UNTERSTÜTZUNG VON DEZENTRALISIERUNGS-MAßNAHMEN IN SUBSAHARA-AFRIKA DURCH DIE DEUTSCHE ENTWICKLUNGSZUSAMMENARBEIT

Methoden-Anhang

2025

Helge Roxin
Dr. Christoph Dworschak
Haik B. Gregorian

DEval
DEUTSCHES EVALUIERUNGSINSTITUT DER ENTWICKLUNGS-ZUSAMMENARBEIT

INHALT

Abbildungen

Tabellen

1. PORTFOLIOREKONSTRUKTION UND -ANALYSE

Zu Beginn der Evaluierung lag kein klar abgegrenztes Dezentralisierungsportfolio der deutschen Entwicklungszusammenarbeit vor. Daher wurde das Portfolio im Rahmen der vorliegenden Evaluierung wie folgt rekonstruiert:

Als erster Bezugspunkt diente der Förderbereichsschlüssel (FBS) 15112 für „Dezentralisierung und subnationale Regierungsführung", auf dessen Basis relevante Vorhaben im Datenportal MeMFIS identifiziert wurden.[1] Auf diese Weise konnten 120 Vorhaben für die Förderregion Afrika und 241 Vorhaben weltweit identifiziert werden. Da der FBS 15112 jedoch nur einen jüngeren Zeitraum, ab circa 2010, abdeckt, sind viele dezentralisierungsrelevante Vorhaben nicht in dem FBS enthalten. Daher wurde der Suchkreis durch eine inhaltsbasierte Stichwortsuche erweitert. Hierfür wurden die in MeMFIS hinterlegten Textfelder „Bezeichnung EZ-Maßnahme (deutsch)" und „Zielsetzung" mithilfe eines Rasters relevanter Begriffe durchsucht.[2] Wie im Bericht dargelegt, konnten durch diesen Ansatz 297 Vorhaben für die Förderregion Afrika und 514 Vorhaben weltweit identifiziert werden.

Im Rahmen der Datenerhebung und -analyse stellten sich manche Vorhaben als relevant für den Evaluierungsgegenstand heraus, die jedoch nicht im rekonstruierten Portfolio enthalten waren. Diese wurden dem rekonstruierten Portfolio aufgrund ihres nachweislichen Dezentralisierungsbezugs nachträglich hinzugefügt. Diese manuellen Änderungen des rekonstruierten Portfolios sind im Folgenden einzeln nach Partnerland aufgeführt.

Sambia

- 2011.6656.0 – „Stärkung der parlamentarischen Kontrolle in Sambia I“
- 2011.2112.8, 2014.2077.7, 2016.2216.6 – „Politische Teilhabe von Zivilgesellschaft in Governance-Reformen und Armutsbekämpfung“
- 2015.6868.2 – „Erweiterung Chalimabana“
- 2020.2097.2 – „Förderung von Transparenz, Partizipation und Zugang zu Recht“

Ghana

- 199465493, 199570060, 1998.6689.8, 1999.6535.1, 2001.6605.8 – „Distriktstädte Serie“
- 2006.2108.6, 2009.2048.8, 2015.2087.3 – „Verbesserung öffentlicher Finanzen“
- 2012.6636.0, 2012.7023.0 – „Ergebnisorientierte Verbesserung der Finanzverwaltung“

Senegal

- 199866716 – „Kommunalentwicklung/Dezentralisierung (Kaolack und Fatick)“
- 200565879 – „Unterstützung Kommunalentwicklung in den Regionen Kaolack und Fatick II“
- 2002.2506, 2004.6549.2, 2007.2013.6, 2010.2201.1, 2013.6732.5 – „Friedensförderung in der Casamance“

[1] Da die Abdeckung des BMZ-Portfolios in MeMFIS am umfangreichsten war, wurde es gegenüber den Datenportalen der OECD-DAC und IATI vorgezogen.

[2] Das Raster an Suchbegriffen wurde in der Inception-Phase kontinuierlich auf Basis von Vorhaben- und Strategiedokumenten sowie Staheholder-Konsultationen und Begriffen aus der ToC angereichert. Das Raster beinhaltete beispielsweise die folgenden Begriffe: ‚Dezentralisierung', ‚subnational', ‚Kommunen', ‚Distriktverwaltung' bzw. ‚-regierung', ‚Selbstverwaltung', ‚Mehrebenensystem', ‚Subsidiarität', ‚Regierungsführung', *‚local governance'* und ‚lokale Regierungsführung'. Die angesprochene Erweiterung umfasste die Begriffe ‚Dekonzentration', ‚Gemeindeverwaltung', ‚Gemeindeentwicklung', ‚Gebietskörperschaften', ‚Raumplanung', ‚Eigeneinnahmen', ‚Provinz', ‚Munizipien', ‚Vertretung', ‚Verband', ‚Verbund', ‚Partizipation', ‚marginal', ‚Daseinsvorsorge', ‚Versorgung'.

Mosambik

- 2002.2125.9 – „Dezentralisierung und Kommunalentwicklung“
- 2003.2091.1 – „PRODDER“

Kamerun

- 2017.2020.0 – „Modernisierung des Personenstandswesens“

Burkina Faso

- 2009.6693.7 – „Kommunalentwicklungsfonds“
- 2019.6946.8, 2003.2131.5 – „Kommunalentwicklungsfonds (FDC)“

Malawi

- 2017.2025.9 – „Stärkung des öffentlichen Finanzmanagements“
- 199665209, 199770298 – „Sekundärstädte“
- 2001.7057.1 – „Ausbau von Sekundärzentren Phase VI (Begleitmaßnahme)“

Für die Anzahl der Nennungen relevanter Vorhaben in den Tabellen 4, 5, 6, 7 und 8 im Evaluierungsbericht wurden MeMFIS-Einträge nach ihrer BMZ-Nummer gebündelt – Vorhaben, deren Zusage-Betrag mit 0€ hinterlegt war, wurden herausgefiltert. Für die Berechnung der Volumina in den Abbildungen 3 und 4 sowie den Tabellen 9 und 10 im Evaluierungsbericht wurden die inflationsbereinigten Beträge von „Zusage-Betrag inkl. Re-Programmierungen" herangezogen.

Der Wirkungsstrang „Förderung der lokalen Wirtschaft" ließ sich im Portfolio nur schwierig abbilden. Zum einen erschwerte die fehlende Trennschärfe zwischen Förderung auf nationalstaatlicher und lokaler Ebene die Identifikation dezentralisierungsrelevanter Vorhaben in den MeMFIS-Daten. Da Wirtschaftsförderung einen großen Teil der deutschen EZ darstellt, würde eine solche undifferenzierte Einbeziehung aller wirtschaftsbezogenen Vorhaben überdies eine Portfolioanalyse des Förderbereichs „Dezentralisierung“ unmöglich machen. Zum anderen konnten, anders als für die anderen Wirkungsstränge, keine weiteren Förderbereichsschlüssel oder übersektorale Kennungen genutzt werden, um die „Förderung der lokalen Wirtschaft" abzubilden. Jedoch konnten mehrere relevante Vorhaben des Wirkungsstrangs aus den letzten 10-15 Jahren über Interviews, Dokumentenanalysen und die Evaluierungssynthese identifiziert werden.

