
Sozial	Deutsch	0,949	0,959	0,941	22
Sozial	Englisch	0,882	0,914	0,857	21
Gesamt (92, gewichtet)	Alle	0,881	—	—	92

Gesamt-F1 ist das domänengewichtete Mittel ((72×0,939+20×0,674)/92≈0,881); die Kopfzahl 0,939 gilt nur für die Sozial-Subgruppe. Gesamt-Precision/Recall werden nicht aggregiert (domänenspezifisch, s. Zeilen oben).

Die Extraktionsqualität für soziale Konzepte übertrifft die mathematische Domäne deutlich: alle drei Sprachen erreichen F1 ⩾ 0,88, mit chinesischen (F1=0,974) und deutschen (F1=0,949) Ergebnissen, die das Qualitätsziel (F1 ⩾ 0,70) weit übertreffen. Dies bestätigt, dass die zuvor beobachtete niedrige deutsche Extraktionsqualität (F1=0,506) domänenspezifisch war und nicht die Modelleignung für die Hauptstudie widerspiegelt.

2.9 Model Comparison

Um zu bestimmen, ob die Extraktionsqualität durch die Pipeline oder die Modellfähigkeit begrenzt ist, vergleichen wir mehrere Modelle auf denselben Goldlabels. Die Tabelle zeigt Ergebnisse für die soziale Konzeptdomäne (72 Label) und die mathematische Domäne (20 Label):

Model	Domne	ZH F1	DE F1	EN F1
qwen3-8B (lokal)	Mathematik	0,857	0,506	0,711
qwen-plus (API)	Mathematik	0,952	0,489	0,778
qwen3.7-max (API)	Mathematik	0,980	0,551	0,778
qwen-plus (API)	Sozial	0,974	0,949	0,882

Die Ergebnisse zeigen einen entscheidenden Befund: Die Extraktionsqualität ist domänenabhängig. Während qwen-plus in der mathematischen Domäne lediglich DE F1=0,489 erreicht, steigt der Wert für soziale Konzepte auf DE F1=0,949. Dies liegt vermutlich an der unterschiedlichen Konzeptstruktur: Mathematische Konzepte sind präziser und domänenspezifischer, während soziale Konzepte alltagssprachlich näher an der Trainingsdistribution der Modelle liegen. Für die Hauptstudie (soziale Konzepte) ist die Extraktionsqualität in allen drei Sprachen als hoch einzustufen.

2.10 Curriculum Coverage Score (CS)

Um die Beziehung zwischen Lehrbuchinhalten und offiziellen Lehrplänen zu quantifizieren, definieren wir den Coverage Score (CS):

$$ CS(G_{textbook}, G_{curriculum}) = \frac{|V_{textbook} \cap V_{curriculum}|}{|V_{curriculum}|} $$

Der Coverage Score misst den Anteil der vom Lehrplan geforderten Konzepte, die im Lehrbuchgraph abgedeckt sind. Die Berechnung erfolgt pro Bildungsstufe und sprachspezifisch.

Aktuelle Ergebnisse für die mathematischen Lehrpläne (Stand 2026-08-08, scripts/compute_all_coverage_v2.py):

Lehrplan	Gesamt-Coverage	Hchste Stufe	Niedrigste Stufe
China (CN)	95,4 %	Grundstufe 1 – 2 (100 %)	Mittelstufe 7 – 9 (91,7 %)
England (UK)	37,3 %	KS2 Y6 (43,9 %)	KS4 (28,3 %)
Vereinigte Staaten (US)	17,2 %	alle Stufen (17,9 %)	—
NRW (DE)	12,7 %	Grundschule 1 – 4 (höchste)	Sekundarstufe II (niedrigste)

Der Coverage Score zeigt erhebliche Unterschiede zwischen Bildungssystemen: Während chinesische Lehrbücher den nationalen Lehrplan nahezu vollständig abdecken (95,4 %), liegt die Abdeckung für NRW bei nur 12,7 %. Dies könnte auf die unterschiedliche Granularität der Lehrpläne oder auf eine größere methodische Lücke zwischen NRW-Lehrplan und den verwendeten Mathematiklehrbüchern hinweisen. Der Coverage Score wird als vierter Indikator neben LDS, CDS und HDS in die Analyse einbezogen.

LinguaGraph — Results: CognitiveSpace Knowledge Graph

Language: German Status: Final (v0.13.3) — §4 (Humanvalidierung) auf N=15 (ersetzt N=8; Entwurfsnotiz archiviert unter _archive/20260811_rnd_review/docs/paper/03_results_human_v2.md)

---

3. Ergebnisse

3.1 CognitiveSpace: Statistische Übersicht

Die Extraktion und Fusion der 68 Lehrbücher ergibt einen Wissensgraphen mit folgenden Kenngrößen:

Metrik	Wert
Gesamtkonzepte	556 (eindeutige Konzepte)
Gesamtrelationen	525 (direkte Relationen)
Lehrbuchquellen	68 (39 ZH + 18 EN + 11 DE)
Bildungsstufen	4 (Grundschule → Universität)

Metrik	Wert
Dichte	0,0015
Isolierte Knoten	2 (< 0,5 %)

3.2 Verteilung nach Bildungsstufe

Die Konzepte verteilen sich erwartungsgemäß über die vier Bildungsstufen, wobei der Schwerpunkt auf Oberstufe und Universität liegt:

Stufe	Konzepte	Anteil	Charakteristik
Grundschule	26	4,7 %	Grundlegende Arithmetik, einfache Geometrie
Mittelschule	57	10,3 %	Algebra, Gleichungen, Funktionen
Oberstufe	200	36,0 %	Analysis, Wahrscheinlichkeit, Vektoren
Universität	273	49,1 %	Höhere Analysis, Lineare Algebra, DGLS

Diese Verteilung spiegelt die zunehmende Spezialisierung und den wachsenden Begriffsumfang in höheren Bildungsstufen wider.

3.3 Sprachübergreifende Abdeckung

Die Alignierung zeigt eine substanzielle dreisprachige Überschneidung:

Abdeckung	Konzepte	Anteil*
ZH + EN + DE (dreisprachige Gruppen)	219	39,4 %
Sprachspezifisch (nicht abgedeckt)	359	—
— davon nur ZH	120	21,6 %
— davon nur DE	94	16,9 %
— davon nur EN	145	26,1 %

*Anteile beziehen sich auf 556 Konzepte. 219 Gruppen ≠ 219 Konzepte: Gruppen- und Konzeptzählung verwenden unterschiedliche Nenner, daher addieren sich 39,4 % + 64,6 % nicht auf 100 %.

Die relativ hohe exklusive ZH-Abdeckung (21,6 %) ist auf die spezifischeren chinesischen Lehrpläne in der Grund- und Mittelschule zurückzuführen, während die Exklusivanteile für EN (26,1 %) und DE (16,9 %) deutlich geringer ausfallen.

3.4 Semesterstruktur-Analyse

Die Analyse des Graphen auf Semesterstruktur zeigt folgende Beobachtungen:

Konnektivität: Der Graph ist dünn verknüpft (238 Links auf 556 Knoten, Dichte 0,0015, 388 Zusammenhangskomponenten): 381 Knoten (68,5 %) haben keine ausgehenden Kanten; die größte Komponente umfasst 121 Knoten. Die Struktur folgt einem Kern-Peripherie-Muster mit wenigen zentralen Ankerkonzepten und vielen peripheren Einträgen.

Hierarchische Struktur: Grundschulkonzepte haben einen hohen Zentralitätsgrad und dienen als Anker für zahlreiche Oberstufen- und Universitätskonzepte. Dies bestätigt das erwartete "Knowledge Core → Expansion"-Muster.

Kreuzsprachliche Kanten: Konzepte, die in mehreren Sprachen vorkommen, weisen strukturell ähnliche Nachbarschaftsbeziehungen auf — ein Indikator dafür, dass die zugrundeliegende mathematische Wissensstruktur sprachunabhängige Invarianten besitzt.

3.5 CognitiveSpace 3D-Visualisierung

Die CognitiveSpace-Visualisierung stellt den Wissensgraphen als interaktive 3D-Kugelschale dar. Die Visualisierung ist unter cognitive-space/web/index.html lokal ausführbar und wird über GitHub Pages automatisch bereitgestellt.

Wichtigste Funktionen: - Kugelschalen-Layout: Vier konzentrische Schalen nach Bildungsstufe, deterministisch positioniert - Sprachfilter: Interaktive Umschaltung ZH/EN/DE/All mit sofortiger graphischer Aktualisierung - Farbcodierung: Grün (Grundschule) → Cyan (Mittelschule) → Blau (Oberstufe) → Violett (Universität) - Knotendetail: Anzeige von Lehrbuchquelle, Kapitel, Abschnitt bei Klick - Such- und Erkundungsmodi: WASD-Navigation, drei Ansichtsmodi, BFS-Expansion bei Klick

3.6 CognitiveSpace-Screenshot

[Abbildung: CognitiveSpace-3D-Visualisierung — 556 Konzepte in konzentrischen Kugelschalen, vier farbcodierte Bildungsstufen, sichtbare 238 Links (von 525 Relationen) als blaue Verbindungslinien]

3.7 LDS-K: Sprachübergreifender Strukturvergleich (Textbook-Pipeline)

Die Pipeline-basierte LDS-K Analyse der Mathematik-Lehrbücher (556 Konzepte, 3 Sprachen) ergibt:

Sprachpaar	LDS-K
ZH-EN	0.934
DE-EN	0.938
ZH-DE	0.519

Der ZH-DE Wert ist auffällig niedrig — chinesische und deutsche Mathematikbücher sind strukturell ähnlicher als jede der beiden mit den englischen Lehrbüchern. Um zu testen, ob diese Werte tatsächlich sprachgetriebene Divergenz messen, wurde eine Null Model Suite mit degree-preserving Randomisierung (Double-Edge Swap, 1000 Iterationen) durchgeführt:

Bedingung	ZH-EN	DE-EN	ZH-DE
Full (LDS-K baseline)	0.934	0.938	0.519
Structure Null (degree-preserving)	0.957	0.957	0.717
Node-Permuted Null	0.934	0.938	0.519
Complete Random	1.000	1.000	1.000

Zentraler Befund: Full LDS-K < Structure Null LDS-K für alle drei Sprachpaare. Unter degree-preserving Randomisierung sind die randomisierten Graphen systematisch unterschiedlicher als die echten Graphen. Dies bedeutet, dass Lehrbuch-Wissensstrukturen über Sprachgrenzen hinweg konvergieren — das Gegenteil einer sprachgetriebenen Divergenz.

Die Interpretation: LDS-K wird von der Gradverteilungsstruktur dominiert (eine Eigenschaft, die von universeller mathematischer Prerequisite-Logik geteilt wird), nicht von sprachspezifischen Inhaltsarrangements. Der wissenschaftliche Kernbeitrag verschiebt sich damit zu ΔLDS = LDS-C LDS-K, der den sprachspezifischen Anteil der menschlichen Kognition isoliert.

3.8 LDS-K Vertiefung: Bildungsebenen, Sensitivität und Cross-Source-Nullmodell

Die LDS-K-Ergebnisse wurden vertieft (reproduzierbar via scripts/lds_k_deepen.py + lds_k_wiki_gloss.py):

Bildungsebenen: Die ZH-DE-Konvergenz (0.52 im Pool) ist systematisch über alle vier Bildungsebenen — Grundschule 0.39, Mittelschule 0.80, Oberstufe 0.54, Universität 0.49. Sie ist kein Artefakt einer einzelnen Ebene. Die Kantenkomponente trägt für ZH-DE dreifach mehr zur Divergenz bei (Δ 0.075) als für ZH-EN/DE-EN (Δ ~0.02) — die Konvergenz zeigt sich auch in der Relationenstruktur, nicht nur in der Konzeptwahl.

Sensitivität: Die Ergebnisse sind robust gegenüber (a) Kantenrichtung (Δ ⩽ 0.004), (b) Alignierungstoleranz — Synonym-/Stem-Mapping vs. exakte Gloss-Übereinstimmung (Δ ⩽ 0.05), (c) Wichtigkeitsschwelle (Δ ⩽ 0.02 auf LLM-Konzepten). Die Rangfolge (ZH-DE < ZH-EN ≈ DE-EN) bleibt in allen Konfigurationen erhalten.

Cross-Source-Nullmodell: Zwei Quellen wurden verglichen: 1. Lehrbuch (Mathematik) vs. Wikipedia (sozial): LDS = 1.00 (null Knotenüberlappung) — dies ist eine Domänenkonfundierung (Mathematik ⊥ Soziales), kein Sprachsignal. Der Vergleich ist methodisch ungeeignet, weil die Quellen verschiedene Wissensdomänen abdecken. 2. Wikipedia(zh) vs. Human(zh), beide sozial (domänenrein): LDS = 0.94 — die Quelle trägt einen großen Anteil zur Divergenz bei (institutionelle Enzyklopädie vs. spontane Kognition überschneiden sich kaum, J_node=0.06).

Wikipedia-Negativkontrolle (korrigiert): Frühere Analysen berichteten LDS = 1.00 für soziale Wikipedia-Konzepte. Dies war ein Alignierungs-Artefakt: chinesische Konzepte (z. B. 自由) werden von canonical_key (das nur lateinische Zeichen extrahiert) zu leeren Schlüsseln reduziert. Nach Glossierung von 96 chinesischen und deutschen Konzepten in denselben kanonischen Schlüsselraum ergeben sich reale Werte:

Quelle	ZH-EN	DE-EN	ZH-DE
Mathematik-Lehrbücher	0.934	0.938	0.519
Wikipedia (sozial, aligniert)	0.698	0.723	0.819

Zentraler Befund (mit Einschränkung): Die Struktur sozialer Konzepte (Freiheit, Gerechtigkeit, Verantwortung, Heimat, Erfolg) weist eine höhere sprachübergreifende Divergenz auf als die Mathematik-Labels — und das Muster ist umgekehrt: Während ZH-DE in Mathematik am stärksten konvergiert (0.52), ist es in der sozialen Domäne am divergentesten (0.82). Einschränkung (P2-Recheck): Dieser Vergleich ist kein belastbarer Beleg unabhängiger sprachübergreifender Konvergenz. Die Mathematik-Knoten sind Alignierungs-Labels (keine unabhängig extrahierten Konzepte), das de-Label-Feld enthält teils chinesische Texte, und ein Size-Matching auf die Wikipedia-Größe kehrt das Muster um (bei k=15–35 ist der Wikipedia-J_node höher; Details: docs/p2_methodology_rechecks.md). Die Mathematik-"Konvergenz" (0.519) ist daher teilweise ein Darstellungs-/Größen-Artefakt; die Aussage ist nur als "die Alignierungs-Labels institutionellen Wissens stimmen sprachübergreifend besser überein als die unabhängig extrahierten sozialen Konzepte" zu lesen — nicht als Beleg, dass institutionelles Wissen "inhaltlich" konvergiert.

4. Humanvalidierung: Kognitive Graphen mehrsprachiger Probanden

Dieser Abschnitt präsentiert LDS-C Ergebnisse auf Basis der erweiterten, QC-geprüften Stichprobe (N=15: 6 DE • 6 ZH • 3 EN), extrahiert mit deepseek-v4-flash (opencode GO). Alle Ergebnisse sind reproduzierbar (Skripte scripts/lds_c_extract.py, lds_c_compute.py, lds_c_thematic.py).

4.1 Versuchsdesign und Datengrundlage

Dimension	Spezifikation
Teilnehmer	N=15 (6 DE, 6 ZH, 3 EN bilingual-ZH)
Stimuli	5 soziale Themen (Freiheit, Gerechtigkeit, Verantwortung, Heimat, Erfolg)
Datenquelle	Freeze-SSOT freeze_survey_20260703 + freeze_survey_20260712 (QC-geprüft, 4/4 in zweitem Batch)
Extraktionsmodell	deepseek-v4-flash (opencode GO), Temperatur 0.3
Extrahierte Antworten	15/15 (100 %), pro Antwort 5–6 Konzepte je Thema mit engl. Gloss
Konzept-Alignierung	Englischer Gloss → kanonischer Schlüssel (Synonym-Karte + Stemmer)

4.2 Konzeptebene LDS-C (zwischen den Sprachgruppen)

Aggregierte Konzeptmengen pro Sprachgruppe, LDS = 1 Jaccard(kanonische Schlüssel).

Sprachpaar	LDS-C (gepoolt)	95 %-CI (Bootstrap)	LDS-K (Konzept)	ΔLDS
ZH-EN	0.963	[0.936, 0.985]	0.977	0.014
DE-EN	0.932	[0.899, 0.986]	0.977	0.045
ZH-DE	0.936	[0.908, 0.960]	0.887	+0.049

Wichtig: Alle drei Sprachpaare zeigen LDS-C ≈ 0.93–0.96. Die Werte liegen nahe der Zufallsverteilung (siehe 4.3) und deutlich höher als die früheren N=8-Schätzungen (0.70–0.75), die mit einer älteren Pipeline berechnet wurden.

● Hinweis zur ΔLDS-Definition (Berichtsformat): Die hier berichteten ΔLDS-Werte (konzeptuelle Ebene, 0.044 bis +0.047) verwenden die konzeptuelle LDS-K (nur Knoten) als Basis, sodass LDS-C und LDS-K auf derselben Skala verglichen werden (apples-to-apples). Die relationale v3-LDS-K (§4.5) ist mit den relationalen LDS-C-Werten zu vergleichen — beide ΔLDS-Formate sind in der Arbeit durchgängig nach Ebene getrennt und nicht miteinander vermischt. Die Titelaussage „ΔLDS ≈ 0 (0.05 bis +0.05)" bezieht sich auf die konzeptuelle Ebene.

4.3 Null-Modell-Prüfung: Kein separierbares Sprachsignal

Zwei Null-Modelle testen, ob die beobachtete LDS-C auf Sprache zurückgeht oder auf Teilnehmervariabilität:

Sprachpaar	LDS-C (beobachtet)	Within-Lang Split-Half (Boden)	Label-Permutation	Perm-p
ZH-EN	0.963	0.958	0.940	0.08
DE-EN	0.932	0.923	0.938	1.00
ZH-DE	0.936	0.922	0.935	1.00

Zentraler Befund: 1. Within-Language-Split-Half (Teilnehmervariabilität innerhalb einer Sprache): 0.92–0.96 ≈ beobachtete LDS-C. Die Variabilität innerhalb einer Sprachgruppe ist genauso groß wie die Divergenz zwischen den Gruppen. 2. Label-Permutation (Permutation der Sprachlabels): 0.94–0.94 ≈ beobachtete LDS-C. Der formale Permutationstest (500 Iterationen, zweiseitig) liefert p = 0.08 / 1.00 / 1.00 — die Sprachlabels tragen kein messbares Signal (kein p-Wert unterhalb konventioneller Schwellen).

