Pixel und Promille. Eine methodisch-explorative Untersuchung zur Erfassung von Alkoholwerbung mittels generativer multimodaler Sprachmodelle in den Computational Social Science.

Merlin Reinhardt

September / Oktober 2025 *

Zusammenfassung

Die Exploration der Möglichkeiten und Grenzen des Einsatzes von multimodalen generativen Sprachmodellen im Bereich der sozialwissenschaftlichen Forschung für den Anwendungsbereich Bildanalyse befindet sich noch im Prozess. Diese Arbeit untersucht die Anwendungsmöglichkeiten und Herausforderungen generativer multimodaler Sprachmodelle am Beispiel einer soziologischen Fallstudie zu Alkoholwerbung in europäischen Supermarktkatalogen. Hierfür werden drei methodische Ansätze systematisch auf ihre praktische Anwendbarkeit und Ergebnisqualität hin verglichen. Die Ergebnisse des methodischen Vergleichs zeigen ein deutliches Spannungsfeld auf: Während proprietäre Modelle exzellente Ergebnisse bei hoher Effizienz liefern, scheitern offene Modelle an technischen Hürden oder unzureichender Performanz. Dies stellt Forschende vor ein Dilemma zwischen praktischer Machbarkeit und wissenschaftlichen beziehungsweise forschungspraktischen Anforderungen wie Reproduzierbarkeit und Datenautonomie. Die inhaltliche Analyse der Werbekataloge deckt Länderunterschiede in der Präsenz, Platzierung und qualitativer Präsentation von Alkoholwerbung auf und deutet auf unterschiedliche nationale Werbekulturen, insbesondere im Hinblick auf den Schutz von Minderjährigen hin.

*Hausarbeit im Forschungspraktikum 'Generative KI für sozialwissenschaftliche Forschung' im Sommersemester 2025 bei Dr. Maximilian Weber an der Goethe-Universität Frankfurt am Main im Rahmen des Master of Arts Soziologie.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1 Einleitung

Die schnellen und fortschreitenden Entwicklungen im Bereich der generativen Sprachmodelle stoßen auch im Bereich der sozialwissenschaftlichen Methodenforschung auf große Resonanz (Weber and Reichardt, 2023, S. 1) und werden durch eine umfangreiche und fortschreitende Exploration der Anwendungsmöglichkeiten begleitet. Im Bereich der Textverarbeitung ist mittlerweile auf diverse Studien hinzuweisen, während die Potenziale für den Bereich der Bildverarbeitung bislang von einzelnen Autor_innen adressiert werden, insgesamt jedoch untererforscht sind (siehe Kapitel 2). Dabei weisen nicht nur generative textbasierte Modelle (LLM), sondern auch generative multimodale Modelle (MM) beachtliche Entwicklungen auf. Bislang bestehende Herausforderungen, welche in der Nutzung von Methoden der Computational Social Science im 'Vision'-Bereich, etwa für umfassende Analysen von Bilddaten, bestanden haben, könnten durch neuere MMs überwunden oder reduziert werden, wodurch sich neue Forschungsmöglichkeiten eröffnen (Law and Roberto, 2025, S. 6-7).

Neben der fachlichen Diskussion um die neuen Möglichkeiten auf technischer Ebene verläuft eine ebenso wichtige wissenschaftstheoretische und ethische Debatte über die konkrete Gestaltung von Forschungsprozessen, welche generative Sprachmodelle einbinden. Während es hier insbesondere aus Gründen der Reproduzierbarkeit und der Datenautonomie (Ollion et al., 2023, zitiert nach Weber and Reichardt, 2023) als präferiert und empfohlen angesehen wird, offene Modelle zu nutzen (Spirling, 2023, zitiert nach Weber and Reichardt, 2023), so gibt es im Bereich der Bildverarbeitung mit MMs in den Sozialwissenschaften aufgrund einer bislang dünnen Studienlage hinsichtlich praktischer Umsetzbarkeit für einzelne Forschende Unklarheiten, was die Wahl eines offenen Modells gegenüber der Nutzung eines proprietären Modells in der Forschungspraxis hinsichtlich technischer Herausforderungen und Performanz bedeuten kann.