2. SYNTHESE WISSENSCHAFTLICHER LITERATUR ZUM THEMENBEREICH DEZENTRALISIERUNG

Das Vorgehen zur Synthese wissenschaftlicher Literatur folgte drei Schritten:

1. Die Erstellung eines Datenauszugs aus dem *Web of Science*,
2. eine Large-Language-Model-basierte Vorauswahl, die Publikationen nach ihrer Relevanz filtert,
3. das händische Kodieren einer zufälligen Stichprobe von Publikationen. Diese Schritte werden im Folgenden näher beschrieben.

2.1 Datenauszug Web of Science

Der erste Schritt zur Synthese wissenschaftlicher Literatur bestand darin, einen Datenauszug aus der *Web of Science* (WoS) *Core Collection* zu erstellen. Der WoS-Auszug wurde mit den folgenden Suchkriterien erstellt:

> *TS=((develop* OR "administrative capacity" OR "democra*" OR ("servic*" AND ("quality" OR "provision")) OR subnational OR sub-national OR "local govern*" OR "accountab*" OR "corruption" OR "investment" OR "political trust" OR "conflict") AND (decentralization OR "regional government splits" OR "administrative splits" OR "local administrati* reform" OR "devolution")) AND WC=(Political Science OR Economics OR Development Studies) NOT PY=(2025)*

Dabei steht „TS“ für das Durchsuchen von Publikationstiteln, Abstracts und Keywords. Im Fokus stehen hier übergreifende Konzepte, die mit Dezentralisierung und relevanten Wirkfeldern der Wirkungslogik[3] übereinstimmen. „WC” steht für WoS-Themenfeldkategorie. Einzelne Einträge können mehr als einer Kategorie zugeordnet sein, wodurch die hier ausgewählten Themenfelder eine ausreichend breite Masse an Publikationen abdecken. In der untenstehenden Grafik sind alle enthaltenen WoS-Themenkategorien mit mehr als 100 Publikationen abgebildet. Dabei enthält die manuell erstellte „Anderen“-Kategorie über 40 Residualkategorien mit weniger als 100 Publikationen. Einschlägige Beispiele sind *Business* (66), *Forestry* (64), *Management* (47), *Transportation* (44), and *Environmental Sciences* (41). Weniger prominente Beispiele sind *History* (17), *Law* (15), *Education* (4), *Pharmacology & Pharmacy* (2) und *Nutrition and Dietetics* (1). Der Schlussteil „NOT PY“ bedeutet, dass alle Publikationsjahre außer 2025 berücksichtigt werden sollen. Der Ausschluss des laufenden Jahres sollte verhindern, dass Neupublikationen zu ständigen Veränderungen des WoS Auszugs und einer späteren fehlenden Reproduzierbarkeit führen.[4]

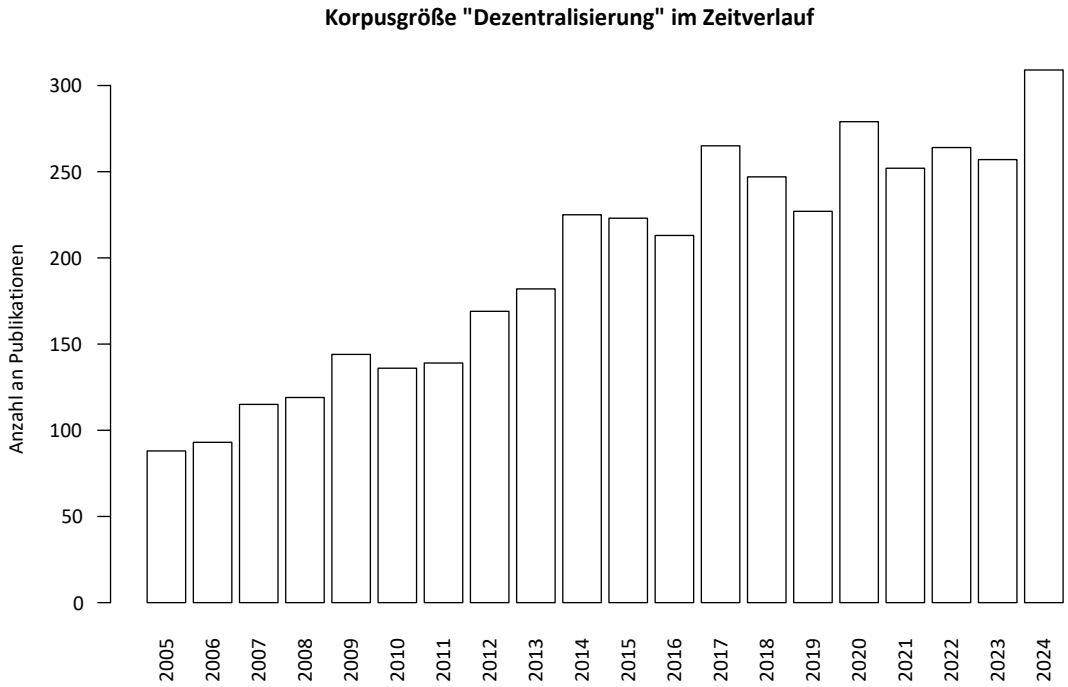
Da im weiteren Verlauf eine Large-Language-Model-basierte Filterstufe eingeplant war, wurde der WoS-Auszug möglichst umfassend angelegt: der Auszug sollte die Zahl der „*False Negatives*“ (relevante Publikationen, die jedoch nicht im Datenauszug enthalten sind) möglichst geringhalten, auch wenn das zu einer höheren Anzahl von „*False Positives*“ (irrelevante Publikationen, die jedoch im Datenauszug enthalten sind) führt. So ergaben die Suchkriterien einen Auszug von 3954 Publikationen mit Informationen zu Publikationstitel, Keywords,

[3] Ein Einbeziehen der Wirkungsfelder war nötig, um beispielsweise Publikationen zu Block-Chain Technologien herauszufiltern. Ein vereinfachter „TS“ Filter mit nur „((decentralization OR "regional government splits" OR "administrative splits" OR "local admi-nistrati* reform" OR "devolution"))” wurde dagegengehalten, um sicherzustellen, dass die sich nicht überschneidenden Teilmengen keinen Verlust an relevanten Publikationen andeuten.

[4] Trotz dieser Maßnahme kam es zu Änderungen im WoS-Datenauszug im Laufe dieser Arbeitsphase. So kam es mehrfach vor, dass die gleichen Suchkriterien bei nochmaliger Suche zu leicht abweichenden Ergebnissen führten. In Zusammenarbeit mit dem WoS Core Collection Team konnte bestätigt werden, dass die WoS Core Collection ständigen Änderungen aufgrund der im Hintergrund ablaufenden Kuration unterliegt, was deren Reproduzierbarkeit stark einschränkt. Die hier dargestellten Ergebnisse basieren auf einem Auszug, der im März 2025 erstellt wurde.