Interpretation: Auf Konzeptebene ist die LDS-C von Teilnehmervariabilität dominiert. Das Between-Subject-Design (jede Person antwortet nur in einer Sprache) vermischt sprachgetriebene Divergenz mit individueller Variabilität. Bei N=15 (6/6/3) dominiert die letztere. Die Hypothese ΔLDS > 0 wird auf Konzeptebene nicht bestätigt.

4.4 Thematische Analyse (6-Kategorien-Codebook)

Eine grobkörnige Codierung (Codebook v1: Legal/Institutional, Individual/Autonomy, Material/Concrete, Social/Relational, Moral/Abstract, Emotional/Affective; LLM-Klassifikation, n=335 Konzepte) vergleicht die Verteilungen der Themenkategorien:

Kategorie	DE	EN	ZH
Legal/Institutional	0.05	0.10	0.12
Individual/Autonomy	0.27	0.21	0.25
Material/Concrete	0.11	0.12	0.09
Social/Relational	0.15	0.21	0.15
Moral/Abstract	0.17	0.15	0.24
Emotional/Affective	0.14	0.10	0.11

Test	Ergebnis
Chi-Quadrat	x^2=9.14, p=0.519 (n.s.), Cramér's V=0.117
Permutationstest	p=0.354 (5000 Iterationen)
Shannon-Entropie	DE 2.68 • EN 2.74 • ZH 2.64

◆ Befund: Die Kategorienverteilungen unterscheiden sich nicht signifikant zwischen den Sprachen. Es zeigen sich jedoch konsistente Richtungstendenzen, die mit einer kulturellen Rahmungshypothese übereinstimmen: ZH rahmt Freiheit häufiger rechtlich-institutionell (Freiheit: 0.24 vs. DE 0.09 / EN 0.07), ZH stärker moralisch-abstrakt (0.24), DE stärker individuell-autonom (0.27) und affektiv (0.14). Diese Tendenzen sind bei N=15 nicht statistisch bestätigbar.

4.5 Relationale Ebene (v3-Knoten+Kante)

Zusätzlich zur Konzeptebene wurden explizite Relationen zwischen den extrahierten Konzepten erfasst (7 Relationstypen, pro Thema). Insgesamt 162 Kanten (DE 59, ZH 67, EN 36). Die v3-LDS kombiniert Knoten- und Kanten-Jaccard (frozen Formel):

Sprachpaar	LDS v3	95 %-CI	Node-Jaccard	Edge-Jaccard	LDS-K (v3)	ΔLDS	Split-Half-Boden
ZH-EN	0.977	[0.962, 0.992]	0.037	0.010	0.934	+0.043	0.979
DE-EN	0.966	[0.949, 0.993]	0.068	0.000	0.938	+0.028	0.959
ZH-DE	0.964	[0.948, 0.980]	0.064	0.008	0.519	+0.445	0.958

Zentraler Befund: Die Kantenstrukturen überlappen sprachübergreifend kaum (Edge-Jaccard ≈ 0, DE-EN exakt 0.000). Der Split-Half-Boden (0.958 – 0.979) entspricht der beobachteten v3-LDS (0.964 – 0.977) — auch auf relationaler Ebene ist die Divergenz vollständig durch Teilnehmervariabilität erklärbar. Der scheinbar große ΔLDS-Wert für ZH-DE (+0.445) ist ein Artefakt des Vergleichs zwischen spärlichen Menschengraphen und dichten, strukturell konvergenten Lehrbuchgraphen — nicht ein Beleg für sprachgetriebene Divergenz (der Split-Half-Boden von 0.958 für ZH-DE widerlegt dies).

Fazit relationale Ebene: Die relationale Struktur bestätigt die Konzept- und Themenebene. Unter Between-Subject-Bedingungen (N=15) ist kein separierbares Sprachsignal nachweisbar — weder auf Knoten-, Kanten-, noch Kategorienebene.

4.6 Zusammenfassung und methodologische Reflexion

Die erweiterte Humanvalidierung (N=15) liefert über drei Analyseebenen hinweg ein konsistentes, ehrliches negatives Ergebnis:

Konzeptebene: LDS-C 0.93 – 0.96, nicht von Teilnehmervariabilität unterscheidbar (Split-Half-Boden 0.92 – 0.96; Label-Permutation 0.94).

Themenebene: Kategorienverteilungen nicht signifikant verschieden (x^2 p=0.52, Permutation p=0.35), aber konsistente Richtungstendenzen (ZH rechtlich/moralisch, DE autonom/affektiv, EN sozial).

Relationale Ebene (v3): LDS-C 0.96 – 0.98, Edge-Jaccard ≈ 0, Split-Half-Boden ≈ beobachtet → ebenfalls kein separierbares Sprachsignal.

Die früheren N=8-Befunde (LDS-C 0.70 – 0.75, ΔLDS>0) werden nicht repliziert: Sie stammen aus einer älteren Pipeline und sind mit dem erweiterten, QC-geprüften Datensatz und der standardisierten Extraktion nicht vereinbar.

Methodologische Lehre (Kernbeitrag dieser Revision): Ein Between-Subject-Design kann sprachgetriebene Divergenz nicht von individueller Variabilität trennen, wenn die Stichprobe klein ist. Dies gilt unabhängig von der Analyseebene (Konzept, Kategorie, Relation). Zukünftige Studien benötigen (a) ein Within-Subject-Design, (b) größere Stichproben pro Sprachgruppe, oder (c) Metriken, die gegen individuelle Konzeptwahl robust sind.

Dieser negative Befund ist wissenschaftlich wertvoll: Er falsifiziert die einfache Hypothese "Sprache → unterschiedliche Konzeptgraphen" auf allen drei Ebenen und präzisiert die Bedingungen, unter denen sprachliche Kognitionseffekte nachweisbar wären.

5. LLM-as-Subject: Sprachsignal im Within-Subject-Design

Der negative Humanbefund (§4) wirft die entscheidende methodologische Frage auf: Ist die fehlende Trennbarkeit ein Artefakt des Between-Subject-Designs (Teilnehmervariabilität überlagert die Sprache) oder ein echtes Nichtvorhandensein sprachlicher Kognitionseffekte? Um diese zu beantworten, wurde ein kontrolliertes Within-Subject-Experiment mit einem Large Language Model als standardisiertem kognitivem Subjekt durchgeführt: dasselbe Modell (deepseek-v4-flash, opencode GO) antwortet in ZH/DE/EN auf dieselben 5 Themen. Da die Gewichte

identisch sind, ist die Sprache die einzige Variation — das Within-Subject-Design gilt konstruktionsbedingt. Reproduzierbar via scripts/lds_c_llm_subject.py, lds_c_llm_analyze.py, lds_c_llm_per_topic.py, lds_c_llm_lmm.py.

5.1 Versuchsdesign

Dimension	Spezifikation
Subjekt	deepseek-v4-flash (opencode GO), Temperatur 0.3, k=10 unabhängige Sessions pro Bedingung
Stimuli	5 soziale Themen (Freiheit, Gerechtigkeit, Verantwortung, Heimat, Erfolg)
Sonden	P1 Sprach-Haupteffekt • P2 Rahmen-Code-Entkopplung • P3 Freie Assoziation • P5 Antwort- vs. Promptsprache
Hygiene	jede Session unabhängig (keine Kontextakkumulation), themenweise Beantwortung, konterbalancierte Reihenfolge
Umfang	220 Einheiten / ~570 API-Aufrufe, 0 leere Extraktionen

5.2 Sprach-Haupteffekt (P1): Signal ist nachweisbar

Sprachpaar	LDS-C (gepoolt)	95 %-CI	Split-Half-Boden	Label-Permutation	Perm-p
ZH-EN	0.955	[0.935, 0.966]	0.875	0.879	<0.01
DE-EN	0.930	[0.912, 0.955]	0.846	0.880	<0.01
ZH-DE	0.945	[0.932, 0.961]	0.862	0.879	<0.01

Zentraler Befund: Die beobachtete LDS-C übersteigt den Split-Half-Boden um +0.08 bis +0.09 und die Label-Permutation um +0.05 bis +0.08. Ein formaler Permutationstest (500 Iterationen; zweiseitig; Anteil der Permutationen mit LDS ≥ beobachtetem LDS) liefert für alle drei Sprachpaare $p < 0.01$ (keine der 500 Permutationen erreichte die beobachtete LDS-C). Anders als bei den menschlichen Daten (N=15, Between-Subject: LDS-C ≈ Boden ≈ Permutation, p = 0.08/1.0/1.0) trägt das Sprachlabel hier messbares Signal.

Dies ist der zentrale Kontrast der Arbeit: Derselbe Messrahmen (LDS-C + Nullmodelle), nur das Design von Between-Subject (Menschen) auf Within-Subject (LLM) umgestellt, verwandelt ein nicht-trennbares Signal in ein trennbares. Der menschliche Negativbefund ist damit als Design-Artefakt charakterisiert — nicht als Beleg für das Fehlen sprachlicher Kognitionseffekte (Charakterisierung auf Exklusion + Konsistenz-Demonstration gestützt; keine quantitative Kausalzuordnung, s. §5.9.2b/§8.14).

5.3 Rahmen-Code-Entkopplung (P2): Rahmen wirkt innerhalb, nicht über Codes

Code	Bedingung		LDS	95 %-CI
ZH	ZH-natürlich	vs. ZH-Code+DE-Rahmen	0.916	[0.905, 0.943]
DE	DE-natürlich	vs. DE-Code+ZH-Rahmen	0.914	[0.901, 0.937]

Die reine Rahmenmanipulation (gleicher Code, anderer kultureller Rahmen) erzeugt LDS ≈ 0.91 – 0.92 — oberhalb des Split-Half-Bodens. Die Rahmeninstruktion verändert also die Konzeptwahl innerhalb eines Codes. Der scheinbare Vergleich mit der sprachübergreifenden LDS-C (0.93 – 0.96) ist jedoch nicht äquivalent: Er vergleicht innerhalb desselben Code-Orbits, nicht zwischen Orbits (siehe LMM, §5.6).

5.4 Freie Assoziation (P3): statistische Basis der Konzeptdivergenz

Sprachpaar	Assoziations-LDS	Konzept-LDS (P1)	Δ
ZH-EN	0.933	0.955	0.022
DE-EN	0.920	0.930	0.011
ZH-DE	0.736	0.945	0.209

Die freien Assoziationen desselben Modells (englisch glossiert in denselben kanonischen Schlüsselraum) erklären die Konzeptdivergenz für EN-Paare nahezu vollständig ($\Delta \approx 0.01$ bis 0.02) — die Konzeptdivergenz ist hier ein statistisches Nebenprodukt der Assoziationsverteilung. Für ZH-DE dagegen liegt die Assoziationsdivergenz (0.736) weit unter der Konzeptdivergenz (0.945): Es existiert eine strukturelle Ebene jenseits der Assoziationsstatistik — konsistent mit der Rahmen-Wirkung auf das kulturell entfernteste Paar.

5.5 Antwort- vs. Promptsprache (P5): Promptsprache ohne eigenen Effekt

Kontrast		LDS	95 %-CI	Innerhalb-Bedingung-Boden
DE-Frage/DE-Antwort	vs. ZH-Frage/DE-Antwort	0.815	[0.800, 0.884]	0.827 / 0.843

Bei fixierter Antwortsprache (DE) erzeugt die Variation der Fragesprache (DE vs. ZH) keine über dem Boden liegende Divergenz. Die Frage-/Promptsprache trägt kein eigenständiges Signal — die Antwortsprache (Produktionssprache) dominiert.

5.6 LMM: Sprachcode ist der dominante Organisator (Mechanismus-Kern)

Um die marginalen Beiträge von Sprachcode und kulturellem Rahmen zu trennen, wurde ein gemischtes Modell über 50 dyadische Zellpaare (5 Zellen × 5 Themen, Response = Jaccard-Similarität) geschätzt:

Similarität ~ same_lang + same_frame + (1|topic)

Effekt	Koeffizient (Jaccard)	SE	t	p	Permutation p	Bootstrap p
Intercept	+0.050	0.011	4.60	<0.001	—	—
same_lang	+0.038	0.009	4.28	<0.001	0.000	<0.01
same_frame	+0.001	0.009	0.13	0.90	0.779	0.82

Marginalbeitrag: Das Teilen des Sprachcodes erhöht die Konzeptähnlichkeit signifikant (+0.038); das Teilen des kulturellen Rahmens hat keinen signifikanten marginalen Beitrag (+0.001). Die Permutationsprüfung (blockweise, 1000 Iterationen) bestätigt die Robustheit. Zusätzlich wurde ein Cell-Cluster-Bootstrap (1000 Iterationen, Resampling der 5 Zellen unter Beibehaltung der Dyaden innerhalb der Zelle) durchgeführt, um der Nicht-Unabhängigkeit der Dyaden (jede Zelle erscheint in 4 Dyaden) Rechnung zu tragen: same_lang bleibt signifikant (p<0.01), same_frame nicht-signifikant (p=0.82) — die Kernsaussage ist gegenüber dieser konservativeren Inferenz robust.

Harmonisierung mit P2 (Zwei-Boden-Konzept): - Boden 1 = Split-Half innerhalb einer Bedingung: J≈0.12 (gleicher Code + Rahmen) - Boden 2 = vollständig verschiedene Zellen: J≈0.05 (verschiedener Code + Rahmen) - Rahmenwechsel (gleicher Code): J≈0.088 → zwischen den Böden: bewegt Konzepte innerhalb des Code-Orbits (0.12→0.088), erreicht aber keine codeübergreifende Divergenz (0.088 0.05).

→ Der Sprachcode ist die dominante Organisationsebene (lexikalisch-statistisch); der kulturelle Rahmen ist ein sekundärer, code-interner Modulator.

5.7 Themenbezogene Decomposition

Thema	Sprachsignal (P1Boden)	Rahmen-Effekt (P2)	LDS-C ZH-DE (D1-Baseline)
Freiheit	0.045	0.89	0.959
Gerechtigkeit	0.114	0.96	0.957
Verantwortung	0.058	0.87	0.980
Heimat	0.082	0.86	0.927
Erfolg	0.115	0.98	0.927

Die D1-Baseline-Spalte (data/lds_c/lds_c_results_20260808.json, per_topic_lds_c.ZH-DE) verankert die Themenabsolutwerte: Alle fünf Themen liegen im Band 0.93 – 0.96 (Verantwortung 0.9804 höchst), konsistent mit LDS-C ≈ 0.93 – 0.96.

Das Sprachsignal und der Rahmen-Effekt sind bei abstrakten, moralisch konnotierten Themen (Erfolg, Gerechtigkeit) am stärksten — konsistent mit der Hypothese, dass abstrakte Konzepte sprach- und kulturabhängiger sind als konkrete.

5.8 Zusammenfassung und Einordnung

1. Sprachsignal existiert im Within-Subject-Design (P1: LDS-C Boden) — der menschliche Negativebefund ist ein Between-Subject-Artefakt.
2. Dominanter Mechanismus: Sprachcode (M2, lexikalisch-assoziativ) — signifikanter marginaler Beitrag im LMM.
3. Sekundärer Mechanismus: kultureller Rahmen (M3) — wirkt innerhalb des Codes, überbrückt Codes nicht.
4. ZH-DE trägt eine strukturelle Ebene jenseits der Assoziationsstatistik (P3: Δ=0.209) — am stärksten beim kulturell entferntesten Paar.
5. Promptsprache (M5) ist ohne eigenen Effekt (P5 falsifiziert).

Die Ergebnisse validieren die zentrale Methodenlehre aus §4.6: ein Within-Subject-Design ist erforderlich, um sprachgetriebene Divergenz zu trennen — und bieten zugleich eine testbare Blaupause für zukünftige Humanstudien mit Within-Subject-Design.

5.9 Design-Effekt-Beweis, Divergenztreiber und Knoten/Kanten-Dekomposition

Reproduzierbar via scripts/lds_c_design_effect.py, lds_c_divergence_drivers.py, lds_c_node_edge_decomp.py; Ergebnisse in data/lds_c/llm_subject/design_effect_*.json, data/lds_c/divergence_drivers_*.json, data/lds_c/lds_k_deep/node_edge_decomp_*.json.

5.9.1 Design-Effekt-Beweis: gleiche Signalamplitude, unterschiedliche Bodenlinie.

Der Vergleich von Mensch (Between-Subject, N=15) und LLM (Within-Subject, k=10) auf identischem Messrahmen zeigt:

Sprachpaar	LDS-C Mensch	Boden Mensch	s/f Mensch	LDS-C LLM	Boden LLM	s/f LLM
ZH-EN	0.963	0.958	1.005	0.955	0.875	1.092
DE-EN	0.932	0.924	1.009	0.930	0.846	1.100

Sprachpaar	LDS-C Mensch	Boden Mensch	s/f Mensch	LDS-C LLM	Boden LLM	s/f LLM
ZH-DE	0.936	0.922	1.015	0.945	0.862	1.096

Die Signalamplitude (LDS-C) ist bei Mensch und LLM nahezu identisch (0.93 – 0.96). Der entscheidende Unterschied liegt in der Bodenlinie (Within-Language Split-Half): Bei Menschen überlagert die Variabilität innerhalb einer Sprachgruppe (0.92 – 0.96) die Divergenz zwischen den Gruppen (Signal/Boden ≈ 1.00 → Signal untergegangen); beim LLM liegt der Boden (0.85 – 0.87) deutlich unter dem Signal (Signal/Boden ≈ 1.09 – 1.10 → Signal sichtbar).

● Hinweis zur Stichprobengröße (Sample-Size-Kontrolle): Die menschlichen Bodenwerte in der Tabelle beruhen auf Halb-Splits mit 3/3 (DE/ZH) bzw. 1/2 (EN) Teilnehmenden; die LLM-Bodenwerte auf 5/5 Stichproben. Ein N-abgestimmter Floor-Scan (§5.9.1, N=6 → 3+3 Halb-Splits, analog zur menschlichen Aufteilung) ergibt für das LLM weiterhin einen Boden von 0.854 – 0.885 — weit unter dem menschlichen Boden (0.922 – 0.958, Differenz 0.05 – 0.07). Die Schlussfolgerung ist daher nicht durch unterschiedliche Stichprobengrößen verursacht.