Um diese praxisrelevante Lücke zu schließen und die Einstiegshürden für zukünftige Forschungsprojekte zu senken, untersucht diese Arbeit systematisch die Anwendbarkeit verschiedener MM-Ansätze für die sozialwissenschaftliche Forschung. Ziel ist es, Unklarheiten in Fragen der technischen Umsetzung hinsichtlich eines Einsatzes von MMs zu reduzieren und eine exemplarische Anwendungsmöglichkeit zu demonstrieren. Die zentrale Forschungsfrage dieser Arbeit lautet:

> *Welche Implikationen ergeben sich aus dem Vergleich von offenen und proprietären multimodalen Modellen für die sozialwissenschaftliche Forschungspraxis, demonstriert am Beispiel einer kultursoziologischen Analyse der visuellen Alkoholwerbung in europäischen Supermarktkatalogen?*

Zur Beantwortung dieser Frage gliedert sich die Arbeit in einen methodischen Vergleich und eine exemplarische Fallstudie, in welcher MMs für zentrale Arbeitsschritte

genutzt werden. Nach Erarbeitung der entsprechenden Forschungsstände (Abschnitt 2) werden im methodischen Teil drei unterschiedliche Forschungsansätze gegenübergestellt, welche sich hinsichtlich Forschungsaufbau, Reproduzierbarkeit und Datenautonomie unterscheiden: ein vollständig lokaler Ansatz mit einem offenen Modell (Abschnitt 3.2.1), ein cloud-basierter Ansatz mit stärkerer Rechenleistung bei gleichzeitiger Nutzung eines offenen Modells (Abschnitt 3.2.2) und ein Ansatz, welcher per Programmierschnittstelle (API) auf ein proprietäres Modell zurückgreift (Abschnitt 3.2.3).

Die damit verschränkte exemplarische kultursoziologische Fragestellung lautet:

> *Welche qualitativen und quantitativen Unterschiede weisen die Werbekataloge großer Einzelhandelsketten verschiedener europäischer Länder in der Präsentation alkoholischer Produkte auf?*

Die Frage bietet Möglichkeiten, die MMs in ihrer Leistungsfähigkeit durch unterschiedliche Analyseaufgaben (bspw. binäre Klassifikation, Objekterkennung, Zählung) herauszufordern. Ausgangspunkt dieser kultursoziologischen Fragestellung war, dass Konsument_innen in Deutschland und anderen europäischen Staaten im Alltag gegenüber alkoholischen Produkten oftmals unfreiwillig stark exponiert sind, etwa im Kassenbereich von Supermärkten. Dabei zählt Europa neben Afrika zu den Weltregionen, in welchen die meisten Todesfälle mit Alkoholbezug festzustellen sind (World Health Organization, 2024, xii) und circa 10 Prozent der Bevölkerung von Alkoholsucht oder gefährlichem Alkoholkonsumverhalten betroffen sind (World Health Organization, 2024, S. 55). Durch die leichte Verfügbarkeit von Werbekatalogen sollen diese im Rahmen dieser Arbeit dazu genutzt werden, die Exponierung gegenüber Alkoholwerbung und somit gesellschaftliche Strukturen in Bezug auf Alkoholumgänge zu erforschen.

Die Ergebnisse des methodischen Vergleichs (Abschnitt 4.1) können als praxisnahe Orientierungshilfe für Forschende dienen und somit zur methodischen Expertise in den Computational Social Science zu beitragen. Die Ergebnisse der exemplarischen Fallstudie (Abschnitt 4.2) liefern empirische Einblicke in Unterschiede der visuellen Vermarktung von Alkohol in Europa.

2 Hintergrund und Forschungsstand

2.1 Nutzung von generativen multimodalen Modellen in den Sozialwissenschaften

Während im Bereich der Textverarbeitung auf diverse Arbeiten sowohl im Bereich der offenen als auch proprietären Modelle mit vielversprechenden Ergebnissen hinsichtlich der Nutzung von LLMs im sozialwissenschaftlichen Bereich hinzuweisen ist (bspw. Alizadeh et al. 2023 & Gilardi et al. 2023, zitiert nach Weber and Reichardt 2023; Alizadeh et al.

2025; Gunes and Florczak 2023), ist der Forschungsstand im Bereich der Nutzung von multimodalen Modellen weniger umfassend ausgeprägt. Dabei nehmen Bilder in sozialen und politischen Prozessen eine immer zentralere Rolle ein, woraus sich wertvolle Forschungspotentiale ergeben (Williams et al., 2020).