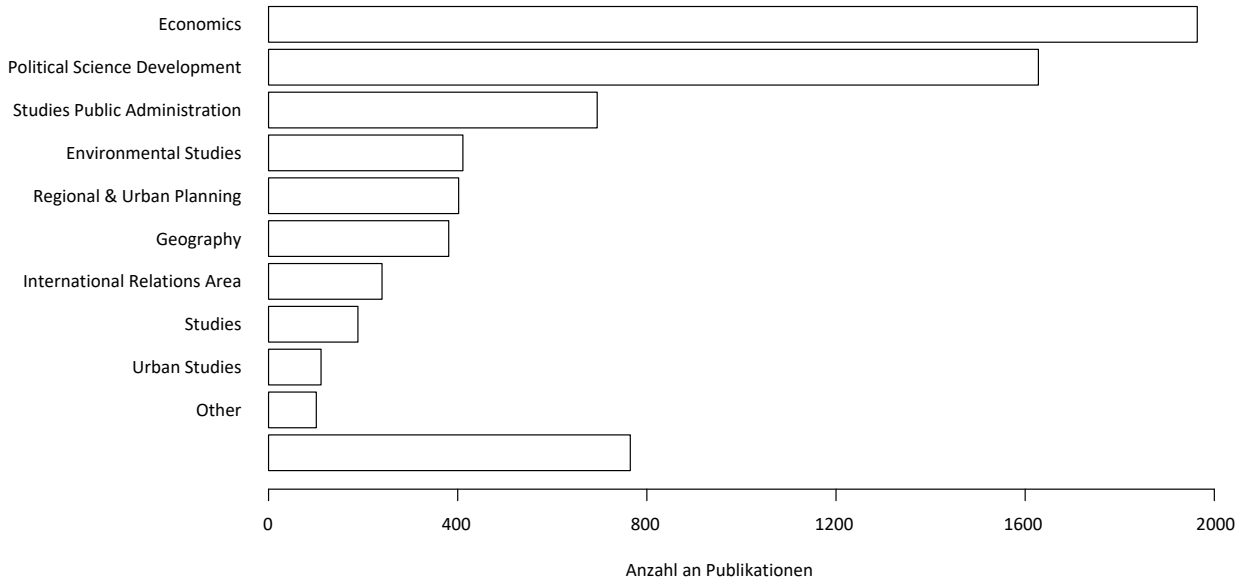
Abstrakt und verschiedenen Metainformationen.[5] Ein anschließender Ausschluss aller Publikationen, die keinerlei international anerkannte Identifikationsnummer aufweisen (zum Beispiel DOI oder ISSN), führte zu einem Korpus von 3946 Publikationen zur Weiterverarbeitung.

Abbildung 1 WoS-Korpus Übersicht



Quelle: DEval, eigene Darstellung

Abbildung 2 Themenkategorien im Korpus „Dezentralisierung“



Quelle: DEval, eigene Darstellung

[5] Ein systematischer Zugriff und Download der Gesamttexte der Publikationen ist in WoS nicht möglich.

2.2 Large-Language-Model-basierte Vorauswahl

Im weiteren Verlauf wurde der oben beschriebene WoS-Auszug mit einem Large-Language-Model (LLM) ausgelesen, um Filterkriterien zu generieren und anzuwenden. Das Ziel dieses Schrittes war es, das darauffolgende händische Kodieren der Studien zu vereinfachen, indem eine Vorauswahl basierend auf den LLM-Ergebnissen getroffen wird, durch die irrelevante Publikationen herausgefiltert werden. Diese Vorauswahl basierte auf Mindestkriterien hinsichtlich der Relevanz und Qualität der Studien.

Um eine möglichst gute („präzise“ und „sensitive“) Vorauswahl zu treffen war das Tuning des Modells von besonderer Bedeutung. Das Tuning bestand aus Performanz-Vergleichen zwischen verschiedenen Modellen, verschiedenen Ausgabewerten und „Prompt Engineering“, welche in Zusammenarbeit mit einem externen IT-Dienstleister in einem iterativen Prozess umgesetzt wurden. Das leistungsstärkste Modell, mit dem die finale Textanalyse durchgeführt wurde, war Meta-Llama-3.3-70B-Instruct-AWQ-INT4. Diesem LLM wurde folgende Aufgabenbeschreibung (Auszug) gegeben:

> *You are a research assistant analyzing scientific studies. Your task is to extract specific information about studies from the provided abstract and related metadata. Answer each question concisely and precisely, following the specified output format. Provide a one-sentence reasoning for each answer. For scoring questions, provide BOTH a matching score (pi) between 0 and 1, AND an uncertainty score (sigma) between 0 and 1. […] Do not make assumptions. Base your answers *only* on the provided abstract and associated metadata. […]*

Darauf folgten eine Vielzahl an Kodierfragen, die sich in drei Kategorien unterteilen lassen: 1) Relevanz des Themas und Designs, 2) wissenschaftliche Güte und 3) Studienergebnisse. Das LLM wurde angewiesen, als Antwort auf jede Frage einen Schätzwert, einen Unsicherheitswert und eine Begründung zu generieren. Der Unsicherheitswert[6] und die Begründung wurden im Kontext des ‚Tuning-Prozesses‘ zum qualitativen Abgleich genutzt, um Schwierigkeiten und Verbesserungspotenziale des LLMs zu identifizieren.

Für die Vorauswahl wurde eine Vielzahl verschiedener Schätzwerte des LLMs als Filterkriterien herangezogen. Die beiden wichtigsten Kriterien waren dabei 1) ein Schätzwert zur allgemeinen Themenrelevanz („Relevance Score“) und 2) ein Schätzwert zur allgemeinen wissenschaftlichen Güte („Science Score“). Die Anweisung an das LLM zur Generierung dieser beiden Schätzwerte lautete:

> ***Relevance Score**: Based on the abstract and the associated metadata, how central is the concept of decentralization to the study? Note: “decentralization” may also be called, for example, deconcentration, delegation, or devolution, and refers to a process of moving a state’s political, fiscal, and/or administrative responsibilities from the national to a subnational level. More precisely: [Detailed definitions of political, fiscal, and administrative decentralization.]*