Ein Floor-Scan (LLM-Signal bei N = 3/5/6/8/10 Stichproben pro Sprache) zeigt, dass das Signal auch bei N=3 nachweisbar bleibt (Marge +0.02 – 0.04) und bei N=10 auf +0.08 anwächst. Die menschliche Null ist also kein Stichprobeneffekt, sondern Folge der Varianzstruktur: menschliche Teilnehmer innerhalb einer Sprache sind heterogen, LLM-Stichproben desselben Gewichtssatzes sind homogen.

5.9.2 Virtuelle Between-Subject-Linse + Heterogenitäts-Injektion (Kausal-Konsistenztest).

(a) Exklusion. Werden die LLM-Stichproben als "virtuelle Teilnehmer" mit exakt der menschlichen Between-Subject-Pipeline bei menschlicher N (=6 pro Sprache) analysiert, überlebt das Sprachsignal (Marge ZH-EN +0.064, DE-EN +0.073, ZH-DE +0.070). D. h. weder "keine Sprachwirkung" noch "Between-Subject-Design an sich" erklärt die menschliche Null.

(b) Direkte Kausalprüfung (Heterogenitäts-Injektion, scripts/lds_c_heterogeneity_injection.py). Mensch und LLM zeigen fast identische within-language Paar-Overlaps (ZH 0.057 vs 0.059; DE 0.106 vs 0.108) — die Heterogenität einzelner Antworten ist vergleichbar. Der Unterschied liegt in der Aggregationssparsität: menschliche Teilnehmer tragen pro Person weniger Konzepte bei, sodass ein 3+3-Halb-Split spärlich ist und beide Hälften kaum überlappen. Injiziert man diese Sparsität in die LLM-Stichproben (Konzept-Dropout mit Wahrscheinlichkeit 1q), kollabiert die Signal-Marge monoton:

q (Behaltewahrscheinlichkeit)	LLM LDS-C (ZH-DE)	LLM Boden	Signal-Marge
1.00 (keine Injektion)	0.944	0.873	+0.071
0.60	0.939	0.905	+0.033
0.40	0.946	0.916	+0.030
0.30	0.955	0.941	+0.014
Mensch (Referenz)	0.936	0.921	+0.015

Bei q=0.30 erreicht der injizierte LLM-Boden (0.941) den menschlichen Boden (0.921), und die Signal-Marge kollabiert auf +0.014 — nahe der menschlichen +0.015. Dies zeigt, dass eine ausreichende Aggregationssparsität die menschliche Null reproduzieren kann. Einschränkung (P2-Recheck): q=0.30 ist der spärlichste der zehn Scan-Punkte und 3× spärlicher als die tatsächliche menschliche Sparsität (der mechanismus-konsistente Zielwert q≈0.76 – 0.91 liefert weiterhin eine Marge von +0.05 – 0.065, nicht kollabiert; die automatische Boden-Kalibrierung des Skripts stoppt bei q=0.35 mit Marge +0.033, dem 2,2-Fachen des menschlichen Werts). Die Injektion ist daher eine Konsistenz-Demonstration ("extreme Sparsität kann die Null erzeugen"), keine vollständige direkte Kausalprüfung mit der realistischen menschlichen Sparsität. Die heterogenitätsbedingte Erklärung der menschlichen Null bleibt gestützt (Exklusion anderer Ursachen + Konsistenz), ist aber nicht als quantitative Kausalzuordnung zu werten.

5.9.3 Divergenztreiber (RQ4): Rahmengeladene Konzepte treiben ZH-DE.

Eine Asymmetrie-Analyse über Konzepte (nur in einer Sprache auftretende Konzepte, nach Erwähnungshäufigkeit gewichtet) identifiziert die Treiber der Sprachdivergenz. Für ZH-DE (LLM-P1):

Thema	Konzept	Richtung	Hufigkeit
Gerechtigkeit	equal opportunity	DE-only	7
Freiheit	freedom limit of	DE-only	6
Heimat	physical space	ZH-only	6
Erfolg	goal own	DE-only	6
Freiheit	boundary freedom of	ZH-only	4
Gerechtigkeit	due treatment	ZH-only	4

Die Treiber sind rahmengeladene Kulturkonzepte: DE tendiert zu Autonomie/Regel/Ziel (equal opportunity, freedom limit, goal own), ZH zu Raum/Grenze/Anspruch (physical space, boundary freedom, due treatment) — konsistent mit den Richtungstendenzen der Themenanalyse (§4.4: DE autonom, ZH rechtlich-institutionell). Geteilte Konzepte sind extrem selten (20 – 27 über 5 Themen), im Einklang mit LDS-C ≈ 0.93 – 0.96. Auf Relationsebene (menschlich): DE-only responsibility → consequence (freq 3), ZH-only success → goal (freq 2) — die strukturelle Arbeitsteilung ist auch relational sichtbar.

5.9.4 Knoten/Kanten-Dekomposition: Die Reversion ist knotengetrieben.

Die soziale/institutionelle Umkehr (ZH-DE am konvergentesten in Mathematik, am divergentesten in sozialen Wikipedia) wird in Knoten- und Kantenkomponente zerlegt (frozen v3):

Quelle	Paar	LDS v3	J_node	J_edge	node-only	Kantenbeitrag
Mathematik	ZH-DE	0.519	0.556	0.407	0.444	0.074
Wikipedia sozial	ZH-DE	0.819	0.200	0.162	0.800	0.019
Mensch kognitiv	ZH-DE	0.954	0.085	0.008	0.915	0.038

Die Umkehr besteht auf der Knotenebene (Mathematik node-only 0.444 → konvergent; sozial 0.800 → divergent), nicht auf der Kantenebene: der Kantenbeitrag ist in der Mathematik am größten (0.074). Institutionelles Wissen konvergiert primär in der Konzeptwahl (Knoten); die Beziehungsorganisation (Kanten) bleibt über alle Quellen hinweg systemisch divergent. Dies vertieft §3.8: der Kantenbeitrag 0.075 ist keine "kantengetriebene Konvergenz", sondern eine "kantengetriebene Divergenz" innerhalb eines konzeptuell konvergenten Fachs.

Einschränkung (P2-Recheck): Die Mathematik-Knoten sind Alignierungs-Labels und damit teils Alignierungsprodukte, nicht unabhängig extrahierte Konzepte; zudem enthält das de-Label-Feld im Alignierungs-Datensatz teils chinesische Texte (110/150 Stichprobe), was den mathematischen J_node künstlich erhöht. Ein Size-Matching auf die Wikipedia-Größe (~45 Knoten/Sprache) kehrt das Muster um: bei k=15–35 zeigt Wikipedia einen höheren J_node (0.07–0.15) als Mathematik (0.03–0.07); die mathematische "Konvergenz" (J_node 0.556) ist ein Artefakt der nahe-vollständigen Universumsabtastung (~195/219 Gruppen) + des Label-Artefakts, nicht ein Beleg unabhängiger sprachübergreifender Konvergenz. Die "institutionelle Konvergenz / kulturelle Divergenz"-Umkehr ist daher kein belastbarer inhaltlicher Befund (Details: docs/p2_methodology_rechecks.md).

5.10 Modellübergreifende Replikation (55 Messungen, 50 eindeutige Modelle).

Reproduzierbar via scripts/lds_c_multi_model.py; Ergebnisse in data/lds_c/llm_subject/multi_model_replication_*.json.

Das identische P1-Protokoll (3 Sprachen × k=10, gleicher Messrahmen LDS-C/Floor/Nullmodelle) wurde auf 42 DashScope-, 7 zen/OpenRouter-Modellen plus D1-Baseline sowie neu je 2 Kilo-, 1 Cohere-, 1 NIM- und 1 opencode-go-Modell angewendet → 55 vollständige Messungen (je 10/10/10), davon 50 eindeutige Modell-Identitäten (5 Modelle auf zwei Hosts gemessen: deepseek-v4-flash, deepseek-v4-pro, glm-5.2, kimi-k2.6, laguna-s-2.1). Die Familien: Qwen (20), DeepSeek (12), GLM (7), Kimi/Moonshot (6), MiniMax (2), mimo/ByteDance, laguna/poolside (×2 Hosts), longcat, nemotron-3-ultra + nemotron-3-super (NVIDIA, US-Ursprung), gpt-oss-20b (OpenAI-Gewichte, US-Ursprung), command-a (Cohere, CA-Ursprung), gpt-5.6-luna (Herkunft ungeklärt, offengelegt). ~87 % der Messungen stammen von chinesischen Anbietern; sieben westliche Messungen (6 Identitäten: NVIDIA ×2, Poolside ×2 Hosts, OpenAI-Gewichte, Cohere, luna) sind enthalten. Die Modellauswahl folgte der API-Verfügbarkeit (kostenlose Tiers, Anbieter-Quotas), nicht einer präregistrierten Matrix — eine Verfügbarkeitsstichprobe (s. Robustheit (d) in §8.15).

◆ Signal — breit repliziert, aber nicht ausnahmslos: Alle 55 ZH-DE-Sprachpaare sind signifikant ($p<0.05$, meist <0.01); die ZH-DE-Marge reicht von +0.03 (deepseek-r1-0528) bis +0.42 (command-a-03-2025). Von den 165 Sprachpaar-Tests sind 8 nicht signifikant ($p\geq0.05$) — alle betreffen englisch-haltige Paare (ZH-EN oder DE-EN), überwiegend DeepSeek-R1/Distill-Modelle (deepseek-r1-0528 DE-EN $p=0.672$). Hinweis zur p-Reportung: Bei n_iter=500 tritt $p=0.0$ auf, wenn der beobachtete Wert über allen Permutationen liegt; streng ist $p<0.004$ zu berichten, und es wurde keine Mehrfachtest-Korrektur vorgenommen. Ein formaler Bonferroni-Test ($\alpha\approx0{,}0003$ bei 165 Tests) übersteigt diese Permutationsauflösung; die Meta-Betrachtung — alle 55 ZH-DE-Paare überschreiten sämtliche 500 Permutationen (erwartet unter der globalen Null $\approx0{,}11$) — stützt die Aussage dennoch (Details §8.15).

Kulturrichtung — über dem Zufallsniveau: Eine Voten-Analyse der ZH-DE-Treiber (wie viele Modelle markieren ein Konzept als nur-in-DE bzw. nur-in-ZH) wird gegen ein frequenz-angepasstes Zufalls-Nullmodell getestet. Die beobachtete Zahl richtungskonsistenter Konzepte ($\max(DE,ZH) \geq t$) übersteigt das Nullmodell bei allen Schwellen ($p<0.001$): ≥3: 1133 vs. 882±11; ≥10: 218 vs. 147±4; ≥20: 61 vs. 12±2. Die stärksten Konzepte (Heimat:safety 46 DE-Stimmen; Heimat:physical space 44 ZH-Stimmen; equal opportunity 42) stützen die DE-Autonomie/ZH-Raum-Orientierung, jedoch mit begrenzter Abdeckung (z. B. produzierten nur 46 von 56 Modellen Heimat:safety als einseitigen Treiber).

6. Fächerübergreifende Validierung: Physik und Chemie

Um zu prüfen, ob sich die Befunde über die Mathematik hinaus verallgemeinern lassen, wurden zwei parallele Wissensgraphen konstruiert: ein Physik-Graph (von der Grundschule bis zur Hochschulebene; Mechanik, Thermodynamik, Elektromagnetismus, Optik, moderne Physik, Strömungsmechanik sowie Kern- und Teilchenphysik; 366 Konzepte, 383 Relationen, 94 Verlagsausgaben) und ein Chemie-Graph (Mittel- und Oberstufe sowie Hochschule; Materiestruktur, chemische Reaktionen, Stöchiometrie, organische/anorganische/physikalische/analytische Chemie; 220 Konzepte, 215 Relationen, 6 Verlage pro Sprache).

6.1 Konzeptdichtestruktur (CDS)

Niveau	Physik CDS	Mathe CDS	Verhltnis
Grundschule	0,222	0,216	1,03
Mittelstufe	0,029	0,271	0,11
Oberstufe	0,013	0,073	0,17
Hochschule	0,011	0,042	0,26

◆ Befund F6: Der Physik-CDS erreicht seinen Höhepunkt in der Grundschule (0,222), während der Mathe-CDS in der Mittelstufe (0,271) kulminiert. Beide Disziplinen zeigen Dichtespitzen in frühen Bildungsstufen, jedoch auf unterschiedlichen Niveaus:

- Physik: Mechanik-Konzepte der Grundstufe (Kraft, Masse, Geschwindigkeit, Dichte, Druck) sind hochgradig miteinander verbunden — nahezu jedes Konzept bezieht sich auf Kraft oder Bewegung. Dies spiegelt den fundamentalen Charakter der klassischen Mechanik wider.

- Mathematik: Die Mittelstufe markiert die Integration von Arithmetik, Algebra, Geometrie und Wahrscheinlichkeitsrechnung zu einem kohärenten Netzwerk, bevor die Aufspaltung in spezialisierte Teilgebiete erfolgt.

Der 3,7-fache Abfall von Physik-Grundschule (0,222) zur Oberstufe (0,013) spiegelt die rasche curriculare Expansion wider: Der Physikunterricht erweitert sich von ca. 10 Kernkonzepten der Mechanik auf 136 Oberstufenkonzepte, die Thermodynamik, Elektromagnetismus, Optik und moderne Physik umfassen.

6.2 Hierarchietiefenstruktur (HDS)

Metrik	Physik	Mathe
Maximale Tiefe	6	8
Mittlere Tiefe	0,85	0,40
Wurzelkonzepte	219 (60 %)	459 (83 %)

◆ Befund F7: Physik weist tiefere Voraussetzungsketten auf (mittlerer HDS 0,85 vs. Mathe 0,40). Dies spiegelt den kumulativen Charakter physikalischen Wissens wider: Das Verständnis elektromagnetischer Induktion erfordert zuerst die Beherrschung von elektrischer Ladung → Strom → Magnetfeld → Faradaysches Gesetz — eine Kette von 4+ Konzepten. Die Mathematik hingegen verfügt über mehr unabhängige Einstiegspunkte (83 % Wurzeln).

6.3 Verlagsabdeckung

Sprache	Verlage	Bildungsstufen
Chinesisch (ZH)	33	Grundschule → Hochschule
Englisch (EN)	34	Grundschule → Hochschule
Deutsch (DE)	27	Grundschule → Hochschule

Zu den wichtigsten Verlagen zählen: - ZH: 人教版, 沪科版, 北师大版, 苏科版, 粤教版, 鲁科版, 马文蔚, 程守洙, 漆安慎, 赵凯华, 汪志诚, 杨福家, 梁昆淼, 郭硕鸿, 曾谨言 - EN: Khan Academy, CK-12, AP Physics 1/2/C, IB, IGCSE, GCSE, Halliday Resnick Walker, Serway Jewett, Young Freedman, Griffiths, Kittel, Feynman Lectures - DE: Duden, Lambacher Schwere, Westermann, Cornelsen, Klett, Auer, Dorn-Bader, Kern, Thieme, Tipler, Demtröder, Jackson, Arfken, Schiff, Greiner

6.4 Interpretation

Die kontrastierenden CDS-Muster über die Disziplinen hinweg deuten darauf hin, dass die Wissensorganisation disziplinabhängig und nicht universell ist:

- Physik folgt einem Konvergieren-dann-Divergieren-Muster: eng verbundene Grundlagenkonzepte (elementare Mechanik) vor der Ausbreitung in unabhängige Teilgebiete (Thermodynamik, E&M, Optik, moderne Physik).
- Mathematik folgt einem Aufbauen-dann-Ausbreiten-Muster: grundlegende Arithmetik integriert sich in der Mittelstufe und verzweigt sich anschließend in spezialisierte Domänen.

Dieses Muster steht im Einklang mit der pädagogischen Struktur des Physikunterrichts, in dem die Mechanik die konzeptuelle Grundlage liefert, die alle späteren Teilgebiete durch Kraftanalyse und Energieerhaltung vereint.

6.5 Abbildungen

- Abbildung 6: CDS-Vergleich — Mathematik vs. Physik nach Bildungsstufe (cognitive-space/figures/fig6_cds_comparison.png; Spiegel: cognitive-space/web/figures/, _deploy/figures/)

6.6 Datenbestände

Bestand	Konzepte	Relationen	Status
Mathe-Wissensgraph	556	525	Vollständig
Physik-Wissensgraph	366	383	Vollständig
Chemie-Wissensgraph	220	215	Vollständig

Weitere Lehrplangraphen (z. B. NRW, chinesische Lehrpläne) liegen außerhalb des Fokus dieser Arbeit.

6.7 Chemie: Konzeptdichtestruktur und Hierarchie (220 Konzepte)

Niveau	Chemie CDS	Mathe CDS	Verhltnis
Mittelstufe	0,042	0,271	0,15
Oberstufe	0,030	0,073	0,41
Hochschule	0,013	0,042	0,30

Der Chemie-Korpus enthält keine Grundstufen-Konzepte (Chemieunterricht beginnt im untersuchten Lehrbuchbestand erst in der Mittelstufe), daher entfällt die Grundstufen-Zeile.

HDS-Metrik	Chemie	Physik	Mathe
Maximale Tiefe	3	6	8
Mittlere Tiefe	0,20	0,85	0,40

HDS-Metrik	Chemie	Physik	Mathe
Wurzelkonzepte	184 (84 %)	219 (60 %)	459 (83 %)

◆ Befund F8: Der Chemie-CDS erreicht seine Spitze in der Mittelstufe (0,042) — dasselbe „Integriere früh, trenne spät"-Muster wie die Mathematik (0,271), jedoch auf deutlich niedrigerem Niveau. Die geringere Dichte und die flache Hierarchie (maximale Tiefe 3, 84 % Wurzelkonzepte) reflektieren den kleineren, weniger kumulativen Korpus (220 Konzepte, keine Grundstufe). Damit bestätigen die Chemie-Daten die Befunde F6 (Physik: Spitze in der Grundschule) und F7 (Physik: tiefere Voraussetzungsketten) als disziplinübergreifend konsistent: Die Wissensorganisation ist disziplinabhängig (Dichtespitzen auf verschiedenen Bildungsstufen), folgt aber in allen drei MINT-Disziplinen demselben frühen Integrationsmuster.

7. Exploratorische Sprachproduktionsanalyse (LPA)

7.1 Einordnung

Parallel zur strukturierten Wissensgraphenanalyse (LDS-K, §3) und der Konzeptanalyse menschlicher Probanden (LDS-C, §4) wurde ein dritter, unabhängiger Datenstrom erhoben: eine explorative Sprachproduktionsstudie. Zehn linguistische Aufgaben wurden an deutsche Muttersprachler verteilt (N=6 Pilot, N≥120 im Feld geplant). Die Aufgaben umfassen freie Assoziation, zweisprachige Konzepterklärung, soziale Skripte, visuelle Beschreibung, Raumsprache (Bewegung + Statik), Zeitreferenz, Ereignisbeschreibung und Namensstrategien.