Im Bereich der computergestützen Bildanalyse standen Forschende bis vor wenigen Jahren vor der Herausforderung entweder selbst ein neuronales Netz zu trainieren oder ein vortrainiertes neuronales Netz durch *fine-tuning* nutzbar zu machen (bspw. Sheng et al., 2021; Kim et al., 2024; Hwang and Naik, 2023). Diese Ansätze, welche dem Feld *machine learning* zuzuordnen sind, erfordern einen erheblichen Arbeitsaufwand, ausgeprägte technische Expertise und bedeutsame Rechenleistung, wodurch neue Möglichkeiten, multimodal-generative *Foundation Models* (MM) für Bildanalysen zu verwenden, als sehr attraktiv erscheinen (Law and Roberto, 2025, S. 17; Torres and Cantú, 2022).

Aktuell ist auf verschiedene Studien hinzuweisen, welche MMs als Analysewerkzeug nutzen und ergründen, ob sich Vorteile, welche im Text-LLM-Bereich festgestellt wurden (Gilardi et al., 2023), auch auf den Bereich der MM-gestützten Bildanalyse übertragen lassen. Zu nennen sind hier Studien der Autor_innen Sarmadi et al. (2025) sowie Law and Roberto (2025), welche beide das GPT-4o Modell (OpenAI, 2024) nutzen, um Satellitenbilder kodieren zu lassen. In beiden Fällen ordnen die Autor_innen die Ergebnisse als vielversprechend ein. Gleichzeitig werden proprietäre Modelle genutzt, wodurch der Forschungsstand hinsichtlich der Performanz offener Modelle unbeleuchtet bleibt. Ferner bleiben die Fragen, welche Ergebnisqualität MMs in Bezug auf andere Bilddatentypen als Satellitenbilder erzeugen und wie gut MMs mit aktuellen Bilddaten umgehen können, welche *garantiert* nicht in den Trainingsdaten enthalten sein können, liegen *status quo* keine ausreichenden Forschungsergebnisse vor.

2.2 Alkoholkulturforschung und Werbekataloge als Datenquelle

Während neben theoretischen soziologischen Betrachtungen von Werbung (bspw. Cluley and Nixon 2019) auf qualitativ-inhaltsanalytische Untersuchungen von Werbedarstellungen in Verschränkung mit alkoholischer Produktpräsentation (bspw. Hall and Kappel 2018) hinzuweisen ist, so scheinen privatwirtschaftliche Werbekataloge als leicht zugängliche, regelmäßig erscheinende und vorstrukturierte Datenquelle in den Sozialwissenschaften bislang nur am Rande die Aufmerksamkeit sozialwissenschaftlich Forschenden zu erfahren. Dabei ist anzunehmen, dass Werbeprospekte durchaus soziale Normen und kollektive Präferenzen abbilden und somit im Zeitverlauf auch sozialstrukturelle Veränderungen erfassbar machen.

Untersuchungen von Werbekatalogen fanden etwa durch eine Studie von Stevenson (2002) statt, in welcher die Darstellung von Schwarzen Personen quantitativ-inhaltsanalytisch untersucht wurde. Qualitativ-inhaltsanalytisch forschten Isański and Leszkowicz (2011)

im Kontext von Eye-Tracking-Versuchen zur Wahrnehmung von Werbekatalogen. Weitere Untersuchungen finden sich im Bereich der Marketing-Forschung, einschließlich Ländervergleichsanalysen (bspw. Tan et al., 2023).

Untersuchungen von Alkoholdarstellungen in Werbekatalogen sind insbesondere im Bereich der *Public Health*-Forschung zu finden. Ähnlich der Forschungsfrage dieser Arbeit untersuchten Johnston et al. über zwölf Monate hinweg zwei australische Supermarktketten, mit dem Ziel ein genaueres Bild über Marketingstrategien und die Produktpräsentationen zu entwickeln (Johnston et al., 2017). Ausgangspunkt für diese Studie ist der Befund, dass der gesellschaftliche Fokus auf das Thema Alkohol aber auch soziale Ungleichheiten und andere gesellschaftliche Strukturen Einflüsse auf individuellen Alkoholkonsumentscheidungen haben (Livingston 2013, S. 113, zitiert nach Johnston et al. 2017).