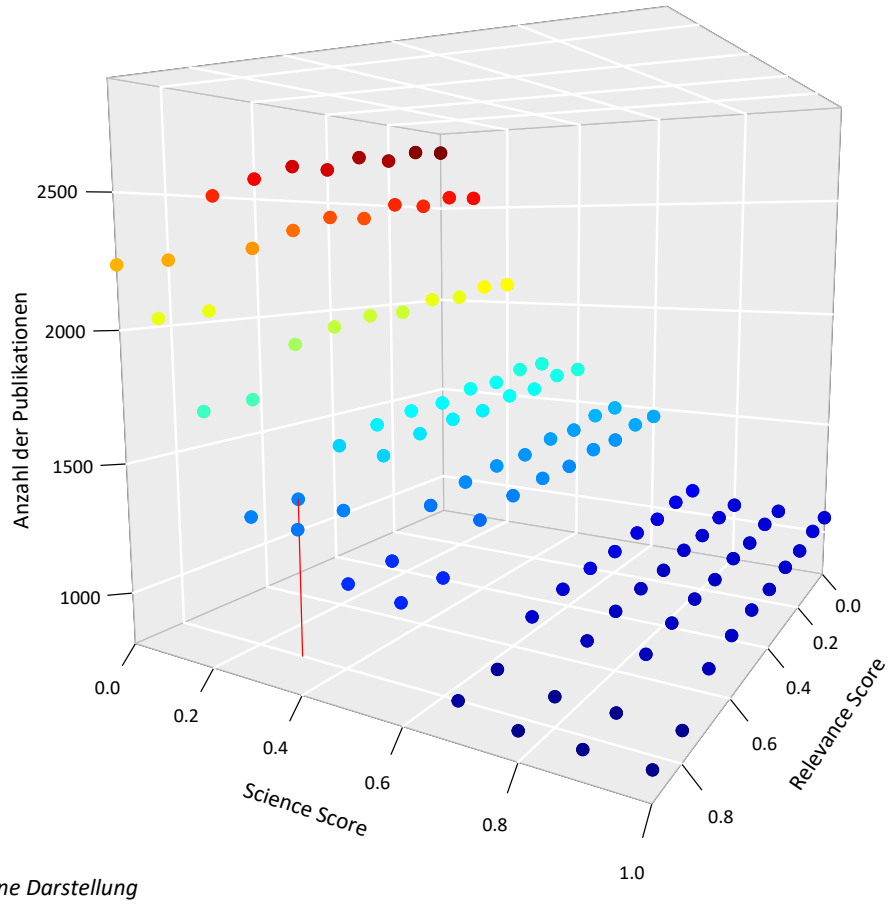
> ***Science Score**: Thinking about the methodology that the study uses to analyze the data identified in the previous prompt, assess the level of scientific rigor of the research design in approximating experimental or observational causal inference. Consider whether the research design allows for strong causal claims. Note: This does not need to be explicit; even in the absence of an explicit commitment towards approximating causal inference, the mode of intra- or inter-unit comparison may allow for a certain level of approximation towards causal claims. Provide matching scores for each level (Low, Medium, High) such that they sum to 1. [Uncertainty instructions.]*

Die beiden Scores wurden auf einer Skala von 0 bis 1 bewertet. Diese kontinuierlichen Schätzwerte erforderten die Bestimmung geeigneter Schwellenwerte, die den Anteil an relevanten Publikationen im verbleibenden Korpus maximieren und den Anteil irrelevanter Publikationen minimieren. Dabei sollten die Schwellenwerte so konservativ wie möglich ausfallen. Abbildung 3 zeigt, wie sich unterschiedliche Schwellenwerte auf die finale Korpusgröße auswirken. Bei den jeweils niedrigsten Schwellenwerten der beiden Scores (0.0 für

[6] Da sich der Unsicherheitswert im Laufe des „Tuning-Prozesses“ als wenig reliabel herausgestellt hat, wurden zusätzlich mehrere Durchläufe mit dem gleichen Modell durchgeführt, um so die Modellreliabilität besser einschätzen zu können. Dabei wurde eine Temperatur („Randomisierungsfaktor“ der die Streuung des Modells reguliert) von 0.3 angewandt.

Relevance und 0.0 für *Science*) enthält der Korpus 2799 Publikationen[7] zum weiteren händischen Kodieren. Die jeweils höchsten Schwellenwerte (0.9 für *Relevance* und 1.0 für *Science*) ergeben einen Korpus von 898 Publikationen. Dabei ist in der drei-dimensionalen Abbildung gut zu erkennen, wie sich die beiden *Scores* gegenseitig moderieren und der Science Score deutlich stärker differenziert als der *Relevance Score*.[8]

Abbildung 3 Korpusgröße entlang verschiedener Schwellenwerte



Quelle: DEval, eigene Darstellung

Die finalen Schwellenwerte wurden auf 0.8 für den *Relevance-Score* und 0.3 für den *Science-Score* festgesetzt. Der daraus resultierende Korpus enthält 1395 Publikationen (rote Linie in Abbildung 3). Dabei wurden die Schwellenwerte so niedrig wie möglich gesetzt, um den Anteil an fälschlich ausgeschlossenen Publikationen zu minimieren. Hier zeigten qualitative Abgleiche von Stichproben entlang der jeweiligen Scores, dass ein weiteres Absenken der gewählten Schwellenwerte zu einem unverhältnismäßig starken Abfall in der relativen Treffergenauigkeit führen würde. Mit 1395 verbliebenen Einträgen konnte so der ursprüngliche WoS-Auszug von 3946 Publikationen mithilfe der LLM-basierten Vorauswahl erheblich reduziert werden. Somit erlaubte diese Vorauswahl einen deutlich gezielteren Einsatz von Ressourcen im anschließenden händischen Kodierverfahren.

[7] Die verbleibende Differenz zum ursprünglichen WoS-Auszug von 3946 Publikationen entsteht durch verschiedene Mindestanforderungen im Hinblick auf andere Schätzwerte. Beispielweise wurde bei einem Großteil der Differenzeinträge Dezentralisierung nicht als Einflussvariable, sondern als Ergebnisvariable eingeordnet. Zudem entfielen 85 Publikationen, die nicht vom LLM evaluiert werden konnten (z.B. aufgrund eines fehlenden *Abstracts*).

[8] Dieser Unterschied im Grad der Differenzierung entspricht den Erwartungen, da die WoS-Filterkriterien bereits zu einer Eingrenzung nach Themenrelevanz geführt hatten, jedoch keinerlei Filter nach wissenschaftlicher Qualität erlaubten.

2.3 Händisches Kodieren und Skalieren

Im weiteren Verlauf wurde der gefilterte Korpus von 1395 Publikationen für das händische Kodieren vorbereitet, kodiert und schließlich skaliert. Aufgrund begrenzter Ressourcen wurde die Reihenfolge der Publikationen, in der sie kodiert würden, randomisiert. Dies sollte sicherstellen, dass, unabhängig vom finalen Umfang der Kodierung, eine repräsentative Zufallsstichprobe vorliegen würde. Das genaue Kodierverfahren und die zu kodierenden Variablen sind in einem detaillierten Kodierleitfaden festgehalten, welcher als Vorlage für die beteiligten studierenden Beschäftigen gedient hat.