Die LPA ist als exploratives Codierframework konzipiert — nicht als validiertes psychometrisches Instrument. Das vollständige Codebook mit Definitionen, Inklusions-/Exklusionskriterien und Beispielen ist in docs/lpa_codebook.md dokumentiert (vgl. Krippendorff, 2018, zur inhaltsanalytischen Methodik).

7.2 Codierdimensionen

Vier Dimensionen wurden mit einem a priori definierten Kriterienkatalog codiert. Jedes Kriterium wird binär codiert (vorhanden/nicht vorhanden); es werden keine zusammengesetzten Scores gebildet.

D1 — Räumliche Granularität (Aufgaben 5–6): Erfasst den Detailgrad räumlicher Beschreibungen. Die Motion-Event-Codierung folgt Talmy (2000, Vol. I) zur Ereignistypologie (Manner vs. Path); die statische Raumcodierung folgt Levinson (2003) zu räumlichen Referenzrahmen und Pederson et al. (1998) zur semantischen Typologie. Die Thinking-for-Speaking-Hypothese (Slobin, 1996) motiviert die Annahme sprachspezifischer Muster der Raumkodierung.

D2 — Zeitliche Rahmung (Aufgabe 7): Erfasst die Auflösung der Ambiguität räumlicher Metaphern für Zeit (Lakoff & Johnson, 1980, Kap. 9). Die Konstruktion moved forward ist ambigue zwischen früher (Time-moving) und später (Ego-moving). Die Kategorien folgen Boroditsky (2000), die diese Ambiguität erstmals systematisch experimentell nachwies.

D3 — Konzeptuelle Flexibilität (Aufgaben 2, 3, 8, 9): Erfasst Strategien bei konzeptuell anspruchsvollen Aufgaben: bilinguale Repräsentation (Pavlenko, 2005), soziale Skripte (Goffman, 1967), Ereigniskonstruktion (Talmy, 2000, Vol. II; Croft, 2012).

D4 — Lexikalische Produktion (Aufgaben 1, 10): Erfasst lexikalische Strategien — semantische Netzwerkstruktur (Collins & Loftus, 1975) und Neologismenbildung (Clark, 1993; Štekauer, 2005).

7.3 Pilotergebnisse (N=6 DE)

D1 Räumliche Granularität: Kriterienerfüllung von 3/9 bis 9/9. DE06 codierte Bewegungspfad (Quelle: aus, Pfad: durch, Ziel: betritt), statische Objekte mit Schreibseitenorientierung (Schreibseite nach unten) und vertikaler Relation (über dem Buch) — eine Kombination intrinsischer und relativer Referenzrahmen (vgl. Levinson, 2003). DE02 codierte nur ein Objekt ohne räumliche Relation. Die beobachtete intra-linguale Variation ist konsistent mit Slobin (2000).

D2 Zeitliche Rahmung: 1/6 eindeutig (T+: vorverlegt), 1/6 nicht-standardisiert (T~: vorwärts bewegt), 3/6 ambigu (T?: verschoben), 1/6 unvollständig (T-). Die Ambiguitätsvermeidung (3/5 Bearbeitende wählten richtungsneutrale Formulierungen) ist konsistent mit Boroditsky (2000) und Gentner et al. (2002).

D3 Konzeptuelle Flexibilität: Bilinguale Parallelerklärungen von 2/6 Probanden, inklusive Code-Switching (DE → EN innerhalb einer Äußerung), konsistent mit Pavlenko (2005) und Grosjean (2010). Sechs soziale Skriptstrategien beobachtet. Perspektivwechsel bei einem Probanden.

D4 Lexikalische Produktion: Assoziative Kategorien umfassten Messung, Institution, Zeitdruck und Kalender. Benennungsstrategien von funktional-präzise bis kreativ-neologistisch und bilingual hybrid (vgl. Štekauer, 2005).

7.4 Methodische Anmerkungen

Diese Analyse ist explorativ. Vor systematischen cross-linguistischen Vergleichen:

1. Inter-Rater-Reliabilität: Cohen's $\kappa \geq 0{,}70$ angestrebt (Landis & Koch, 1977). Zwei unabhängige Codierer, 20 Antworten.
2. Dimensionsanalyse: Abhängig von Stichprobengröße und Verteilungseigenschaften wird exploratorische Faktorenanalyse oder PCA durchgeführt.
3. Cross-linguistische Erweiterung: Äquivalente ZH/EN-Fragebögen erstellt.

7.5 Methodologische Einordnung

Die drei Evidenzströme befinden sich in unterschiedlichen Reifestadien:

Strom	Typ	Reife
LDS-K	Graphentheoretischer Vergleich, validierte Nullmodelle	Hoch
LDS-C	Konzeptgraph-Vergleich	Mittel (N≥30 ausstehend)
LPA	Exploratives qualitatives Codierframework	Früh (Codebook, IRR ausstehend)

Die LPA-Ergebnisse werden unabhängig von LDS-K und LDS-C berichtet. Konvergente Muster können auf Zusammenhänge hinweisen; divergente Muster sind gleichermaßen informativ.

8. Diskussion

8.1 Zusammenfassung der Ergebnisse

Diese Studie führte LinguaGraph ein, ein wissensgraphbasiertes Rahmenwerk zur Analyse, wie mathematisches Wissen über Sprachen (Chinesisch, Deutsch, Englisch), Bildungsstufen (Grundschule bis Universität) und zuletzt auch Disziplinen (Mathematik vs. Physik) hinweg organisiert ist. Zwölf zentrale Befunde ergaben sich (F11 und F12 wurden auf Basis der erweiterten N=15-Humanvalidierung revidiert):

#	Befund	Evidenz
F1	CDS erreicht Spitze in der Mittelstufe (0,271), nicht in der Grundschule	Nicht-monotonisches Dichtemuster
F2	3,7-facher Dichteabfall von der Mittel- zur Oberstufe	CDS 0,271 → 0,073; Konzeptanzahl 4,2-facher Anstieg
F3	HDS ⩽ 8 (Mittelwert 0,40); 83 % der Konzepte sind Wurzeln	Mathematik ist ein flaches Netz, kein tiefer Baum
F4	Lehrbuch-LDS-K variiert stark (0,519 ZH-DE bis 0,938 DE-EN)	Strukturdominiert, nicht sprachgetrieben
F5	Nullmodell falsifiziert LDS-K als Sprachmetrik: Voll < Struktur-Null	Gradverteilung dominiert; ΔLDS ist zentral
F6	Verschiedene Disziplinen zeigen verschiedene CDS-Muster	Mathematik Spitze in Mittelstufe; Physik Spitze in Grundschule
F7	Physik weist tiefere Voraussetzungsketten auf (HDS-Mittel 0,85 vs. 0,40)	Physikalisches Wissen ist stärker kumulativ
F8	Chemie-CDS erreicht ebenfalls Spitze in der Mittelstufe (0,042)	Konsistent mit „Integriere früh, trenne spät“-Muster
F9	Coverage-Scores variieren drastisch zwischen Bildungssystemen (12,7 – 95,4 %)	Lehrplandesign-Philosophie treibt Unterschiede
F10	China zeigt nahezu perfekte Übereinstimmung (95,4 %); NRW am niedrigsten (12,7 %)	Zentralisiertes vs. föderales Systemmerkmal
F11	Falsifiziert (N=15): LDS-C auf Konzeptebene (0.93 – 0.96) ist nicht von Teilnehmervariabilität unterscheidbar (Split-Half-Boden 0.92 – 0.96; Label-Permutation 0.94). Keine konsistente sprachspezifische Rangordnung.	Nullmodell-Suite (Within-Lang Split-Half, Label-Permutation)
F12	Überarbeitet (N=15): Der Unterschied zur Simulationsbasislinie (0.647) ist ein Artefakt der Extraktions-/Alignierungsmethode (höhere LDS-Skala), nicht ein Beleg für sprachliche Kognition. LDS-C ≈ Within-Language-Null → kein über Zufall hinausgehendes Sprachsignal.	LDS-C ≈ Split-Half-Boden; ΔLDS ≈ 0 (0.05 bis +0.05)

8.2 Interpretation des CDS-Gipfels

Der Befund, dass der Concept Density Score seinen Höhepunkt in der Mittelstufe erreicht (F1) und nicht in der Grund- oder Hochschule, bedarf einer sorgfältigen Interpretation. Eine naive Erwartung könnte lauten, dass „höher entwickeltes Wissen dichter vernetzt ist.“ Die Daten widersprechen dem: Der Mathematiklehrplan der Mittelstufe fungiert als Wissensdichteknotenpunkt, an dem grundlegende Arithmetik, einfache Algebra, Geometrie und Wahrscheinlichkeitskonzepte eng miteinander verbunden sind. Dieses Muster ist konsistent mit Ausubels Assimilationstheorie [12], die vorhersagt, dass Wissensstrukturen in Phasen der Konsolidierung vor der Aufspaltung in Spezialisierungen eine maximale Integration erreichen.

Der anschließende 3,7-fache Abfall von der Mittel- zur Oberstufe (F2) fällt mit einer 4,2-fachen Ausweitung der Konzeptanzahl zusammen, was darauf hindeutet, dass der Mathematiklehrplan an diesem Übergang bewusst diversifiziert wird. Dies könnte ein pädagogisches Gestaltungsprinzip widerspiegeln: Die Mittelstufe vermittelt eine integrierte Grundlage; die Oberstufe führt spezialisierte Teilbereiche ein (Analysis, Vektorgeometrie, Statistik), die in relativer Isolation gelehrt werden, bevor eine mögliche Reintegration auf universitärem Niveau erfolgt.

Die Robustheit dieses Befundes über drei Sprachen hinweg (ZH, EN, DE) deutet darauf hin, dass es sich nicht um ein Artefakt einer bestimmten Lehrbuchtradition handelt. Vielmehr könnte es eine universelle Eigenschaft der mathematischen Lehrplangestaltung widerspiegeln — oder zumindest eine Konvergenz über drei unterschiedliche Bildungssysteme hinweg.

8.3 Sprachübergreifende strukturelle Divergenz: Eine Nullmodell-Kritik

Die LDS-K-Ergebnisse (F4) zeigen eine erhebliche Variation zwischen den Sprachpaaren: ZH-EN=0,934, DE-EN=0,938, ZH-DE=0,519. Der ZH-DE-Wert sticht hervor — chinesische und deutsche Lehrbuchwissensstrukturen sind beträchtlich ähnlicher (niedrigerer LDS-K) als jede von beiden im Vergleich zum Englischen. Dies stellt unmittelbar die naive Erwartung in Frage, dass typologisch entfernte Sprachen (ZH-DE) die größte Divergenz aufweisen würden.

Um zu bestimmen, ob diese Werte echte sprachgetriebene strukturelle Unterschiede darstellen, wandten wir eine Nullmodell-Suite mit vier Bedingungen an:

Bedingung	Beschreibung	ZH-EN	DE-EN	ZH-DE
Voll (LDS-K-Baseline)	Realer Graphenvergleich	0,934	0,938	0,519
Struktur-Null	Grad-erhaltende Kantenumordnung (×1000)	0,957	0,957	0,717
Knotenpermutations-Null	Zufällige Neuzuweisung von Knotenbezeichnungen	0,934	0,938	0,519
Vollständig zufällig	Erds – Rényi-Graph	1,000	1,000	1,000

Der entscheidende Befund: Vollständiges LDS-K < Struktur-Null-LDS-K für alle drei Sprachpaare. Unter grad-erhaltender Randomisierung (Doppelkantentausch, 1000 Iterationen) sind die randomisierten Graphen systematisch unterschiedlicher voneinander als die realen Graphen. Dies bedeutet, dass Lehrbuchwissensstrukturen stärker konvergieren, als der Zufall vorhersagen würde — das Gegenteil dessen, was eine sprachgetriebene Divergenzhypothese erwarten würde.

Dieses Ergebnis falsifiziert die Interpretation, dass LDS-K sprachgetriebene kognitive Divergenz misst. Stattdessen werden die hohen LDS-K-Werte von der Gradverteilungsstruktur dominiert — einer Eigenschaft, die sprachübergreifend geteilt wird, weil mathematische Voraussetzungslogik universell ist. Wenn die Gradverteilungen erhalten bleiben (Struktur-Null), sinkt die strukturelle Ähnlichkeit, was zeigt, dass das, was Lehrbuchgraphen „ähnlich“ macht, ihre gemeinsame Gradstruktur ist, nicht die sprachspezifische Inhaltsanordnung.

Die theoretische Implikation ist bedeutsam: Während mathematische Wahrheit universell ist, ist der hier erzielte Befund stärker — auch die organisatorischen Strukturen von Lehrbüchern sind sprachübergreifend bemerkenswert konvergent. Drei unterschiedliche Bildungstraditionen (Chinesisch, Deutsch, Englisch) produzieren unabhängig voneinander Lehrbuchwissensgraphen, deren strukturelle Eigenschaften (Gradverteilungen, Dichteprofile) einander ähnlicher sind als vergleichbare Graphen mit derselben Gradsequenz.

Dies bedeutet, dass der Korpusanalyse-Ansatz (LDS-K) per se keine sprachrelativistischen Effekte auf die Wissensorganisation messen kann. Er misst in erster Linie strukturelle Konvergenz, die von der universellen Logik mathematischer Voraussetzungen angetrieben wird. Um ein genuines Sprachsignal zu isolieren, müssen wir zur kognitiven Ebene übergehen — dem Vergleich, wie Menschen ihr Wissen in ihrer Muttersprache ausdrücken — erfasst durch ΔLDS = LDS-C LDS-K.

Die erweiterte Humanvalidierung (N=15, F11) liefert jedoch ein negatives Ergebnis für diese Verschiebung: Die menschlichen LDS-C-Werte (Konzeptebene ZH-EN=0,963, DE-EN=0,932, ZH-DE=0,936) liegen nahe der Within-Language-Null (Split-Half-Boden 0,92 – 0,96) und sind von Label-Permutation nicht unterscheidbar (0,94). ΔLDS $\approx$ 0 (0,05 bis +0,05), nicht > 0. Damit wird die naive Hypothese „Sprache → unterschiedliche Konzeptgraphen″ auf allen drei Analyseebenen (Konzept, Kategorie, Relation) falsifiziert.

8.4 Disziplinübergreifende Validierung

Die Hinzunahme der Physik (F6, F7) bestätigt, dass die CDS- und HDS-Metriken echte strukturelle Eigenschaften der Wissensorganisation erfassen und nicht lediglich Artefakte des Mathematik-Korpus darstellen. Die kontrastierenden Muster — Mathematik gipfelt in der Mittelstufe, Physik in der Grundschule — zeigen, dass Wissensorganisation disziplinabhängig ist, wobei beide demselben „Integriere früh, trenne spät“-Muster folgen, jedoch auf unterschiedlichen Bildungsstufen.

Dieser Befund hat Implikationen für die Lehrplangestaltung. Wenn Mathematik- und Physikstudierende grundlegend unterschiedliche Wissensdichtetrajektorien durchlaufen, dann sind pädagogische Strategien, die für eine Disziplin wirken, möglicherweise nicht auf die andere übertragbar. Der Mathematikunterricht könnte frühe Integration betonen; der Physikunterricht könnte akzeptieren, dass Integration auf fortgeschrittenem Niveau ein natürlicher Bestandteil des Lernverlaufs ist.

8.5 Die Lehrplanebene

Die Integration von Lehrplanstandards (Kernlehrplan NRW, UK National Curriculum, US NGSS/CCSS) in das Wissensgraphen-Rahmenwerk offenbart einen systematischen Befund: Die Übereinstimmung zwischen Lehrbuch und Lehrplan variiert dramatisch zwischen den Bildungssystemen:

System	Coverage-Score	Muster
China (CN)	95,4 %	Nahezu perfekte Ausrichtung (zentraler Lehrplan)
England (UK)	37,3 %	Mäßig; am höchsten in der Sekundarstufe II
Vereinigte Staaten (US)	17,2 %	Niedrig (breite Richtlinien, lokale Variation)
NRW Deutschland	12,7 %	Niedrigste (detaillierte, studiengangspezifische Vorgaben)

Der Coverage-Score misst die Lehrplan-→Lehrbuch-Übereinstimmung: Findet sich zu jedem Lehrplankonzept ein entsprechendes Konzept im Lehrbuchgraph? Die dramatische Spanne — von 12,7 % (NRW) bis 95,4 % (CN) — spiegelt grundlegende Unterschiede in der Bildungsgovernance wider: Zentrale Systeme erzeugen enge Ausrichtung; föderale Systeme mit studiengangspezifischen Vorgaben erzeugen von Natur aus niedrigere messbare Übereinstimmung.

8.6 Warum erzeugen Bildungssysteme unterschiedliche Wissensstrukturen?

Die erhebliche systemübergreifende Variation der Coverage-Scores (12,7 – 95,4 %) wirft eine über die Messung hinausgehende Frage auf: Was erklärt diese Unterschiede? Wir betrachten drei konkurrierende Erklärungsansätze.

Erklärungsansatz A: Granularität des Lehrplans (am besten gestützt)

Die sparsamste Erklärung ist, dass sich Lehrpläne in ihrer Granularität unterscheiden. Der NRW-Kernlehrplan spezifiziert 299 Mathematikkonzepte über 6 Stufen, während der UK National Curriculum ähnliche Inhalte mit 397 breiteren Deskriptoren abdeckt. Wenn ein Lehrplan Konzepte in feinerer Granularität definiert, kann jedes Lehrbuchkonzept definitionsgemäß weniger Lehrplankonzepte abdecken — was niedrigere Coverage-Scores unabhängig von der tatsächlichen Übereinstimmung der Inhalte erzeugt.

Dies wird durch das NRW-stufenweise Muster gestützt: Die Abdeckung erreicht ihren Höhepunkt in der Sek I (Klassen 7 – 8), wo sich der Lehrplan auf gemeinsame Kerninhalte konzentriert, und fällt in der Sek II (Klassen 11 – 13) ab, wo der Lehrplan spezialisierte Kurse (Grundkurse, Leistungskurse) mit feinkörnigeren Kompetenzerwartungen einführt.

Erklärungsansatz B: Bildungsphilosophie und Prüfungsstruktur (höherer Interpretationswert)

Das britische Muster (37,3 %) und das US-Muster (17,2 %) spiegeln unterschiedliche Bildungsphilosophien wider. Der UK National Curriculum bietet einen mäßig vorschreibenden Rahmen, an dem sich Lehrbücher auf Sekundarstufenebene orientieren. Die USA zeigen eine geringere Ausrichtung (17,2 %), was mit breiten, nicht vorschreibenden Richtlinien (NGSS/CCSS) konsistent ist, die lokale Anpassungen ermöglichen. Chinas nahezu perfekte Ausrichtung (95,4 %) ist konsistent mit einem zentralisierten Lehrplansystem, in dem Lehrbücher nach expliziten nationalen Standards verfasst werden.