Unter Betrachtung der Vorteile von Werbekatalogen als Datenquelle (leichter Datenzugang, Erscheinen seit Jahrzehnten in regelmäßigen Abständen, internationale Verbreitung, klare Strukturiertheit, numerische Preiskodierungen, *et cetera*) scheinen Werbekataloge trotz beschriebener Studien eine unterschätzte Datenquelle zu sein, da sie durch Zeitverlaufsanalysen und relative Ländervergleiche als Indikator für gesellschaftliche Entwicklungen und Einstellungen dienen können.

3 Daten und Methoden

Law and Roberto (2025, S. 9-12) bieten einen methodischen Rahmen zur Arbeit von multimodalen Modellen in den Sozialwissenschaften, welcher auf dem 'Agnostic Approach'-Ansatz von (Grimmer et al. 2022, zitiert nach Law and Roberto 2025, S. 9-10) aufbaut. Dieser methodische Rahmen bietet in angepasster Form die methodische Struktur der vorliegenden Arbeit und wird in Abbildung 1 schematisch dargestellt. Die vorbereitenden Arbeiten umfassen neben der Entwicklung der Fragestellung die Auswahl der Supermärkte, das Sammeln und Speichern der Prospekte in geeignetem Format, sowie eine erste Modellexploration. Im zweiten Schritt erfolgte theoriegeleitet die Entwicklung der Bild-Labels beziehungsweise der Variablen sowie die Prompt-Entwicklung. Aus Gründen der Nachvollziehbarkeit und um selbst nicht den Überblick zu verlieren, ist an dieser Stelle der Prozess systematisch zu dokumentieren. Im Rahmen von Pre-Tests mit einzelnen Modellen passte ich an dieser Stelle die Labels mehrfach an die Modellkapazitäten an. Beispielsweise versuchte ich durch Komplexitätsreduktionen die Ergebnisse der Pre-Tests zu verbessern. Die Label-Entwicklung und -Testung stand dabei bereits in enger Verschränkung mit der Prompt-Entwicklung und der Entwicklung der Python-Skripte. Ein weiterer zentraler Arbeitsschritt ist an dieser Stelle die Generierung einer randomisierten Stichprobe (hier n=150), welche manuell annotiert wird und später als *Human Goldstandard* zur Evaluierung der Modellergebnisse herangezogen wird. In einem dritten

Schritt wurde dann die Annotation je nach Ansatz (siehe Abschnitt 3.2) durch die MMs durchgeführt, die Ergebnisse evaluiert und der Ansatz mit den besten Ergebnissen zur Annotation des Gesamtdatensatzes und zur Inferenz genutzt.

Dieses Kapitel dient, nach einer detaillierteren Erarbeitung des methodischen Vorgehens (Abschnitt 3.1), dazu, die drei verschiedenen Ansätze, welche in Abschnitt 3.2 erläutert werden, vorzustellen. Für detaillierte Einblicke und aus Gründen der Nachvollziehbarkeit sind die entwickelten und verwendeten Python-Skripte, sowie die Prompts in meinem GitHub-Repository öffentlich abrufbar.