Das händische Kodierverfahren wurde von zwei studierenden Beschäftigen durchgeführt und vom Evaluierungsteam begleitet. So wurden in regelmäßigen Austauschrunden schwierig einzuordnende Publikationen besprochen und der allgemeine Kodierleitfaden weiter angereichert. Der fachliche Austausch ergab zudem, dass, wie oben angemerkt, die Outcome-Kategorien „Soziale und wirtschaftliche Entwicklung“ (D01) und „Armutsminderung“ (D03), sowie „Bessere und inklusivere Dienstleistungen für Bürger“ (C02) und „Dienstleistungserbringung zur Daseinsvorsorge/Bereitstellung öffentl. Güter“ (B02), in der wissenschaftlichen Literatur kaum voneinander getrennt werden können und deshalb in der Ergebnispräsentation in gemeinsame Kategorien zusammenzufassen sind.

Eine weitere Komponente des händischen Kodierverfahrens waren zwei „inter-coder reliability (ICR)“-Tests: einer zum Abgleich der Kodier-Reliabilität zwischen den beiden studierenden Beschäftigen und einer zur Überprüfung der Reliabilität der LLM-Vorauswahl. Für den ersten ICR-Test wurden 21 Einträge von jeweils beiden Beschäftigen kodiert. Für den zweiten ICR-Test haben die Beschäftigen 20 Publikationen kodiert, die in der LLM Vorauswahl herausgefiltert wurden und die nicht in die Synthese-Ergebnisse einfließen. Beide ICR-Tests wurden blind durchgeführt: die zusätzlichen Publikationen wurden zufällig in die jeweilig zu kodierenden Stichproben gemischt, ohne dass die Beschäftigen wussten, um welche Einträge es sich dabei handelt. Die Ergebnisse zeigten ein hinreichendes Maß an Übereinstimmung und wurden zur weiteren Steigerung der Zuverlässigkeit des Kodierens genutzt. Alle Details zu individuellen Kodier-Entscheidungen sind in Kodierprotokollen festgehalten.

Um eine repräsentative Ergebnisdarstellung zu erzielen, wurde die zufällige Stichprobe, die die studierenden Beschäftigen kodiert haben, auf die ursprüngliche Population von 1395 relevanten Publikationen hochskaliert. Hierfür wurde die kodierte Stichprobe von 1099 Publikationen (79 %) durch Ziehung mit Zurücklegen („bootstrapping“) 5000-mal neu generiert und die jeweiligen Anteile an zugeordneten Einfluss- und Ergebnisvariablen für jede der 5000 Stichproben neu berechnet.[9] Der Durchschnittswert, sowie die 2,5 %- und 97,5 %-Quantile, der daraus entstandenen Verteilung an Zuordnungen, wurde sodann auf die Gesamtzahl von 1395 Publikationen angewandt. Die Balkendiagramme im Berichtskapitel „Wissensstand“ zeigen diese auf den Durchschnittswerten basierende Hochrechnung. Die tabellarische Ansicht in Tabelle 1 (s.u.) zeigt neben diesen Schätzwerten auch die jeweiligen Konfidenzintervalle.

Die weitere grafische Darstellung der Verteilung von Effektausprägungen ist im Evaluierungsbericht dokumentiert. Diese folgt einem einfachen „*vote counting*“-Ansatz, in dem die aggregierten Ergebnisse entsprechend der ungewichteten Anzahl an Studien dargestellt werden. Da dieser Ansatz weder die geschätzten Unsicherheiten der Effektausprägungen, die Rigorosität der Studien, noch deren zugrundeliegenden Stichprobengrößen berücksichtigt, erfordern die Ergebnisse eine differenzierte Interpretation: So können Zusammenhänge, für die unterschiedliche Effektausprägungen jeweils zu einer gewissen Menge vorliegen, lediglich als „gemischt“ gedeutet werden. In diesem Fall sind Aussagen über relative Mehrheiten womöglich irreführend. Nur eindeutige Studienlagen – bei denen die große Mehrheit an Studienergebnissen in eine gemeinsame Richtung weist – können als Indiz für eine mögliche Effektrichtung gelesen werden.

[9] Eine alternative Herangehensweise bestand in der Nutzung der händisch kodierten Stichprobe als Trainingsset zur Klassifizierung des restlichen Korpus auf Basis der LLM-basierten Themenzuordnung, die für jede Observation geschätzt wurde. Die entsprechenden Klassifikationsschwellenwerte (F1 Maximum) waren dabei allerdings nicht so ergiebig wie erhofft (siehe Abbildung 4), was auf eine geringe Trennschärfe der LLM-Schätzwerte innerhalb des vorgefilterten Publikationskorpus hinweist. Somit versprach das durch die Zufallsauswahl der Stichprobe ermöglichte ‚Bootstrapping-Verfahren‘ eine deutlich höhere Genauigkeit.

Tabelle 1 Schätzwerte und deren Unsicherheit

	Impacts		Highly aggregated outcomes						Outcomes (selection)
	D01/D03	D02	C01	C02/B02	C03	C04	C05	C06	B06
Political positives	25 [10; 45]	54 [30; 84]	0 [0; 0]	60 [35; 89]	5 [1; 15]	30 [10; 50]	15 [5; 30]	10 [2; 25]	75 [45; 109]
Political negatives	20 [5; 40]	25 [10; 45]	5 [1; 15]	40 [20; 65]	5 [1; 15]	5 [1; 15]	20 [5; 35]	10 [2; 20]	15 [3; 30]
Political other	25 [10; 45]	10 [2; 20]	0 [0; 0]	44 [20; 70]	15 [3; 30]	5 [1; 15]	5 [1; 15]	0 [0; 0]	49 [25; 74]
Fiscal positives	80 [50; 114]	10 [2; 25]	60 [35; 89]	85 [55; 119]	5 [1; 15]	0 [0; 0]	10 [2; 25]	5 [1; 15]	0 [0; 0]
Fiscal negatives	40 [20; 65]	5 [1; 15]	10 [2; 25]	40 [20; 65]	10 [2; 25]	0 [0; 0]	20 [5; 35]	0 [0; 0]	0 [0; 0]
Fiscal other	60 [35; 89]	0 [0; 0]	20 [5; 40]	50 [25; 74]	25 [10; 45]	0 [0; 0]	0 [0; 0]	0 [0; 0]	0 [0; 0]
Administrative positives	30 [10; 50]	10 [2; 25]	15 [5; 30]	59 [35; 89]	15 [3; 30]	5 [1; 15]	5 [1; 15]	0 [0; 0]	10 [2; 25]
Administrative negatives	5 [1; 15]	0 [0; 0]	0 [0; 0]	35 [15; 60]	5 [1; 15]	0 [0; 0]	5 [1; 15]	0 [0; 0]	0 [0; 0]
Administrative other	10 [2; 25]	0 [0; 0]	10 [2; 25]	45 [25; 70]	5 [1; 15]	0 [0; 0]	0 [0; 0]	0 [0; 0]	0 [0; 0]
Total	135, 65, 95	74, 30, 10	75, 15, 30	204, 115, 139	25, 20, 45	35, 5, 5	30, 45, 5	15, 10, 0	85, 15, 49