Diese Interpretation deckt sich mit der vergleichenden Bildungsforschung: Schmidt et al. [54] fanden, dass die Kohärenz von Lehrplänen zwischen TIMSS-Ländern erheblich variiert, wobei China eine hohe Übereinstimmung zwischen intendierten und implementierten Lehrplänen aufweist. In jüngerer Zeit dokumentiert die OECD-Veröffentlichung „Education at a Glance" [7][32], dass föderale Strukturen eine stärker variierende Lehrplanumsetzung hervorbringen als zentrale Systeme.

Erklärungsansatz C: Arbeitsteilung zwischen Lehrplan und Lehrbuch (am differenziertesten)

Eine dritte Möglichkeit ist, dass sich die Beziehung zwischen Lehrbuch und Lehrplan in den verschiedenen Systemen grundlegend unterscheidet. In der deutschen Tradition legen Lehrpläne minimale Kompetenzstandards fest, während Lehrbücher erhebliche Autonomie in der Wissensorganisation ausüben. Im chinesischen System werden Lehrbücher direkt aus dem nationalen Lehrplan entwickelt, was eine nahezu perfekte Ausrichtung (95,4 %) erzeugt.

Nach dieser Interpretation ist der niedrige Coverage-Score NRWs (12,7 %) kein Mangel, sondern ein Merkmal: Deutsche Lehrbücher sind darauf ausgelegt, alternative organisatorische Strukturen anzubieten, die den Lehrplan ergänzen statt zu duplizieren. Dies würde vorhersagen, dass NRW-Lehrbücher eine HÖHERE interne strukturelle Diversität (mehr Variation zwischen Verlagen) aufweisen als chinesische Lehrbücher — eine testbare Hypothese für zukünftige Arbeiten.

Synthese

Die drei Erklärungen schließen sich nicht gegenseitig aus. Die Lehrplangranularität (A) ist die sicherste Interpretation, die Bildungsphilosophie (B) bietet die reichhaltigste Erzählung und die Arbeitsteilung zwischen

Lehrplan und Lehrbuch (C) eröffnet die interessantesten Forschungsfragen. Unsere Daten sind mit allen drei konsistent, aber ihre Beurteilung erfordert zusätzliche Evidenz — insbesondere systemübergreifende Analysen der Lehrplankonzept-Granularität und der inhaltlichen Diversität von Lehrbüchern.

Diese Herausforderung — die Trennung von Messungseffekten und echten strukturellen Unterschieden — ist selbst ein Beitrag: Sie zeigt, dass systemübergreifende Bildungsvergleiche eine sorgfältige Beachtung der Struktur des Referenzstandards erfordern, nicht nur des Lehrbuchgraphen.

8.7 Extraktionszuverlässigkeit und Fehleranalyse

Ein potenzielles Bedenken bei jeder LLM-basierten Analyse ist, ob Messfehler die berichteten Ergebnisse verursachen könnten. Unsere Extraktionsvalidierung anhand von 92 Goldstandard-Annotationen (ZH F1=0,974, DE F1=0,949, EN F1=0,882) deutet darauf hin, dass die Extraktionsqualität des früheren primären Extraktionsmodells (qwen-plus) insgesamt hoch ist. Die Fehleranalyse zeigt, dass 29 % der Extraktionsfehler bei sehr kurzen Antworten (1 – 2 Wörter) auftreten, bei denen eine leere Extraktion tatsächlich angemessen ist. Bei den verbleibenden Fehlern handelt es sich überwiegend um partielle Auslassungen — 1 – 2 Konzepte aus einer Liste von 3 – 4 fehlen — und nicht um systematische Fehlleitungen.

● Hinweis zur N=15-Erweiterung: Die erweiterte Humanvalidierung (§4) wurde mit deepseek-v4-flash (opencode GO) extrahiert; eine erneute F1-Validierung gegen die Goldstandards steht aus. Der 19-Modell-Benchmark (§8.9) zeigt jedoch, dass Konzeptextraktionsqualität über Modellfamilien hinweg eine Eigenschaft der Aufgabe ist (F1-Bereich 0,55 – 0,67), sodass eine modellinduzierte Skalenverschiebung der LDS-C zwar möglich, eine systematische Verzerrung der Null-Relation (beobachtet ≈ Split-Half) aber unwahrscheinlich ist.

Diese Fehlerverteilung bedeutet, dass die strukturellen Metriken (CDS, HDS, LDS, Coverage-Score) robust gegenüber Extraktionsrauschen sind: Partielle Auslassungen reduzieren die Konzeptanzahlen leicht, verzerren jedoch nicht systematisch die Graphentopologie oder die sprachübergreifenden Vergleiche. Wir halten es daher für unwahrscheinlich, dass die berichteten Befunde Artefakte der Extraktionsmethodik sind.

8.8 Robustheitsprüfung: Rechnerische Basislinie — methodisch revidiert

Zur Kontrolle, ob die beobachteten LDS-Werte echte strukturelle Unterschiede und nicht zufällige Konzeptvariation widerspiegeln, wurde ursprünglich eine rechnerische Basislinie aus 300 simulierten Antworten (20 pro Bedingung × 5 Themen × 3 Sprachen) berechnet. Die frühere Version dieser Analyse (N=8, qwen-plus-Extraktion) ergab einen scheinbaren Unterschied zwischen menschlichem LDS (0,727) und Simulation (0,647, p=0,05).

Diese Vergleich ist mit der N=15-Erweiterung nicht mehr haltbar, aus zwei Gründen:

1. Skalenverschiebung: Die neue, standardisierte Extraktion (deepseek-v4-flash) verschiebt LDS-C auf eine andere Skala (0,93 – 0,96); ein direkter Vergleich mit der alten Simulationsverteilung mischt zwei unterschiedliche Messverfahren.
2. Null-Test-Ergebnis: Die korrekte Kontrolle ist nicht der Vergleich zweier Absolutwerte, sondern die Lage der beobachteten LDS-C relativ zur Within-Language-Null (Split-Half 0,92 – 0,96; Label-Permutation 0,94). Beobachtete LDS-C liegt auf diesem Boden → vollständig durch Teilnehmervariabilität erklärbar.

Ein aussagekräftiger Mensch-vs-Simulation-Vergleich erfordert dieselbe Extraktions-/Alignierungsmethode für beide Bedingungen. Dies ist eine Aufgabe für zukünftige Arbeit (nicht mehr eine Stütze der ΔLDS > 0-Hypothese).

8.9 Robustheitsprüfung: Modellübergreifende Extraktionskonsistenz

Um zu überprüfen, ob die LDS-Ergebnisse nicht von einem einzelnen Extraktionsmodell getrieben werden, führten wir einen 19-Modell-Benchmark über drei API-Plattformen durch (DashScope, DeepSeek API, OpenCode GO). Alle Modelle extrahierten Konzepte aus denselben N≥50 Goldstandard-Items:

Rang	Modell	F1	N	Quelle
1	hy3-preview	0,6741	57	OpenCode GO
2	mimo-v2.5-pro	0,6735	75	OpenCode GO
3	qwen-plus	0,6659	92	DashScope
4	qwen-max	0,6610	92	DashScope
5 – 10	Gemischte Modelle	0,59 – 0,63	79 – 92	Gemischt
11 – 19	Untere Stufe	0,55 – 0,59	89 – 92	Gemischt

Zentrale Ergebnisse: (1) Alle 19 Modelle erreichen F1 > 0,55, was bestätigt, dass die Konzeptextraktion modellfamilienübergreifend robust ist. (2) Qwen-plus (das primäre Extraktionsmodell) belegt Rang 3 mit F1=0,666 und liegt damit innerhalb des oberen Clusters. (3) DeepSeek-Modelle (v4-pro mit 0,593, v4-flash mit 0,608) sind wettbewerbsfähig. (4) Der enge F1-Bereich (0,55 – 0,67) über verschiedene Architekturen hinweg (Qwen, DeepSeek, GLM, MinMax, Kimi, Mimo) deutet darauf hin, dass die Extraktionsqualität eine Eigenschaft der Aufgabe ist, nicht eines bestimmten Modells.

Ein sekundärer Befund: 186 zusätzliche DashScope-Modelle (Text, Vision, Sprache) erzeugten alle F1=0,0, was bestätigt, dass diese andere Prompting-Strategien erfordern. GPT-4o und GPT-4o-mini waren aufgrund von Guthabenbeschränkungen während des Benchmarks nicht mehr verfügbar.

8.10 Threats to Validity

Wir identifizieren sechs Hauptbedrohungen für die Validität der berichteten Ergebnisse.

Abhängigkeit vom Extraktionsmodell. Alle LDS-Berechnungen hängen von der Konzeptextraktion mittels qwen-plus ab. Während der 19-Modell-Benchmark die modellübergreifende Konsistenz bestätigt (F1-Bereich 0,55 – 0,67), könnte eine andere Extraktionsarchitektur systematisch andere Konzeptmengen erzeugen und möglicherweise die LDS-Werte verändern.

Diese Bedrohung wird durch den engen F1-Bereich über verschiedene Modellfamilien hinweg teilweise abgemildert.

Repräsentativität des Korpus. Der Mathematik-Korpus (68 Lehrbücher) ist umfassend, jedoch verzerrt: Chinesische Lehrbücher stammen überwiegend von einem einzigen Verlag (Renjiao), deutsche Lehrbücher sind in Richtung universitärer Materialien verschoben und der englische Korpus ist auf IGCSE-/IB-Rahmenwerke beschränkt. Die sprachinterne Split-Half-Null (LDS≈0,97) quantifiziert diese Bedrohung.

Übersetzungsasymmetrie bei der Konzeptausrichtung. Die sprachübergreifende Konzeptausrichtung hängt von Expertenurteilen ab. Fehlausrichtungen erhöhen den LDS, indem sie zur Vereinigungsmenge beitragen, ohne die Schnittmenge zu erhöhen. Diese Bedrohung wird durch die Struktur der aligned_groups teilweise kontrolliert, aber die asymmetrische Abdeckung bleibt eine Quelle von Messfehlern.

Auswahlverzerrung der Lehrpläne. Lehrplandokumente variieren in ihrer Granularität (NRW: 299 Konzepte, UK: 397, US: 2.124, CN: 87). Höhere Granularität senkt mechanisch die Coverage-Scores. Unsere stufenweise Analyse kontrolliert dies teilweise.

Begrenzte Stichprobengröße und Between-Subject-Design (Pilot-Menschendaten). Die ΔLDS-Analyse beruht auf N=15 Teilnehmenden (6 DE, 6 ZH, 3 EN) in einem Between-Subject-Design. Die Nullmodell-Suite zeigt, dass bei dieser Stichprobengröße die Teilnehmervariabilität die Sprachgruppe dominiert: sprachgetriebene Divergenz kann nicht von individueller Variabilität getrennt werden. Eine Post-hoc-Power-Rechnung (scripts/sw_fix_analyses.py, t-Test-Approximation) bestätigt dies: Bei kleinen bis mittleren Effekten (d=0,2 – 0,5) hat ein 6-gegen-6-Between-Vergleich nur Power 0,06 – 0,12, ein gepaarter Within-Vergleich mit N=15 dagegen 0,11 – 0,44 (N=30: 0,19 – 0,75). Die menschliche Null ist damit auch power-analytisch erwartbar — kein Beleg für Abwesenheit, sondern eine Grenze des Designs selbst. Die registrierte Folgeuntersuchung (R1) nutzt folgerichtig Within-Subject mit N≥30 pro Arm.

Reichweite des Nullmodells. Die grad-erhaltende Struktur-Null testet die Kantenanordnung über die Gradstruktur hinaus, jedoch nicht, ob die Gradstruktur selbst sprachbeeinflusst ist. Ein zukünftiges hierarchisches Nullmodell könnte diese Frage angehen.

8.11 LDS-Interpretationsrahmen

Anstatt willkürliche Schwellenwerte für LDS-Werte festzulegen, verankern wir die Interpretation an der Nullmodell-Suite:

LDS-Bereich	Interpretation	Anker
> 0,97	Vollständige Divergenz	Über dem sprachinternen Rauschboden
0,90 – 0,97	Typische sprachübergreifende Divergenz	Nahe am Rauschboden
0,50 – 0,90	Partielle Konvergenz	Unterhalb des Rauschbodens, oberhalb der Struktur-Null
0,00 – 0,50	Erhebliche Konvergenz	Deutlich unter allen Nullenwartungen

In diesem Rahmen: - Sprachinterner Rauschboden ≈ 0,97 (Split-Half) → Obergrenze für aussagekräftigen Vergleich - Struktur-Null ≈ 0,96 (graderhaltend) → strukturelle Basislinie - Vollständig zufällig = 1,00 → Plausibilitätsprüfung

ZH-DE (0,519) liegt im Bereich der „partiellen Konvergenz“ — deutlich unter den Nullenwartungen. ZH-EN (0,934) und DE-EN (0,938) liegen im Bereich „nahe am Rauschboden“ — nicht unterscheidbar von zwei zufälligen Hälften desselben sprachlichen Lehrbuchgraphen. Hierbei handelt es sich in erster Linie um Beobachtungen und nicht um Erklärungen; der Mechanismus, der die Heterogenität zwischen den Paaren antreibt, erfordert weitere Untersuchungen.

8.12 Einschränkungen

Mehrere Einschränkungen sollten anerkannt werden:

Umfang der Daten. Der Mathematik-Korpus (68 Lehrbücher, 556 Konzepte) ist umfassend. Der Physik-Korpus (366 Konzepte) und der Chemie-Korpus (220 Konzepte) ermöglichen eine disziplinübergreifende Validierung, bleiben aber kleiner. Die Lehrplangraphen decken zwar vier Systeme ab (NRW, UK, US, China), verwenden jedoch unterschiedliche Abgleichsmethoden, die die Vergleichbarkeit beeinträchtigen können.

Extraktionsmethodik. Während unsere Goldstandard-Validierung eine insgesamt hohe Qualität belegt (F1=0,881 gewichtet gesamt; 0,939 nur Sozial-Subgruppe, n=72), wurden die sozialwissenschaftlichen Golddaten mittels halbautomatischen Schlüsselwortabgleichs mit anschließender manueller Überprüfung validiert. Im Goldstandard selbst könnten einige Fehler verbleiben.

Stichprobengröße und Design der menschlichen Validierung. Die erweiterte Humanstudie (N=15, 6 DE • 6 ZH • 3 EN) bestätigt die frühere Einschränkung und verschärft sie: Bei N=15 in einem Between-Subject-Design dominiert Teilnehmervariabilität die messbare LDS-C vollständig (Split-Half-Boden ≈ beobachtete LDS-C ≈ Label-Permutation). Dies ist kein Stichprobenfehler, sondern eine strukturelle Grenze des Designs: Da jede Person nur in einer Sprache antwortet, können sprachgetriebene Divergenz und individuelle Variabilität nicht getrennt werden. Die zentrale methodologische Lehre: Zukünftige Studien benötigen (a) ein Within-Subject-Design, (b) größere Stichproben pro Sprachgruppe (N≥30), oder (c) Metriken, die gegen individuelle Konzeptwahl robust sind. Die thematischen Richtungstendenzen (ZH rechtlich/moralisch, DE autonom/affektiv, EN sozial) bieten eine testbare Hypothese für ein Within-Subject-Design.

Interpretation des Nullmodells. Während die graderhaltende Struktur-Null zeigt, dass LDS-K von strukturellen Faktoren und nicht von der Sprache dominiert wird, bewahrt der Doppelkantentausch-Algorithmus die exakte Gradsequenz jedes Graphen. Dies ist eine konservative Null: Sie testet, ob sprachspezifische Kantenanordnungen über die Gradstruktur hinaus Informationen hinzufügen, testet jedoch nicht, ob die Gradstruktur selbst sprachbeeinflusst sein könnte. Ein zukünftiges hierarchisches Nullmodell könnte diese geschichtete Frage angehen.

Kantenfreie Graphen in früheren menschlichen Daten. Die frühere qwen-plus-Extraktion erzeugte eine rein konzeptbasierte Ausgabe (keine Relationen) für menschliche Antworten. In der N=15-Erweiterung wurde dies behoben: Eine Relationsextraktion (7 Typen, scripts/lds_c_extract_relations.py) lieferte 162 Kanten (DE 59, ZH 67, EN 36), sodass die vollständige v3-LDS-Formel (Node- + Edge-Jaccard) angewendet werden konnte. Das Ergebnis bestätigt die Konzeptebene: Edge-Jaccard ≈ 0 und Split-Half-Boden ≈ beobachtete v3-LDS — kein separierbares Sprachsignal. Die Kantenzahl pro Antwort bleibt jedoch spärlich, sodass Kantenaussagen auf Gruppenebene robust, auf individueller Ebene aber eingeschränkt interpretierbar sind.

Kausalität. Unsere Analyse ist korrelativ. Wir messen strukturelle Unterschiede zwischen Systemen, können diese jedoch nicht unabhängig voneinander auf Lehrplangestaltung, Lehrbuchtradition oder Bildungsphilosophie zurückführen.

Generalisierbarkeit. Mathematik, Physik und Chemie könnten strukturelle Eigenschaften teilen, die in geistes- oder sozialwissenschaftlichen Disziplinen nicht vorhanden sind. Die Erweiterung auf zusätzliche Domänen ist eine Priorität.

LDS-Interpretation. Der LDS-Interpretationsrahmen (Abschnitt 8.11) verankert numerische Werte an Nullmodell-Baselines, jedoch sind die Schwellenwerte (0,90, 0,50) deskriptiv statt inferenziell. Mit zunehmenden menschlichen Daten sollten Bootstrap-Konfidenzintervalle diese deskriptiven Schwellenwerte für Hypothesentests ersetzen.

8.13 Implikationen

Trotz dieser Einschränkungen haben die vorliegenden Ergebnisse Implikationen für drei Fachgemeinschaften:

Für die Bildungsforschung: Die CDS- und HDS-Metriken bieten quantitative Werkzeuge für die Lehrplananalyse, die bestehende qualitative Rahmenwerke (TIMSS, PISA) ergänzen. Eine Lehrplangestalterin könnte CDS verwenden, um Dichteengpässe zu identifizieren, und HDS, um übermäßig lange Voraussetzungsketten zu erkennen.