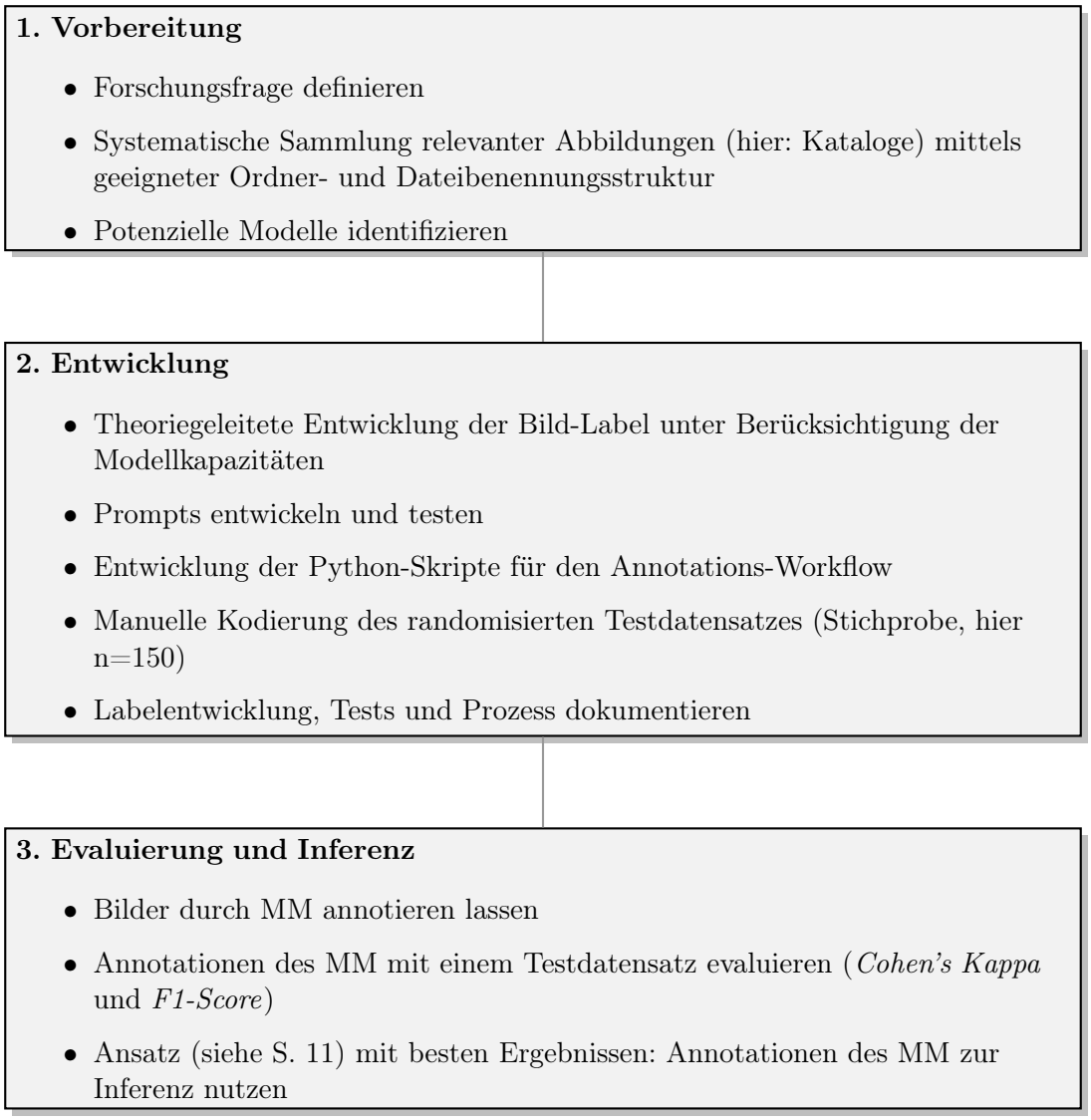


Abbildung 1: Framework zur Bildanalyse mit generativen multimodalen Modellen.

Quelle: Eigene, angepasste Darstellung in Anlehnung an Law and Roberto (2025, S. 9-12).

3.1 Methodisches Vorgehen

3.1.1 Datensatz und Variablenkodierung

Der im Rahmen dieser Arbeit entwickelte Datensatz umfasst 1562 Werbekatalogseiten aus insgesamt 40 Werbekatalogen. Für jedes der fünf untersuchten Länder lagen mindestens für fünf Supermärkte ein oder mehrere Prospekt(e) vor. Die Prospekte wurden in zwei Zeiträumen gesammelt, einer davon im August und einer im September. Dadurch ist auszuschließen, dass die Kataloge in den Trainingsdaten der verschiedenen Modelle enthalten sein könnten. Die Kataloge, vorliegend im PDF-Dateiformat enthielten 1562 Katalogseiten, aus welchen randomisiert 150 Seiten ausgewählt und manuell annotiert wurden.