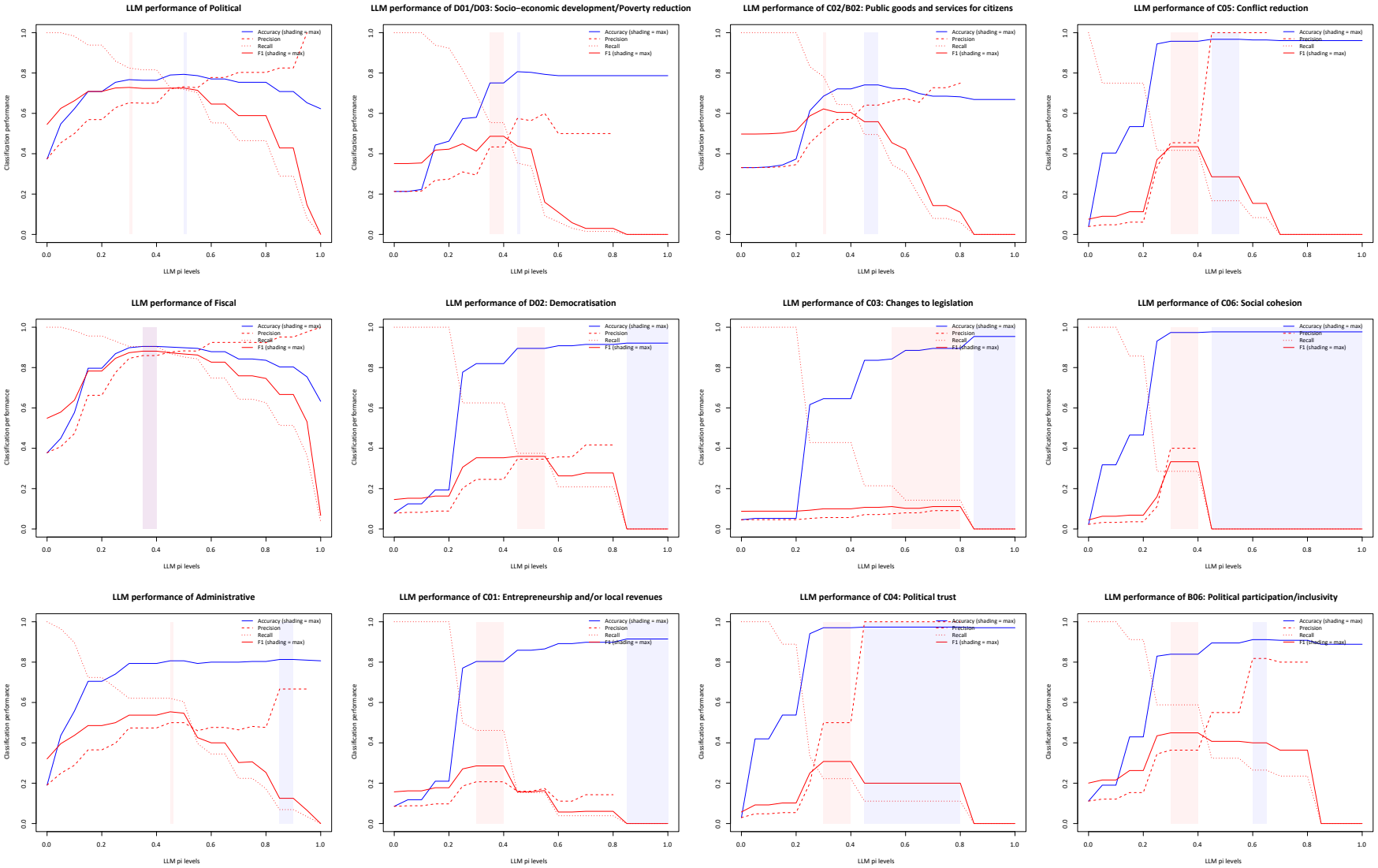
Evidence Gap Map (EGM) on the effect of dimensions of decentralization on relevant outcomes. Square brackets show adjusted 95 % confidence intervals, with floor constraint on lower bounds to not fall below the actual respective sample count.

Quelle: DEval, eigene Darstellung

Impact labels: *D01/D03 = Contribution to socio-economic development/Poverty reduction; D02 = Contribution to democratisation.*

Outcome labels: *C01 = Increases in entrepreneurship and/or local revenues; C02/B02 = Improved provision of services of general interest/provision of public goods; C03 = Changes to legislation that allow for more decentralization; C04 = Increased political trust; C05 = Conflict reduction; C06 = Increased social cohesion; B06 = Increased political participation and/or inclusivity.*

Abbildung 4 In-Sample-Performanz der LLM-Schätzwerte



Quelle: DEval, eigene Darstellung

Anmerkung: Vergleich der In-sample-Performanz des LLMs bei der Zuordnung einzelner Publikationseinträge zu den jeweiligen Einfluss- und Ergebnisvariablen. Blau zeigt die „Accuracy" (Schattierung: Maximalwert), rot zeigt die „Precision" (gestrichelte Linie), den „Recall" (gepunktete Linie) und den F1 Wert (durchgezogene Linie; Schattierung: Maximalwert).

3. EVALUIERUNGSSYNTHESE

3.1 Datengrundlage

Für sämtliche, auf Grundlage des Dezentralisierungsportfolios identifizierte Vorhaben, wurden Evaluierungsberichte bei BMZ, GIZ und KfW angefordert. Diese Anfrage wurde durch eine systematische Recherche öffentlich zugänglicher Quellen abgeglichen und ggf. ergänzt – darunter das BMZ-Transparenzportal, die KfW-Datenbank „Ideal“ sowie die GIZ-Publikationsdatenbank. Gemäß der DEval-Leitlinie zum sicheren Umgang mit generativer künstlicher Intelligenz (DEval, 2025) konnten letztlich ausschließlich öffentliche Dokumente in die Analyse einbezogen werden. Nach Abgleich mit den Durchführungsorganisationen KfW und GIZ umfasste der finale Analysekorpus somit 80 veröffentlichte Evaluierungsberichte (52 von der GIZ, 28 von der KfW).

Diese Berichte wurden im Zeitraum zwischen 1999 und 2022 publiziert und beziehen sich auf insgesamt 104 unterschiedliche Vorhaben, die zwischen 1994 und 2022 durchgeführt wurden. Einige der evaluierten Vorhaben waren zum Zeitpunkt der Berichtserstellung noch nicht abgeschlossen. Die Berichte variieren erheblich in ihrem Umfang und lassen sich in drei Kategorien einteilen: Übersichtsformate mit durchschnittlich zwei Seiten, Kurzformate mit sechs bis 13 Seiten sowie vollständige Evaluierungsberichte mit mehr als 45 Seiten. Zusammengenommen beläuft sich der Gesamtumfang des Analysekorpus auf 1.133 Seiten.