Für KI in der Bildung: Die automatisierte Pipeline zeigt, dass die groß angelegte, sprachübergreifende Konstruktion von Wissensgraphen aus Lehrbüchern mit aktuellen LLMs machbar ist. Dies eröffnet die Möglichkeit einer Lehrplanebenen-Wissensanalyse in einem Umfang, den die manuelle Inhaltsanalyse nicht erreichen kann.

Für die Erforschung des linguistischen Relativitätsprinzips: Unsere Daten stützen keine einheitliche „Sprache formt Wissen“-Behauptung. Stattdessen zeigen sie, dass sprachübergreifende strukturelle Beziehungen heterogen sind — einige Sprachpaare konvergieren erheblich (ZH-DE), während andere auf Rauschniveau liegen (ZH-EN, DE-EN). Die erweiterte Humanvalidierung (N=15) erlaubt eine präzise Antwort auf der Ebene menschlicher Konzeptäußerungen: In einem Between-Subject-Design ist bei dieser Stichprobe kein separierbares Sprachsignal nachweisbar. Das LLM-as-Subject-Experiment (§5) zeigt jedoch, dass dies ein Design-Artefakt ist: Unter Within-Subject-Bedingungen ist das Sprachsignal klar nachweisbar (LDS-C 0.93–0.96 Boden 0.85–0.87), mit dem Sprachcode als dominanter Organisationsebene. Dies falsifiziert die einfache Hypothese „Sprache → unterschiedliche Konzeptgraphen″ nur für das Between-Subject-Design, nicht für die Existenz sprachlicher Kognitionseffekte überhaupt. Die Richtungstendenzen (ZH rechtlich/moralisch, DE autonom/affektiv, EN sozial) bleiben als testbare Hypothesen für Within-Subject-Studien erhalten.

8.14 LLM-as-Subject: Methodologische Einordnung des Within-Subject-Befunds

Das LLM-as-Subject-Experiment (§5) liefert den methodologischen Schlüssel zur Interpretation des negativen Humanbefunds:

Design über Signal: Derselbe Messrahmen (LDS-C + Nullmodelle) produziert unter Between-Subject (Menschen) kein trennbares Signal, unter Within-Subject (LLM) ein deutliches (LDS-C übersteigt den Boden um +0.08 bis +0.09). Die Sprachlabels tragen unter Within-Subject-Bedingungen messbare Information.

Mechanismus: Das lineare gemischte Modell trennt die marginalen Beiträge — der Sprachcode dominiert (same_lang +0.038, $p<0.001$), der kulturelle Rahmen hat keinen signifikanten codeübergreifenden Beitrag (same_frame +0.001, $p=0.90$). Der Rahmen wirkt als code-interner Modulator; die codeübergreifende Divergenz wird von der lexikalisch-assoziativen Ebene getragen. Die freien Assoziationen desselben Modells erklären die Konzeptdivergenz für EN-Paare (M2), nicht aber für ZH-DE, wo eine strukturelle Ebene jenseits der Assoziationsstatistik existiert.

Validität des LLM als Subjekt: Die Verwendung eines LLM als kontrolliertem kognitivem Subjekt folgt etablierten Paradigmen [47][48] und beantwortet eine Frage, die mit menschlichen Between-Subject-Daten prinzipiell nicht beantwortbar ist. Das Modell ist kein Ersatz für menschliche Kognition, sondern ein Werkzeug zur Isolation der Sprachvariable bei konstantem „Gewichtssatz″. Diese Grenze ist zu betonen: Die Ergebnisse zeigen, dass Sprachcode und Rahmen die Konzeptwahl eines LLM verändern — sie extrapolieren nicht direkt auf menschliche Kognition, liefern aber eine testbare Blaupause (Within-Subject-Design, erwartete Effektrichtung) für zukünftige Humanstudien.

Präzisierung des Design-Artefakts (Design-Effekt-Beweis, §5.9.1–5.9.2): Der Vergleich auf identischem Messrahmen zeigt eine gleiche Signalamplitude bei Mensch und LLM (LDS-C jeweils 0.93–0.96); die menschliche Null ist nicht auf ″keine Sprachwirkung″ und nicht auf ″Between-Subject-Design an sich″ zurückzuführen (eine virtuelle Between-Subject-Analyse der LLM-Daten bei menschlicher N überlebt das Signal). Entscheidend ist die individuelle Heterogenität innerhalb einer Sprachgruppe: Sie hebt die Bodenlinie (0.92–0.96) auf das Signalniveau. Die Heterogenitäts-Injektion (§5.9.2b) zeigt als Konsistenz-Demonstration, dass eine ausreichende Aggregationssparsität die menschliche Null reproduzieren kann ($q=0.30$ → Marge +0.014 vs. menschlich +0.015) — wobei der mechanismus-konsistente Zielwert ($q\approx0.76$–0.91) die Marge nicht vollständig kollabiert (P2-Recheck). Die heterogenitätsbedingte Erklärung ist damit gestützt (Exklusion + Konsistenz), aber nicht als quantitative Kausalzuordnung zu werten. Dies hat eine direkte Designimplikation — Between-Subject-Studien zum Sprachrelativismus benötigen entweder ausreichende N pro Gruppe, um die Heterogenität zu schlagen, oder ein Within-Subject-Design, das die Varianzquelle eliminiert.

Strukturelle vs. lexikalische Divergenz (Knoten/Kanten, §5.9.4): Die soziale/institutionelle Umkehr ist knotengetrieben (Konzeptwahl), der Kantenbeitrag in der Mathematik am größten (0.074). Institutionelles Wissen konvergiert im Was (Konzepte), nicht im Wie (Beziehungen) — die relationale Organisation bleibt über alle Quellen

systemisch divergent. Dies trennt den Mechanismus "Konzeptwahl" von "Beziehungsstruktur" und stützt die M2-Dominanz (§5.6) auf der Korpusseite: die Divergenz entsteht primär auf der assoziativ-lexikalischen Ebene, nicht auf der relationalen. Einschränkung (P2-Recheck): Die Mathematik-Knoten sind Alignierungs-Labels (nicht unabhängig extrahierte Konzepte), das de-Label-Feld enthält teils chinesische Texte, und ein Size-Matching auf die Wikipedia-Größe kehrt das Muster um (bei k=15–35 ist der Wikipedia-J_node höher) — der mathematische J_node (0.556) ist damit teilweise ein Darstellungs-/Größen-Artefakt, nicht ein Beleg unabhängiger sprachübergreifender Konvergenz (Details: docs/p2_methodology_rechecks.md).

8.15 Modellübergreifende Replikation: Die Divergenz ist eine breit replizierbare, modellabhängige Eigenschaft

Ein zentraler Vorbehalt gegen den Within-Subject-Befund war dessen Beschränkung auf ein einzelnes Modell (deepseek-v4-flash). Eine Replikation des identischen P1-Protokolls (3 Sprachen × k=10, gleicher Messrahmen LDS-C/Floor/Nullmodelle) auf 54 weiteren Modellen (50 eindeutige Modell-Identitäten; 42 über DashScope, 7 über zen/OpenRouter plus D1-Baseline, 2 über Kilo, je 1 über Cohere/NIM/opencode-go) — DeepSeek-, GLM-, Kimi-, MiniMax- und Qwen-Familien sowie nemotron-3-ultra + nemotron-3-super (NVIDIA, US-Ursprung), laguna-s-2.1 (Poolside, US-Ursprung, auf zwei Hosts), gpt-oss-20b (OpenAI-Gewichte, US-Ursprung), command-a (Cohere, CA-Ursprung) und gpt-5.6-luna (Herkunft ungeklärt, offengelegt) — erweitert die Evidenz:

◆ Befund 1 — Signale replizieren breit, aber nicht ausnahmslos: Alle 55 Messungen zeigen LDS-C über dem jeweiligen Within-Language-Boden, und alle 55 ZH-DE-Paare sind signifikant ($p<0.05$; die meisten $p<0.01$). Von den 165 Sprachpaar-Tests insgesamt sind 8 nicht signifikant ($p\geqslant 0.05$) — alle acht betreffen englisch-haltige Paare (ZH-EN oder DE-EN), überwiegend bei DeepSeek-R1/Distill-Modellen (deepseek-r1-0528 DE-EN $p=0.672$). Die sprachübergreifende Wertedivergenz ist damit eine breit replizierbare Eigenschaft mehrsprachiger LLMs, aber mit modell- und sprachpaar-spezifischer Stärke — insbesondere ist die englisch-bezogene Divergenz in einigen Modellen statistisch nicht von der sprachinternen Varianz trennbar.

◆ Befund 2 — Divergenzmagnitude variiert modellabhängig (Rangordnung): Die ZH-DE-Marge reicht von +0.03 (deepseek-r1-0528) bis +0.42 (command-a-03-2025) — ein Faktor ~13 zwischen den extremsten Modellen. Für die Anwendung als KI-Validierungswerkzeug bedeutet dies eine quantifizierbare Rangordnung der sprachübergreifenden Drift-Sensitivität über Modelle hinweg. Hinweis: Absolute Margen sind über Modelle mit sehr unterschiedlicher Extraktionsdichte (z. B. command-a: Floor ~0,49 vs. typisch 0,70–0,87) nicht direkt vergleichbar; Signifikanz und Kulturrichtung sind es.

◆ Befund 3 — Die Kulturrichtung übersteigt das Zufallsniveau, am stärksten bei hoher Übereinstimmung: Eine Voten-Analyse der ZH-DE-Divergenztreiber (wie viele Modelle markieren ein Konzept als nur-in-DE bzw. nur-in-ZH) wird gegen ein frequenz-angepasstes Zufalls-Nullmodell getestet (beobachtete Gesamthäufigkeit jedes Konzepts fixiert, Richtung jeder Stimme zufällig mit $p=0.5$). Die beobachtete Zahl richtungskonsistenter Konzepte ($\max(DE,ZH)$-Stimmen $\geqslant t$) übersteigt das Nullmodell bei allen Schwellen ($p<0.001$), am stärksten bei hoher Übereinstimmung: $\geqslant$10 Stimmen 218 vs. 147±4; $\geqslant$20 Stimmen 61 vs. 12±2. Die stärksten Konzepte (Heimat:safety 46 Stimmen DE; Heimat:physical space 44 ZH-Stimmen; equal opportunity 42) stützen die DE-Autonomie/ZH-Raum-Orientierung. Wichtig: Die Stärke ist begrenzt — z. B. produzierten nur 46 von 56 Modellen Heimat:safety überhaupt als einseitigen Treiber, und 5 Modelle wurden auf zwei Hosts gemessen (Redundanz).

▲ Ehrliche Abgrenzung: Die Replikation umfasst 50 eindeutige Modell-Identitäten; 5 Modelle wurden auf zwei Hosts gemessen (zen + DashScope ×4, laguna zen + Kilo), was die Effektivität der "55" verringert; ~87 % der Messungen stammen von chinesischen Anbietern — der westliche Anteil umfasst sieben Messungen (NVIDIA nemotron-3-ultra + nemotron-3-super, Poolside laguna-s-2.1 auf zwei Hosts, OpenAI-Gewichte gpt-oss-20b, Cohere command-a, gpt-5.6-luna/Herkunft ungeklärt). Eine Anbieter-Stratifizierung der 55 vollständigen Messungen (scripts/sw_fix_analyses.py): CN-Stratum 48/48 ZH-DE-Paare signifikant (mittlere Marge 0,130), West-Stratum 7/7 signifikant (mittlere Marge 0,182) — richtungskonsistent, aber als Anbieter-Vergleich weiter unterpowert, daher Transparenzbericht. Dass Modelle herkunftsunabhängig stärker mit US-Werten alignieren, ist als Branchenmuster dokumentiert [42]; die Herstellererweiterung (Modell-Matrix) bleibt registrierte Folgearbeit. Die Divergenz ist als Messgröße (nicht als normatives "Problem") zu interpretieren: Ein Modell, das Wertkonzepte in der Zielsprache kulturadäquat rahmt, ist nicht notwendigerweise fehlerhaft. Als Vorschlag für einen operativen Schwellenwert ("hohe sprachübergreifende Divergenz") dient die ZH-DE-Marge (LDS-C Boden) $\geqslant$ 0.10 — bezogen auf die beobachtete Spanne von +0.03 bis +0.42 über 55 Messungen trennt dieser Wert die hoch-divergenten Modelle (command-a, kimi-k2.6, deepseek-v3.x, nemotron) von den niedrig-divergenten (deepseek-v4-flash/pro, r1-Distill). Dies ist eine heuristische Festlegung für die Anwendung, keine validierte Validitätsgrenze; eine kalibrierte Schwelle bleibt zukünftiger Arbeit überlassen. Eine Youden-Kalibrierung [46] auf den 55 vollständigen ZH-DE-Margen (Median-Split-Klassen, Bootstrap-CI über Modelle; scripts/sw_fix_analyses.py) stützt die Größenordnung: balancierte Trennung bei 0,13 (95 %-CI 0,13–0,14); die Heuristik 0,10 liegt knapp unterhalb des CI und wirkt damit sensitiv-inklusiv — quantifiziert, aber weiterhin keine validierte Grenze. Ehrlichkeitshinweis: Die Klassen stammen aus einem Median-Split derselben 55 Margen, sodass die Trennung ($J=1,0$) teilweise tautologisch ist; die Kalibrierung ist eine Konsistenzprüfung, keine unabhängige Validierung.

◆ Robustheit der Replikationsaussage (Sensitivitäten): Vier vorausgenommene Einwände wurden quantitativ geprüft. (a) Multiples Testen: Bei 165 Tests ohne Korrektur wären unter der globalen Null ~8 falsch-positive Treffer zu erwarten; die Permutationsauflösung (500 Iterationen) kann ein Bonferroni-Niveau ($0,05/165\approx 0,0003$) formal nicht auflösen. Entscheidend ist die Meta-Betrachtung: Alle 55 ZH-DE-Paare überschreiten sämtliche 500 Permutationen ($p<0,002$ einseitig je Test); unter der globalen Null wären $55\times 0,002\approx 0,11$ solcher Treffer erwartet — beobachtet wurden 55. (b) Doppelzählung: Dedupliziert auf 50 eindeutige Identitäten (je Dual-Host eine Messung) bleiben 50/50 ZH-DE-Paare signifikant; mittlere Marge 0,140 (vs. 0,136 über 55), Spanne unverändert +0,03…+0,42. Der Voten-Bias durch Doppelstimmen ist auf $\leqslant$10 Stimmen begrenzt und ändert keine Schwellen-Aussage ($\geqslant$10: 218 vs. 147±4; $\geqslant$20: 61 vs. 12±2). (c) luna-Provenienz: Ohne das Herkunft-ungeklärte luna-Modell bleiben 6/6 West-Messungen signifikant (mittlere Marge 0,176 vs. 0,182 mit luna). (d) Selektionsbias: Von 86 begonnenen Messungen sind 55 vollständig (je 10/10/10); 31 blieben unvollständig (API-Limits, Kosten-Stopp, Modell-EOL) — der Ausfall korreliert mit westlichen/ratelimitierten Kanälen und ist in docs/session_handoff_20260910.md dokumentiert. Die Complete-Case-Analyse bleibt damit eine Verfügbarkeitsstichprobe, kein präregistriertes Design. (e) Floor-Normierung: Die Rangordnungs-Aussage (Befund 2) wurde zusätzlich floor-normiert geprüft (Ratio = LDS-C/Boden, je Modell in den Replikationsdaten vorhanden): Die Rangkorrelation zwischen Marge und Ratio beträgt Spearman 0,997, die Top-5 sind identisch, command-a bleibt auch normiert Rang 1 (Ratio 1,87); West-Ratio im Mittel 1,284 vs. CN 1,169. Die Rangordnung wird daher nicht vom Floor-Niveau getrieben — der Kollaps-Hinweis zu command-a (Boden ~0,49) betrifft die absolute Deutung, nicht die Ordnung. (f) Familie×Sprachpaar: Die 8 nicht-signifikanten EN-Tests konzentrieren sich auf die R1/Distill-Familie —

6/10 EN-Tests bei 5 R1/Distill-Modellen vs. 2/100 bei allen übrigen 50 Modellen; der mittlere EN-Boden liegt bei R1/Distill höher (0,844 vs. 0,784). Die Konfundierung Sprache×Modellfamilie ist damit benannt und quantifiziert: Die EN-Hub-Deutung bleibt Hypothese, die Alternative „Reasoning-Verbosität hebt den Boden und schluckt das EN-Signal" ist nicht ausgeschlossen.

8.16 Kulturell-psychologische Einordnung der Divergenzrichtung

Die beobachtete Divergenzrichtung — DE-seitig Autonomie/Regeln/Ziel (equal opportunity, freedom limit, decision, goal, safety), ZH-seitig Raum/Boundary/Anspruch/Harmonie (physical space, boundary, deserved treatment, the weak, indulgence) — ist konsistent mit etablierten kulturell-psychologischen Rahmen:

Selbstkonstruktion (Markus & Kitayama, 1991) [37]: Die Theorie unterscheidet eine unabhängige (independent) von einer interdependenten Selbstkonstruktion. Individuell-orientierte Kulturen (typisch westlich) rahmen Werte über persönliche Autonomie, Leistung und regelbasierte Gerechtigkeit — deckungsgleich mit der DE-seitigen Treiberstruktur (Autonomie, eigene Ziele, Chancengleichheit). Interdependent-orientierte Kulturen (typisch ostasiatisch) rahmen Werte über relationale Grenzen, Harmonie und Fürsorge für die Gruppe — deckungsgleich mit der ZH-seitigen Struktur (Raum, Grenzen, Anspruch, Schutz der Schwachen, Nachsicht).

Individualismus/Kollektivismus (Hofstede, 1980, 2001) [38]: Deutschland weist einen hohen Individualismus-Index auf (≈67), China einen deutlich niedrigeren (≈20). Dies sagt vorher, dass deutsche Rahmungen Wertkonzepte individualistisch (Autonomie, Leistung) und chinesische kollektivistisch (Beziehung, Harmonie, Gruppenpflicht) organisieren.

Konfuzianischer Beziehungsraum: Die ZH-seitige Betonung von "Raum/Grenze/Anspruch" (物理空间, 边界, 应得) ist mit dem konfuzianisch geprägten relationalen Selbst vereinbar, das soziale Grenzen und den eigenen "Platz" in der Beziehungsordnung betont (vgl. Nisbett, 2003: holistische vs. analytische Kognition) [39].

▲ Ehrliche Einordnung: Diese Übereinstimmung ist eine interpretative Hypothese (die Richtung ist mit den Rahmen konsistent), kein Test der Theorien. Die Richtung ist modell- und trainingsdatenspezifisch und wird als beschreibend berichtet; eine kausale Zuordnung (Sprache → Kulturrahmen) ist nicht beansprucht.