Entsprechend den Forschungsfragen wurden sieben Variablen entwickelt, anhand dessen die Werbekatalogseiten annotiert werden sollten. Davon waren fünf kategoriale (*alc, product, warning, reduc, child*) und zwei metrisch-skalierte Variablen (*prod_pp, prod_alc*) enthalten. Die genaue Kodierung ist im Codebuch (Tabelle 6) abgebildet. Für die Entwicklung der Variable *product*, welche zur Kategorisierung alkoholische Produkte dient, bezog ich mich auf die Klassifikationen des Deutschen Krebsforschungszentrums, sowie des Statistischen Bundesamtes, welche alkoholische Produkte anhand ihrer unterschiedlichen Herstellungsarten sowie der sich unterscheidenden Alkoholgehalte in die drei Kategorien Bier, Wein und weinverwandte Erzeugnisse und Spirituosen einteilen (Schaller et al., 2022, S. 1-6; Statistisches Bundesamt, 2008, S. 80).

3.1.2 Modellauswahl

Die Modellauswahl ist im Bereich der MMs aufgrund höherer Entwicklungskosten gegenüber der Auswahl im Bereich der LLMs zwar geringer. Doch auch hier gibt es verschiedene öffentlich nutzbare Modelle, sowohl proprietären als auch offenen Charakters. Insbesondere die offenen Modelle werden außerdem häufig in unterschiedlichen Modell- beziehungsweise Parametergrößen angeboten (Law and Roberto, 2025, S. 6-7). Bei der Auswahl meiner Modelle habe ich mich an der Übersicht und den Auswahlkriterien von Law and Roberto (2025, S. 7-8) orientiert und mit aktuelleren Modellen verglichen. Für die Auswahl des proprietären Modelles orientierte ich mich neben den Werten des etablierten MMMU-Benchmarks (Yue et al. 2024, zitiert nach Law and Roberto 2025, S. 7) auch an den Kosten für die Nutzung der Modell-Programmierschnittstelle (API). Im Bereich der offenen Modelle orientierte ich mich an den Modellgrößen und den unterschiedlichen Hardware-Ausstattungen der Ansätze. Insbesondere für Ansatz 1 nutzte ich einen iterativen Prozess des Ausprobierens und Evaluierens.

3.1.3 Prompt-Entwicklung

Die Aufforderung, Prompt genannt, welche an das jeweilige Modell übergeben wird, beeinflusst durch ihren Inhalt, ihre Struktur und ihren Umfang in Bezug auf Beispiele und Handlungsanweisungen die Ergebnisse des Modells. Bei der Erstellung der Prompts für die Anforderungen dieser Arbeit wurden dem Modell eine *Persona* zugewiesen, eine strukturierte Antwortvorlage definiert, sowie explizite Anweisungen übermittelt, welche Teile des Kontextes berücksichtigt und welche ignoriert werden sollen (White et al., 2023, S. 7, 12, 16). Weitergehend ist festzustellen, dass sich Few-Shot Prompting, also das Übermitteln von musterhaft-idealtypischen Beispielausgaben, in vielen Fällen als effektiver erweist als Zero-Shot Prompting (Brown et al. 2020, zitiert nach Weber and Reichardt 2023). Da jedoch im Bereich der Bildverarbeitung die Anzahl der vom Modell zu verarbeitenden Tokens signifikant höher ist, als im Bereich der Textverbeitung stoßen insbesondere lokale und semi-lokale Ansätze, wie im Rahmen dieser Arbeit umgesetzt, an Grenzen der technischen Umsetzbarkeit (hierzu siehe Ergebnisse in Kapitel 4.1.1). Somit wird aus Gründen der technischen Umsetzbarkeit in dieser Arbeit auf Methoden des Few-Shot-Prompting verzichtet und One-Shot-Prompting angewendet.

3.1.4 Reproduzierbarkeit und Evaluierung

Reproduzierbarkeit als zentrales wissenschaftliche Qualitätskriterium konnte für die drei Ansätze in unterschiedlichem Maße erfüllt werden. Für die offenen Ansätze (Ansatz 1 und 2) war es möglich einen sogenannten 'Seed' im Ausführungsskript einzubauen, wodurch eine genaue Reproduzierbarkeit bei vorliegendem, gespeicherten Modell gewährleistet werden kann. Für den proprietären Ansatz (Ansatz 3) stand die Option eines Seeds nicht zur Verfügung, ebenso wenig wie die Option, das Modell lokal zu speichern. Um dem Reproduzierbarkeitsanspruch dennoch in einem möglichst hohen Ausmaß gerecht zu werden, setze ich den Parameter *temperature* bei allen drei Ansätzen auf den niedrigsten Wert, wodurch der Output des Modells bei gleichem Input weniger Varianz aufweist (Law and Roberto, 2025, S. 22).