3.2 Qualitätskriterien

Um in der nachfolgenden Analyse den erheblichen Unterschieden in Umfang und Format der Berichte Rechnung zu tragen, wurden alle Berichte entlang der folgenden Qualitätsmerkmale auf einer vierstufigen Skala („nicht erfüllt", „eher nicht erfüllt", „eher erfüllt" und „vollständig erfüllt") bewertet:

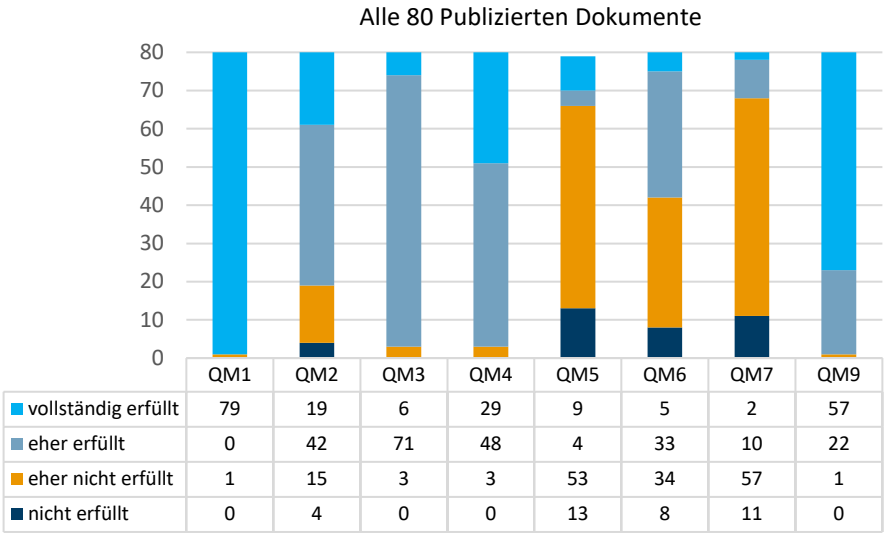
- QM1 Beschreibung des Evaluierungsgegenstands
- QM2 Beschreibung des Kontexts der Entwicklungsmaßnahme
- QM3 Berücksichtigung des Kontexts bei Ergebnissen
- QM4 Darstellung der Wirkungszusammenhänge
- QM5 Beschreibung des Erkenntnisinteresses
- QM6 Nachvollziehbarkeit der Informationsquellen
- QM7 Darstellung der Angemessenheit des methodischen Vorgehens
- QM9 Kohärenz von Daten-Ergebnissen-Schlussfolgerungen

Diese Qualitätsmerkmale basieren auf der Analysemethodik der „Meta-Evaluierung zur Qualität von (Projekt) Evaluierungen in der Deutschen Entwicklungszusammenarbeit" (DEval, 2022, Online-Anhang, S. 81-91[10]). Die Bewertung der Qualitätsmerkmale erfolgte mittels eines Large Language Models (LLama 3.3 70B), welche durch das händische Kodieren einer zufälligen Stichprobe von 20 Berichten validiert wurde. Die Einbindung der KI erfolgte in Zusammenarbeit mit dem Dienstleister Neofonie GmbH.

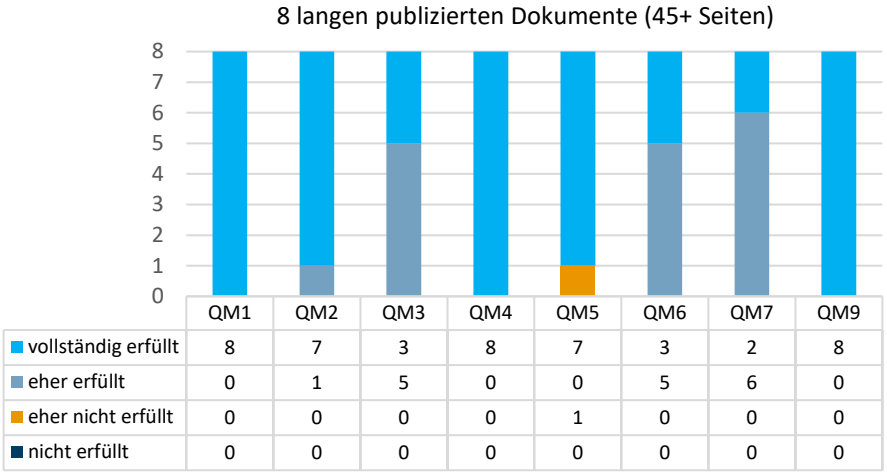
Die Ergebnisse zeigen, dass einige der Qualitätsmerkmerkmale von einem Teil der Evaluierungsberichte – insbesondere den Übersichts- und Kurzformaten – nicht erfüllt werden. Auf Basis dieser Ergebnisse wurden alle Berichte von der weiteren Analyse ausgeschlossen, die mindestens zwei Qualitätsmerkmale „nicht erfüllen“ oder mindestens drei Qualitätsmerkmale „eher nicht erfüllen“. So verblieben 35 Evaluierungsdokumente von GIZ und KfW für die nachfolgende Synthese, die sich auf 61 Vorhaben aus einem Zeitraum von 30 Jahren beziehen.

[10] Siehe *https://www.deval.org/fileadmin/Redaktion/PDF/05-Publikationen/Berichte/2022_Meta-Q/2022_DEval_Meta-Q_Onlineanhang.pdf*

Abbildung 5 Resultate Qualitätsmerkmale



	QM1	QM2	QM3	QM4	QM5	QM6	QM7	QM9
vollständig erfüllt	79	19	6	29	9	5	2	57
eher erfüllt	0	42	71	48	4	33	10	22
eher nicht erfüllt	1	15	3	3	53	34	57	1
nicht erfüllt	0	4	0	0	13	8	11	0



	QM1	QM2	QM3	QM4	QM5	QM6	QM7	QM9
vollständig erfüllt	8	7	3	8	7	3	2	8
eher erfüllt	0	1	5	0	0	5	6	0
eher nicht erfüllt	0	0	0	0	1	0	0	0
nicht erfüllt	0	0	0	0	0	0	0	0