9. Schlussfolgerung

9.1 Zusammenfassung der Beobachtungen

Diese Studie stellt LinguaGraph vor, ein wissensgraphbasiertes Framework zur Messung, wie Wissen über Sprachen (Chinesisch, Deutsch, Englisch), Disziplinen (Mathematik, Physik, Chemie) und Bildungssysteme (NRW, UK, US, China) organisiert ist. Das konsistente strukturelle Muster über alle drei Disziplinen hinweg ist:

Die Organisation von Bildungswissen folgt einem universellen „früh integrieren, spät divergieren"-Muster: Die maximale Verbindungsdichte tritt in den grundlegenden Stufen (Grundschule/Mittelstufe) auf — über alle Disziplinen und alle drei Sprachen hinweg — und nimmt dann mit zunehmender Spezialisierung monoton ab.

Die sprachübergreifende LDS-Analyse zeigt ein differenzierteres Bild. Anstelle eines einheitlichen „Spracheffekts" beobachten wir heterogene sprachübergreifende Strukturbeziehungen:

Sprachpaar	LDS-K (Lehrbuch)	Interpretation
ZH – DE	0,519	Weitgehend konvergent (indikativ) — weit unterhalb der sprachinternen Rauschschwelle (0,97), aber mit P2-Recheck-Einschränkung (Alignierungs-Labels, s. §8.14.5)
ZH – EN	0,934	Nahe der Rauschschwelle — nicht von sprachinterner Variation unterscheidbar
DE – EN	0,938	Nahe der Rauschschwelle — nicht von sprachinterner Variation unterscheidbar

Die entscheidende Erkenntnis ist nicht, dass „Sprache die Wissensstruktur beeinflusst", sondern dass verschiedene Sprachpaare systematisch unterschiedliche Grade struktureller Konvergenz aufweisen, wobei ZH-DE ein Muster zeigt, das Standard-Nullmodelle (grad-erhaltende Randomisierung, sprachinterne Split-Half) nicht erklären können. Diese Heterogenität — nicht Uniformität — ist die primäre Beobachtung, die unser Framework ermöglicht.

9.2 Drei Dimensionen der Struktur

Dimension	Erkenntnis	Grenze
Dichte (CDS)	ALLE Disziplinen erreichen Höhepunkt in frühen Stufen, dann Abfall	Mathe: 0,271 @ Mittelstufe; Physik: 0,222 @ Grundschule; Chemie: 0,042 @ Mittelstufe

Dimension	Erkenntnis	Grenze
Tiefe (HDS)	Voraussetzungsketten sind universell begrenzt	Max 8 (Mathe am tiefsten bei 8, Physik bei 6)
Divergenz (LDS-K)	Heterogen: ZH-DE konvergiert (0,52); ZH-EN und DE-EN auf Rauschniveau (0,93 – 0,94)	Sprachinterne Rauschschwelle: ~0,97
Abdeckung (CS)	Lehrbuch-Lehrplan-Abgleich variiert nach Governance-Modell	NRW 12,7%, UK 37,3%, US 17,2%, CN 95,4%

9.3 Wissenschaftlicher Kernbeitrag: Ein Framework zur Messung sprachübergreifender Strukturheterogenität

Der Kernbeitrag dieser Studie ist keine universelle Erkenntnis über Sprache und Kognition, sondern vielmehr ein methodologisches Framework, das heterogene sprachübergreifende Strukturbeziehungen sichtbar und quantifizierbar macht. Im Einzelnen:

1. LDS allein ist unzureichend — die Nullmodell-Suite zeigt, dass LDS-K-Werte gegen mehrere Basislinien interpretiert werden müssen (Struktur-Null, sprachinterne Rauschschwelle, kompletter Zufall)
2. LDS-K zeigt strukturelle Konvergenz, nicht Divergenz — alle drei Sprachpaare zeigen LDS-K-Werte auf oder unterhalb ihrer sprachinternen Rauschschwellen
3. ΔLDS = LDS-C LDS-K wird als interpretierbares Sprachsignal vorgeschlagen — die erweiterte N=15-Analyse zeigt jedoch ΔLDS ≈ 0 auf Konzeptebene (0.05 bis +0.05); die Relationsebene ist über unterschiedliche Sparsity-Regime hinweg nicht vergleichbar (ZH-DE +0.445 als Dichte-Artefakt, s. §5.9.4)
4. Die N=15-Humanvalidierung (6 DE • 6 ZH • 3 EN) falsifiziert ΔLDS > 0 unter Between-Subject-Bedingungen (Konzeptebene LDS-C 0.93 – 0.96 ≈ Split-Half-Boden 0.92 – 0.96 ≈ Label-Permutation 0.94; Themenebene x^2 p=0.52; Relationsebene Edge-Jaccard ≈ 0). Die früheren N=8-Pilotdaten (DE-ZH +0,232) werden nicht repliziert.
5. Das LLM-as-Subject-Experiment (Within-Subject, §5) weist das Sprachsignal nach — und stützt die Design-Artefakt-Hypothese für den Human-Negativebefund (Exklusion + Konsistenz-Demonstration; keine quantitative Kausalzuordnung, s. §8.14.4): Derselbe LDS-C-Messrahmen liefert unter Within-Subject-Bedingungen LDS-C 0.93 – 0.96 Boden 0.85 – 0.87, mit Sprachcode als dominantem Organisator (LMM: same_lang +0.038, p<0.001; same_frame +0.001, p=0.90).

Die zentrale methodologische Lehre: Between-Subject-Designs können sprachgetriebene Divergenz nicht von individueller Variabilität trennen; ein Within-Subject-Design ist hierfür erforderlich.

Der 19-Modell-Benchmark (F1-Bereich 0,55 – 0,67) und die Wikipedia-Negativkontrolle (nach Alignierung der sozialen Konzepte: reale LDS-Werte statt Artefakt 1,00 — siehe §3.8) bestätigen, dass diese Beobachtungen keine Artefakte der Extraktionsmethodik sind.

9.4 Beiträge

Wir stellen LinguaGraph vor, ein Framework, das:

1. Automatisch mehrsprachige Bildungswissensgraphen aus Lehrbüchern über ZH/EN/DE erstellt
2. Strukturelle Muster mittels vier graphbasierter Metriken (CDS, HDS, LDS, CS) quantifiziert
3. Eine Nullmodell-Grundlage zur Interpretation von LDS bereitstellt, die heterogene sprachübergreifende Strukturbeziehungen statt eines uniformen Spracheffekts aufdeckt
4. Über drei MINT-Disziplinen (Mathematik, Physik, Chemie) kreuzvalidiert
5. Lehrplanabgleich über vier Bildungssysteme integriert (NRW 12,7%, UK 37,3%, US 17,2%, CN 95,4%)
6. 19 mehrsprachige LLMs für Konzeptextraktion benchmarkt (F1-Bereich 0,55 – 0,67) und so die modellübergreifende Robustheit bestätigt

9.5 Einschränkungen

Die Studie hat fünf wesentliche Einschränkungen:

1. Extraktionsqualität variiert nach Domäne: Soziale Konzeptextraktion erreicht ZH F1=0,974, DE F1=0,949, EN F1=0,882 (72 Goldlabels im sozialen Bereich; 92 insgesamt inkl. Mathematik). Die mathematische Domänenextraktion ist niedriger (DE F1=0,506, 20 Goldlabels), was domänenspezifische Variation bestätigt.
2. Stichprobengröße und Design der menschlichen Validierung: Die erweiterte Humanstudie (N=15, 6 DE • 6 ZH • 3 EN) in einem Between-Subject-Design zeigt, dass Teilnehmervariabilität die LDS-C dominiert — sprachgetriebene Divergenz ist von individueller Variabilität nicht trennbar. Eine populationsbezogene Aussage erfordert ein Within-Subject-Design oder N ≥ 30 pro Sprachgruppe.
3. Umfang des Nullmodells: Das graderhaltende Nullmodell ist konservativ — es testet Kantenanordnung jenseits der Gradstruktur, aber nicht, ob die Gradstruktur selbst sprachbeeinflusst ist.
4. Lehrplanvergleich: Der Coverage Score verwendet keyword-basiertes Matching; zukünftige Versionen sollten semantische Alignierung integrieren.
5. Golddatensatzgröße: Aktuelle 92 Goldlabels insgesamt liefern zuverlässige Schätzungen über Domänen hinweg. Eine Erweiterung auf 200+ würde die statistische Aussagekraft für Untergruppenanalysen weiter stärken.

9.6 Zukünftige Arbeit

- Within-Subject-Humanstudie (dieselbe Person antwortet in mehreren Sprachen) zur Trennung von Sprach- und Teilnehmereffekten — das LLM-as-Subject-Experiment (§5) liefert hierfür eine Blaupause: erwartete Effektrichtung (Sprachcode dominant, abstrakte Themen am stärksten) und Design-Vorgaben (k=10 Wiederholungen, Split-Half-Boden als Referenz)

- Mehrsprachig feinabgestimmte Modelle für sprachübergreifende Konzeptextraktion mit höherem F1
- Thematische Richtungstendenzen als testbare Hypothesen (ZH rechtlich/moralisch, DE autonom/affektiv, EN sozial) in einem Within-Subject-Design
- Strukturrobuste Metriken zur Konzeptwahl-unabhängigen Divergenzmessung
- Zusätzliche Disziplinen (Biologie, Geschichte) zur Testung der „früh integrieren, spät divergieren″-Hypothese über Wissenstypen hinweg
- Semantischer Coverage Score mittels embedding-basierten Konzeptabgleichs
- Hierarchische Nullmodelle zur Trennung von Gradstruktureffekten von Kantenanordnungseffekten
- CognitiveSpace-Visualisierung zur interaktiven Erkundung sprachübergreifender Strukturunterschiede

9.7 Anwendung: Prüfung mehrsprachiger KI-Systeme

Über den akademischen Beitrag hinaus etabliert LinguaGraph ein Audit-Werkzeug für mehrsprachige KI-Systeme. Die Motivation ist konkret: LLMs werden heute in Dutzenden Sprachen bereitgestellt, aber überwiegend mit englischen Daten trainiert. Das LLM-as-Subject-Experiment zeigt, dass ein solches Modell wertbeladene abstrakte Konzepte (Gerechtigkeit, Freiheit, Verantwortung, Heimat, Erfolg) sprachabhängig strukturiert — mit statistisch signifikantem Sprachsignal (Permutationstest $p<0.01$) und kulturell gemusterten Divergenztreibern (DE: Autonomie/Regeln; ZH: Raum/Anspruch). Der Domänen-Vergleich (soziale Konzepte divergieren stärker als institutionelle Labels) ist hierfür indikativ, aber wegen der Alignierungsabhängigkeit der Mathematik-Knoten nicht als eigenständiger Beweis zu werten (P2-Recheck; Details: docs/p2_methodology_rechecks.md).

Der Output ist ein interpretierbarer Divergenzbericht pro Modell und Sprachpaar:

Nutzer	Entscheidung	Beitrag
Entwickler	Vorabprüfung vor mehrsprachiger Bereitstellung	LDS-C + konkret divergierende Konzeptbestandteile → gezielte Kalibrierung
Regulierer	Transparenz- und Dokumentationspflichten (z. B. EU AI Act)	Nachweis sprachabhängiger Modellstrukturierung
Forscher	Studie kultureller Werte in KI	Quantitative, interpretierbare Divergenzmessung

▲ Ehrliche Abgrenzung: Die Arbeit ist ein Messverfahren (Methodik) mit breiter, aber begrenzter Evidenz: Das Within-Subject-Experiment zeigt ein robustes Sprachsignal im Basismodell (deepseek-v4-flash), und die modellübergreifende Replikation (§5.10) bestätigt es in 55 Messungen (50 eindeutige Modelle, ~87 % chinesische Anbieter, sieben westliche Messungen aus US/CA-Herstellern; alle ZH-DE-Paare signifikant, 8 englisch-haltige Paare nicht; Kulturrichtung über Zufallsniveau). Die Arbeit ist jedoch kein fertiges Audit-Instrument: Eine produktive Anwendung erfordert die Definition eines Schwellenwerts für „kritische″ Divergenz, die Erweiterung auf westliche Modelle und weitere Konzepte/Sprachen sowie die normative Klärung, ob sprachspezifische Rahmung als „Drift″ oder als kulturadäquate Anpassung zu werten ist — dies ist Gegenstand zukünftiger Arbeit (§9.6).

Abschlusserklärung

Wissen in der Bildung folgt einer nichtlinearen strukturellen Organisation, die in ihrer frühzeitigen Integration universell und in ihrer Divergenzrate disziplinabhängig ist. Sprachübergreifende Strukturbeziehungen sind heterogen: Lehrbuchstrukturen konvergieren in unterschiedlichem Maße über Sprachpaare hinweg, wobei ZH-DE wesentlich stärkere Konvergenz zeigt als ZH-EN oder DE-EN relativ zu sprachinternen Basislinien (indikativ; P2-Recheck-Einschränkung zu Alignierungs-Labels s. §8.14.5). Die erweiterte Humanvalidierung (N=15) zeigt, dass auf Ebene menschlicher Konzeptäußerungen unter Between-Subject-Bedingungen kein von Teilnehmervariabilität separierbares Sprachsignal nachweisbar ist. Das LLM-as-Subject-Experiment (Within-Subject) weist das Sprachsignal dagegen klar nach und identifiziert den Sprachcode als dominante Organisationsebene mit dem kulturellen Rahmen als sekundärem, code-internem Modulator — die negative Humanhypothese ist damit mit der Design-Artefakt-Hypothese vereinbar (gestützt, nicht quantitativ kausal zugeordnet), kein Beleg für das Fehlen sprachlicher Kognitionseffekte. Das LinguaGraph-Framework macht diese unsichtbaren strukturellen Muster sichtbar, messbar und vergleichbar — und bietet eine methodologische Grundlage zur Untersuchung, wann, warum und in welchem Ausmaß sprachspezifische Wissensorganisation existiert. Als Audit-Werkzeug für mehrsprachige KI-Systeme überträgt LinguaGraph diese Grundlage auf die Praxis: Es macht sichtbar, ob und wo ein Modell wertbeladene Konzepte sprachabhängig strukturiert — ein bislang blinder Fleck der KI-Validierung — mit direktem Nutzen für Entwickler, Regulierer und Forscher.

LinguaGraph — 3 Kernschlussfolgerungen (Poster/Pitch-fähig)

Komprimierung der 12 Forschungsbefunde (F1–F12) auf 3 Hauptschlussfolgerungen Geeignet für: Poster, 3-Minuten-Vortrag, Abstract, Begutachtungsunterlagen

Schlussfolgerung 1: Die Wissensdichte ist nicht-monoton und disziplinabhängig

Was wir fanden:

Die Wissensdichte erreicht ihren Höhepunkt früh in beiden Disziplinen — Mathematik in der Mittelstufe (CDS=0,271), Physik in der Grundschule (CDS=0,222) — und fällt danach monoton ab.

Warum es relevant ist:

Die Annahme, dass „fortgeschritteneres Wissen dichter vernetzt ist", ist für beide Disziplinen falsch. Beide folgen einem Gipfel-und-Abfall-Muster, wobei die maximale strukturelle Integration auf der Einführungsstufe auftritt. Diese Konvergenz deutet auf eine universelle Eigenschaft der pädagogischen Wissensorganisation hin: Curricula sind darauf ausgelegt, die Verbindungsdichte in den Grundlagenphasen zu maximieren, bevor sie in die Spezialisierung divergieren.

Evidenz: | Disziplin | Spitzenniveau | Spitzen-CDS | Verlaufsform | |-----------|-----------|---------|-------| | Mathematik | Mittelstufe | 0,271 | ↑ Gipfel ↓ stetiger Abfall | | Physik | Grundschule | 0,222 | ↑ Gipfel ↓ rascher Abfall |

◆ Robustheit: Unabhängig in ZH, EN, DE bestätigt

Schlussfolgerung 2: Die Wissenstiefe hat eine universelle Obergrenze

Was wir fanden:

Die maximale Tiefe der Voraussetzungsketten ist auf HDS ⩽ 8 beschränkt, unabhängig von der Disziplin. Die mittlere Tiefe unterscheidet sich je nach Disziplin (Mathe: 0,40, Physik: 0,85).

Warum es relevant ist:

Pädagogisches Wissen scheint einer natürlichen Tiefengrenze für Voraussetzungsstrukturen zu unterliegen. Beide Disziplinen teilen dieselbe Obergrenze. Physik weist eine 2,1-fach höhere sequenzielle Tiefe auf als Mathematik (64 % Wurzeln in Physik vs. 83 % in Mathe), jedoch nicht die zuvor mit einer kleineren Stichprobe geschätzten 2,8-fache.

Evidenz: | Metrik | Mathe | Physik | |--------|:----:|:-------:| | Max. HDS | 8 | 6 | | Mittlere HDS | 0,40 | 0,85 | | Wurzelkonzepte | 83 % | 64 % | | Strukturtyp | Flaches Netz | Tiefere Kette |

Schlussfolgerung 3: Lehrbuchwissenschaftliche Strukturen konvergieren sprachübergreifend; ΔLDS isoliert das Sprachsignal

Was wir fanden:

Die lehrbuchbasierten LDS-K-Werte reichen von 0,519 (ZH-DE) bis 0,938 (DE-EN), doch eine Nullmodell-Kritik zeigt, dass diese Werte vollständig von der Gradverteilungsstruktur dominiert werden — nicht von der Sprache. Unter gradbewahrender Randomisierung (Structure Null) sind die realen Graphen systematisch ähnlicher als randomisierte Graphen (ZH-EN: 0,934 < 0,957; DE-EN: 0,938 < 0,957; ZH-DE: 0,519 < 0,717). Dies falsifiziert die Hypothese, dass LDS-K sprachbedingte Divergenz misst.

Warum es relevant ist:

Obwohl mathematische Wahrheit universell ist, sind lehrbuchwissenschaftliche Wissensstrukturen über Sprachen hinweg bemerkenswert ähnlich — ähnlicher, als gradrandomisierte Graphen erwarten ließen. Diese Konvergenz deutet darauf hin, dass die mathematische Voraussetzungslogik — und nicht die Sprache — die institutionelle Wissensorganisation dominiert. Einschränkung: Die Mathematik-Knoten sind teils Alignierungs-Labels (P2-Recheck: de-Feld teils chinesisch, Size-Matching kehrt das Muster um) — die ZH-DE-„Konvergenz" ist damit indikativ, kein unabhängiger Beleg. Der zentrale wissenschaftliche Beitrag verschiebt sich von LDS-K (Lehrbuchdivergenz) zu ΔLDS = LDS-C − LDS-K (kognitive Divergenz jenseits der Wissensstruktur, Konzeptebene; Relationsebene nicht vergleichbar). Die N=15-Humanvalidierung falsifiziert ΔLDS > 0 unter Between-Subject-Bedingungen — und präzisiert damit die Bedingungen (Within-Subject, N⩾30), unter denen ein sprachspezifischer Anteil menschlichen Wissensausdrucks nachweisbar wäre.

Evidenz (Nullmodell-Suite): | Bedingung | ZH-EN | DE-EN | ZH-DE | |:----------|:-----:|:-----:|:-----:| | Voll (LDS-K-Basislinie) | 0,934 | 0,938 | 0,519 | | Structure Null (gradbewahrend) | 0,957 | 0,957 | 0,717 | | Node-Permuted Null | 0,934 | 0,938 | 0,519 | | Complete Random | 1,000 | 1,000 | 1,000 | | Interpretation | Struktur dominiert | Struktur dominiert | Real ÄHNLICHER als Zufall |

Humanvalidierung (N=15, erweitert): Die konzeptuelle LDS-C liegt bei 0.93 – 0.96 — nicht unterscheidbar vom Within-Language-Split-Half-Boden (0.92 – 0.96) und von Label-Permutation (0.94). ΔLDS ≈ 0 (0.05 bis +0.05). Die früheren N=8-Werte (0.70 – 0.75) werden nicht repliziert.

LLM-as-Subject (Within-Subject, §5): Dasselbe LLM (deepseek-v4-flash) antwortet in ZH/DE/EN → LDS-C 0.93 – 0.96 Split-Half-Boden 0.85 – 0.87 → Sprachsignal ist unter Within-Subject-Bedingungen klar nachweisbar. LMM: Sprachcode ist der dominante Organisator (same_lang +0.038, p<0.001), kultureller Rahmen sekundär (same_frame +0.001, p=0.90); ZH-DE trägt eine strukturelle Ebene jenseits der Assoziationsstatistik. Zentrale Lehre: Der Human-Negativebefund ist mit der Heterogenitäts-/Design-Artefakt-Hypothese vereinbar (Exklusion + Konsistenz-Demonstration, q=0.30; mechanismus-konsistente q-Werte kollabieren die Marge nicht vollständig — keine quantitative Kausalzuordnung), kein Beleg für das Fehlen sprachlicher Kognitionseffekte.

Einheitliche Narration (30-Sekunden-Pitch, AI-Audit-Framing)

Mehrsprachige KI-Systeme werden überwiegend mit englischen Daten trainiert — ob sie wertbeladene Konzepte (Gerechtigkeit, Freiheit, Verantwortung, Heimat, Erfolg) sprachübergreifend konsistent verstehen, ist ein blinder Fleck gängiger KI-Evaluation. LinguaGraph macht dieses Modellverhalten prüfbar: Das LLM-as-Subject-Experiment (Within-Subject) weist ein statistisch signifikantes, kulturell gemustertes Sprachsignal nach (LDS-C 0.93 – 0.96 Boden 0.85 – 0.87, Permutationstest $p<0.01$; DE: Autonomie/Regeln, ZH: Raum/Anspruch). Der Domänen-Vergleich (soziale Konzepte divergieren stärker als institutionelle Labels) ist hierfür indikativ, aber wegen der Alignierungsabhängigkeit der Mathematik-Knoten nicht als eigenständiger Beweis zu werten (P2-Recheck; Details: docs/p2_methodology_rechecks.md). Der Output ist ein interpretierbarer Divergenzbericht pro Modell und Sprachpaar — für Entwickler (Vorabprüfung), Regulierer (Transparenz, z. B. im Kontext von Dokumentationspflichten wie dem EU AI Act) und Forscher (kulturelle Werte in KI). Die operative Schwelle (Marge ≥ 0.10) ist heuristisch, keine validierte Grenze — kein fertiges Audit-Instrument.

LinguaGraph macht unsichtbare sprachliche Strukturmuster in KI-Systemen sichtbar, messbar und vergleichbar.

Akademische Version (für Begutachtungsunterlagen)

Mathematik, Physik und Chemie folgen unterschiedlichen Dichtetrajektorien, respektieren eine universelle Tiefengrenze und zeigen — kontraintuitiv — bemerkenswerte strukturelle Konvergenz über Sprachen hinweg auf Lehrbuchebene (mit P2-Recheck-Einschränkung: Mathematik-Knoten teils Alignierungsprodukte). Auf menschlicher Ebene (N=15) ist in einem Between-Subject-Design kein separierbares Sprachsignal nachweisbar. Ein LLM-as-Subject-Experiment (Within-Subject) weist das Sprachsignal dagegen klar nach: Sprachcode dominiert, kultureller Rahmen ist sekundärer Modulator. Damit ist der Human-Negativebefund mit der Design-Artefakt-Hypothese vereinbar (gestützt, nicht quantitativ kausal zugeordnet) — kein Beleg für das Fehlen sprachlicher Kognitionseffekte.

Theorie → Evidenz-Kette

Schlussfolgerung	Theorie	Evidenzquelle	Unterstützt
C1	Ausubel (1963) — Wissensintegration vor Spezialisierung	CDS nach Niveau (Mathe + Physik + Chemie)	Curriculumentwicklung
C2	Novak & Cañas (2008) — Concept Maps als propositionale Netzwerke	HDS-Verteilung (Mathe + Physik)	Lernprogression
C3	Nullmodell (gradbewahrendes Rewiring)	LDS-K vs. Structure Null (Voll < Null)	Falsifikation — LDS-K misst KEINE Sprachdivergenz
ΔLDS-Hypothese	Linguistische Relativität [49][50]	LDS-C vs. LDS-K: Human (N=15, Between) + LLM (Within)	Human: falsifiziert (Between-Subject); LLM Within-Subject: Sprachsignal nachweisbar — Code dominiert, Rahmen sekundär

Anhang: Unterstützungs-Offenlegung

Erklärung der Unterstützung — Declaration of Support

Version: 2.0 | Stand: 2026-09-11 (vor Einreichung final geprüft) Grundlage: BWKI 2026 Teilnahmebedingungen, Klausel „Eigenständigkeit" (Stand 28.04.2023) Zweck: Dieses Dokument legt alle externen Unterstützungsquellen des Projekts transparent und vollständig offen. Es ist mit der schriftlichen Ausarbeitung einzureichen. v1.0 → v2.0 Änderungen: alle KI-Anbieter ergänzt (opencode zen / OpenRouter / DashScope / NIM / Cohere / Kilo), 55-Modell-Replikation, Methodenentscheidung Extraktionsmodell = Versuchsmodell, ein Drittanbieter-API-Abrechnungsvorfall; F1-Zahlen-Zuordnung korrigiert.

0. Zusicherung

Das Projekt (LinguaGraph) wurde in seiner wissenschaftlichen Kernarbeit — Fragestellung, Versuchsdesign, Definition der LDS-Metrik, Datenerhebungsplan, Ergebnisinterpretation und Schlussfolgerungen — von den teilnehmenden Schülern eigenständig erbracht. Alle unten genannten Unterstützungsquellen sind transparent offengelegt; keine Form der Hilfe

wurde verschwiegen.

1. Nutzung von KI-Modellen

Das Projekt nutzt große Sprachmodelle (LLMs) in zwei klar getrennten Rollen:

(1) Als Extraktionswerkzeug (Text → Konzeptgraph): für Konzept-/Relationsextraktion aus Human-Fragebögen und Wikipedia-Korpora.

(2) Als Versuchspersonen (LLM-as-Subject, Within-Subject-Design): Dies ist die wissenschaftliche Kernmethode der Arbeit — dasselbe (bzw. verschiedene) Modelle beantworten 5 Gesellschaftsthemen auf ZH/DE/EN, um die Hypothese „Sprache treibt kognitive Strukturdivergenz" zu prüfen. Dies folgt dem durch Binz & Schulz (2023, PNAS) etablierten LLM-as-Subject-Paradigma. Antworten und Extraktionen der Versuchsmodelle werden im Papier als kontrollierte Experimente berichtet, nicht als verborgene Hilfestellung.

1.1 Verwendete Modelle und Anbieter (alle)

Anbieter/Endpunkt	Modelle	Verwendung
opencode GO (https://opencode.ai/zen/go/v1)	deepseek-v4-flash	Baseline-Versuchsmodell (D1-Hauptexperiment) • Konzept-/Relationsextraktion • Glossar-Anglisierung
opencode zen/v1 (https://opencode.ai/zen/v1)	u. a. deepseek-v4-flash, nemotron-3-ultra-free (NVIDIA), mimo-v2.5-free, laguna-s-2.1-free, longcat-2.0-free	Replikations-Versuchsmodelle
OpenRouter (https://openrouter.ai/api/v1, kostenlose Stufe)	u. a. gpt-oss-20b:free (teilweise laufend), nemotron-/laguna-/gemma-Modelle	Replikations-Versuchsmodelle (teilweise unvollständig)
Alibaba Cloud DashScope/Qwen (https://dashscope.aliyuncs.com/compatible-mode/v1)	deepseek-v3/v3.1/v3.2/v4/r1-Familie, GLM-4.5 – 5.2, Kimi, MiniMax, Qwen3.x u. a., 42 Modelle	Replikations-Versuchsmodelle
NVIDIA NIM	gpt-oss-20b	West-Erweiterung (vollständig)
Cohere	command-a-03-2025	West-Erweiterung (vollständig)
Kilo	laguna-s-2.1, nemotron-3-super-120b-a12b	West-Erweiterung (vollständig)
opencode-Terminal	gpt-5.6-luna (Herkunft ungeklärt)	West-Erweiterung (vollständig)

Mehrmodell-Replikation (55 vollständige Messungen / 50 eindeutige Modelle + 31 laufende): 2026-08-09/10 wurden 42 DashScope-Modelle + 7 zen/OpenRouter-Modelle + D1-Baseline unter identischem P1-Protokoll (3 Sprachen × k=10) als LLM-as-Subject vermessen. West-Erweiterung 2026-09-09/10 (gleiches Protokoll, alle vollständig): gpt-oss-20b (NVIDIA NIM), command-a-03-2025 (Cohere), laguna-s-2.1 + nemotron-3-super (Kilo), gpt-5.6-luna (opencode-Terminal) → 55 vollständige Messungen / 50 eindeutige Modelle, alle ZH-DE-Paare signifikant. Alle Modelle wurden als kontrollierte Versuchsmodelle vermessen; ihre Antworten und Extraktionen bilden den Datenkörper des Papiers (data/lds_c/llm_subject/). Dies sind Forschungsdaten, keine „Hilfe bei der Anfertigung".

1.2 Extraktionsqualität und F1-Zahlen (v1-Verwechslung korrigiert)

- Human-validierte F1 (Extraktion sozialer Konzepte): ZH 0,974 / DE 0,949 / EN 0,882, sozial gesamt 0,939 (72 soziale Gold-Annotationen), gewichtet über Domänen 0,881 (n=92) — dies validiert, ob der Extraktor humane Konzeptannotationen wiederherstellt.
- 19-Modell-Extraktionsbenchmark: F1-Spanne 0,55 – 0,67 über Modellfamilien (data/model_comparison/) — zur Modellauswahl, kein berichteter Projektwert.
- v1-Fehler: v1 vermischte beides zu einem einzigen „0,939 (19-Modell-Benchmark)" — korrigiert. Papier §8.7/§8.9 entsprechen den beiden obigen Zahlengruppen.

1.3 offengelegte Methodenentscheidung: Extraktionsmodell = Versuchsmodell

In der Mehrmodell-Replikation extrahiert jedes Versuchsmodell die Konzepte aus seinen eigenen Antworten selbst (lds_c_llm_subject.py, zwei Stufen: Antworten → Extraktion durch dasselbe Modell). Dies hält die „Konzeptstruktur desselben Modells" konsistent, bedeutet aber auch, dass die gemessene sprachübergreifende Divergenz Extraktionverhaltensunterschiede je Sprache enthält. Papier §5 legt diese Konfundierung offen; Extraktionsdichte und Divergenzmarge sind unkorreliert (corr 0,23), und die Domänenkontrolle (institutionelle Konvergenz vs. kulturelle Divergenz) mildert sie teilweise.

1.4 Leerergebnisse und Wiederholungen

- Leerergebnisse werden explizit wiederholt (max. 3 Versuche je Nachricht); nach 3 aufeinanderfolgenden Fehlereinheiten wird das Modell automatisch verworfen (LDS_ABORT_AFTER)
- Baseline-D1-Experiment: 220/220 Einheiten vollständig; Mehrmodell-Sammlung mit Checkpoint-Fortsetzung je Einheit

2. Nutzung KI-basierter Hilfsmittel

Werkzeug	Verwendung	Umfang
Claude Code (Anthropic)	unterstütztes Programmieren, Schreiben von Datenanalyseskripten, Dokumentation, Code-Review, dieser Review-/Reparaturprozess	Kerncode/Analyseskripte wurden mit Claude Code generiert und debuggt; Forschungsdesign und Schlussfolgerungen wurden von den teilnehmenden Schülern geleitet

Hinweis: Claude Code wirkte als KI-Programmierassistent bei Skriptentwicklung, Datenanalyse und Dokumentenlayout mit. Wissenschaftliche Urteile, Versuchsdesign-Entscheidungen und Ergebnisinterpretation wurden von den teilnehmenden Schülern eigenständig erbracht und werden gemäß Wettbewerbsregeln transparent offengelegt.

3. Datensatzquellen

Datensatz	Quelle	Lizenz	Verwendung
Human-Fragebogenantworten (N=15)	eigene Erhebung (Online-Fragebogen, 2026-07)	eigene Daten der Teilnehmenden; Einwilligungserklärungen (docs/ethics/consent_{de,en,zh}.md), DSGVO-konform	Analyse kognitiver Äußerungen (Between-Subject-Design)
Wikipedia-Artikel (ZH/EN/DE, 5 Gesellschaftsthemen)	Wikipedia	CC-BY-SA (Namensnennung erforderlich)	soziokulturelle Wissenskontrolle (Negativkontrolle/Vertiefung)
Mathematik-Lehrbuch-Wissensstruktur	CN/DE/EN-Lehrpläne und -Lehrbücher	Konzeptgraphen sind Strukturbeschreibungen, keine Textkopien	institutionelle Wissenskontrolle (LDS-K)

Wikipedia-Namensnennung (CC-BY-SA-Konformität)

Das Projekt extrahierte dreisprachige Konzeptstrukturen zu 5 Gesellschaftsthemen (Freiheit/Gerechtigkeit/Verantwortung/Heimat/Erfolg). Gemäß CC-BY-SA werden die Quellartikel hier genannt: - Jede Extraktionsdatei data/wikipedia_extractions/{topic}_{lang}.json enthält ein source_url-Feld mit dem konkreten Artikellink. - Bei Einreichung: Das Literaturverzeichnis des Papiers listet je Datei die Wikipedia-Artikel mit Extraktionsdatum (2026-08-08) auf.

4. Personen und Institutionen

Beteiligte	Rolle	Anmerkung
Teilnehmende Schüler	Projektleitung, Forschung	Forschungsdesign, Datenerhebung und Ergebnisinterpretation eigenständig erbracht
Schule (deutsches Gymnasium)	stellt Rahmen	keine inhaltlich-wissenschaftliche Einwirkung
Weitere Teams/Institutionen	keine	keine externe Forschungskooperation

5. Rechenleistung, Infrastruktur und Drittanbieter-API-Abrechnung

Posten	Angabe
Rechenressourcen	lokale CPU (Python-Analyseskripte) + Drittanbieter-LLM-APIs
GPU	keine externe GPU-Leistung genutzt
Drittanbieter-API-Abrechnung	opencode GO (nutzungsbasiert), DashScope (Freibetrag + ein Sperrvorfall), OpenRouter (kostenlose Stufe), opencode zen/v1 (kostenlose Stufe)

5.1 Drittanbieter-API-Abrechnungsvorfall (ehrliche Offenlegung)

Am 2026-08-10 führte bei einer DashScope-Batch-Sammlung die versehentliche Nutzung eines nicht kostenfreien Modells zu einer kurzen Kontosperrung wegen Zahlungsrückstands. Das Problem wurde am selben Tag erkannt: Nach Umstellung auf ausschließlich Freibetrags-Modelle wurde die Sammlung fortgesetzt und nach Begleichung wieder aufgenommen. Der Vorfall ist im internen Übergabedokument (docs/session_handoff_20260810.md §3.3) festgehalten. Er beeinträchtigt die Validität der Forschungsdaten nicht (erhobene Daten wurden verifiziert), wird aber der Vollständigkeit halber offengelegt.

6. Urheberrecht und Musik (Videomaterial)

- Das Einreichungsvideo verwendet ausschließlich selbst erstellte Bilder und systemgenerierte Inhalte.
- Falls im Video urheberrechtlich geschütztes Material verwendet wird, werden in der Videobeschreibung Autor/Rechteinhaber und Link genannt (gemäß Regelwerk).
- LDS-Definition, Fragebogen und Versuchsdesign der Arbeit sind originär (einschlägige Literatur im Papier zitiert).

7. Selbstprüfung und Compliance-Bestätigung

Regelanforderung	Status
Offenlegung der KI-Modellnutzung	diese Datei §1 (alle Anbieter + 55-Modell-Replikation + Extraktions-=Versuchsmodell-Entscheidung)
Offenlegung KI-basierter Hilfsmittel	diese Datei §2
Offenlegung der Datensatzquellen	diese Datei §3
Offenlegung Personen/Institutionen	diese Datei §4
Offenlegung Rechenleistung/Infrastruktur	diese Datei §5 (inkl. API-Abrechnungsvorfall)
Namensnennung von Video-Copyright-Material	bei Aufnahme auszuführen
EU-AI-Act-Konformität	LLM-Nutzung für Forschungsanalyse, keine beschränkten Zwecke (keine Deepfakes u. Ä.)
Keine diskriminierenden/antidemokratischen/militärischen Inhalte	nicht zutreffend

Dieses Dokument wird mit der schriftlichen Ausarbeitung eingereicht. Übersehene Hilfsquellen werden vor Einreichung nachgetragen.

8. Unterschrift (bei Einreichung ausfüllen)

Feld	Angabe
Ort, Datum	____, 2026-09-__
Name (Druckschrift)	____
Unterschrift	____

Erklärung: Die obige Offenlegung ist vollständig und wahrheitsgemäß; alle KI-/Werkzeug-/Datenquellen der Arbeit sind in §§1–5 genannt.