Quelle: DEval, eigene Darstellung

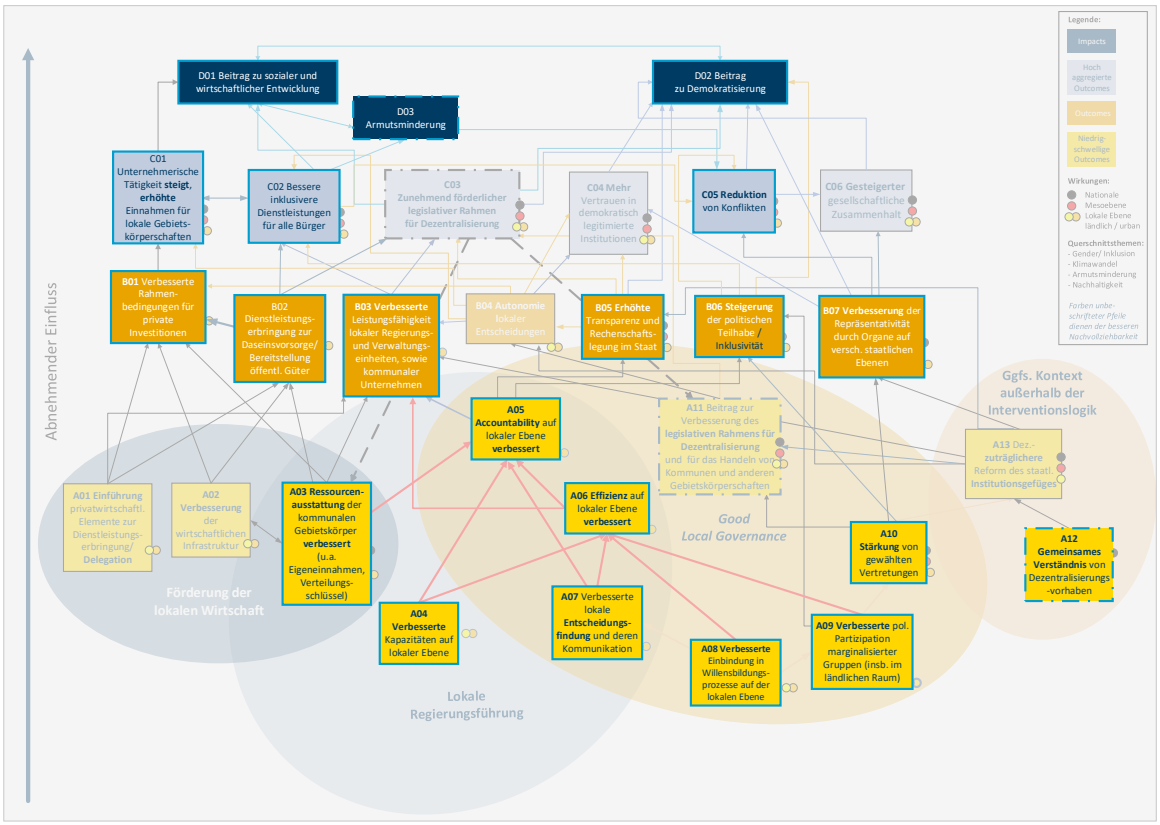
3.3 Synthese

Im weiteren Verlauf wurden die verbliebenen Berichte auf Zielsetzungen und Belege für Wirkungszusammenhänge durchsucht. Die so identifizierten Datenpunkte wurden zusammengetragen und im Abschnitt 7.1 „Effektivität“ im Evaluierungsbericht aggregiert dargestellt.

Zudem wurden sie für die Fallstudienländer grafisch auf Vorhabenebene abgebildet, um so die vorhandene Evidenz – und mögliche Evidenzlücken – beispielhaft nachzuvollziehen. Diese Abbildungen wurden daraufhin innerhalb der Fallstudienländer übereinandergelegt, sodass ein Gesamtbild der Zielsetzungen und Wirkungsbelege für jedes Fallstudienland entsteht. Abbildung 6 und Abbildung 7 zeigen diese Abbildungen beispielhaft für Sambia.[11] Ausgegraute Elemente stehen hier für Inhalte, die nicht in den Berichten wiedergefunden werden konnten.

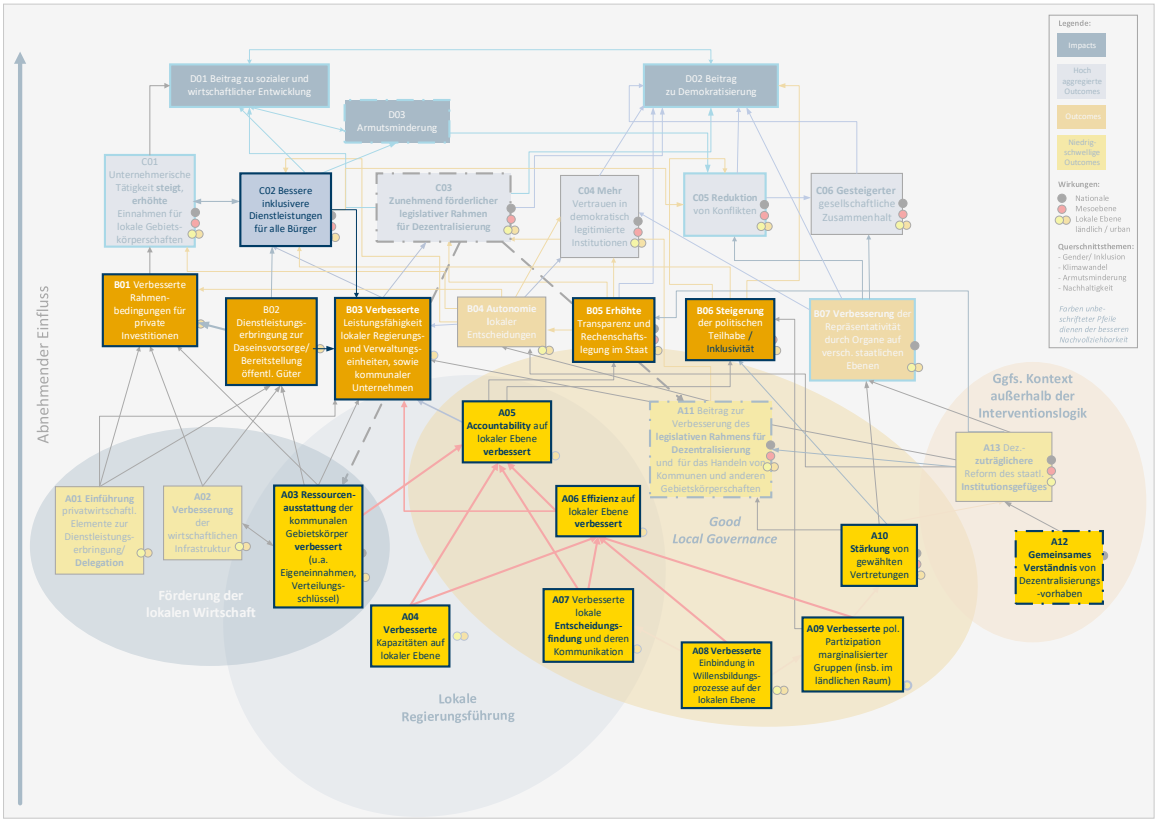
[11] Vgl. auch die im Evaluierungsbericht verwendete vollständige Wirkungslogik, Abbildung 1 auf S.12.

Abbildung 6 Sambia, Zielsetzungen aller analysierten Vorhaben



Quelle: DEval, eigene Darstellung

Abbildung 7 Sambia, Wirkungsbelege aller analysierten Vorhaben



Quelle: DEval, eigene Darstellung