Angelehnt an Weber and Reichardt (2023) und Law and Roberto (2025) werden zur Evaluierung der Annotationsergebnisse verschiedene Evaluationsmaße herangezogen. In einem ersten Schritt wird die grundsätzliche Eignung des Modells mittels *Cohens Kappa* (nachfolgend: Kappa-Score) untersucht. Hierbei wird die Klassifikation des MMs mit den manuellen Kodierungen, welche als Gold-Standard beziehungsweise Grundwahrheit herangezogen werden, unter Berücksichtigung der Möglichkeit zufälliger Übereinstimmungen, verglichen (Weber and Reichardt, 2023, S. 7). In einem zweiten Schritt wird der *F1-Score*, bestehend aus *Precision* und *Recall*, berechnet, welcher spezifischer auf das Fehlerverhalten des Modells eingeht. Der *Recall*-Wert beschreibt, wie viele der relevanten Fälle das Modell identifizieren konnte, während der *Precision*-Wert untersucht, wie viele

der identifizierten Items tatsächlich relevant sind. Um unterschiedliche Klassenverteilungen relativ nach Größe proportional abbilden zu können, wird der gewichtete *F1-Score* (auch: *weighted F1-Score*) verwendet (Weber and Reichardt, 2023, S. 7). Der Wertebereich des *Kappa-Score* reicht von -1 bis 1, wobei Werte ab 0,6 und höher bereits als 'substantiell-positiv' betrachtet werden (Landis and Koch, 1977, S. 165). Der Wertebereich des *F1-Score* reicht von 0 bis 1. Im Kontext dieser Arbeit sollen *F1*-Werte ab 0,7 als zufriedenstellend beziehungsweise als Indikator für eine sinnvolle wissenschaftliche Nutzbarkeit des Modells interpretiert werden[1]. Für die beiden metrisch-skalierten Variablen *prod_pp* und *prod_alc*, für welche das Modell die Anzahl an Produkten auf einer Seite zu ermitteln hatte, nutze ich zur Evaluierung der Ergebnisse die absolute und die quadrierte Fehlerabweichung, *Mean Absolute Error (MAE)* und *Root Mean Square Error (RMSE)*. Durch die Kombination der beiden Metriken kann ein solider Eindruck gewonnen werden, wie stark die Abweichungen des Modells im Mittel sind.

3.2 Ansätze

Im Verlauf dieser Arbeit entwickelte ich verschiedene Ansätze, um auf Herausforderungen, die im Prozess der Arbeit mit den MM entstanden sind, zu reagieren. Die entwickelten Ansätze unterscheiden sich konzeptionell stark hinsichtlich Reproduzierbarkeit und Datenautonomie.

3.2.1 Ansatz 1: offen-lokal

Ansatz 1 (offen-lokal) zeichnet sich durch den höchsten Grad an Reproduzierbarkeit und Datenautonomie aus. Die Hardwareausstattung bestand dabei aus meinem eigenen Laptop mit *16GB*-Arbeitsspeicher, einem *Intel i7-6600U × 4*-Prozessor und keiner NVIDIA- oder AMD-Grafikkarte. Die genutzten Modelle waren kleine offene Modelle aus der *Ollama*-Umgebung, welche lokal auf dem Laptop gespeichert wurden. Da Geräte dieser Art mittlerweile zur "Standardausstattung" von Forschenden gehören, bietet dieser Ansatz niedrigschwellige und wissenschaftstheoretisch wünschenswerte Forschungsmöglichkeiten. Denn durch die lokalen Modellberechnungen findet keine Weitergabe von Forschungsdaten an Dritte statt. Nachteilig können sich bei diesem Ansatz jedoch die geringe Rechenleistung und Arbeitsspeichergrößen auswirken, sodass lediglich sehr kleine, kondensierte Modelle genutzt werden können. Die im Rahmen dieses Ansatzes genutzten offenen Modelle wurden nach den in Abschnitt 3.1.2 dargestellten Kriterien ausgewählt und sind in Kapitel 4.1.1 erläutert.

[1] Diese Einschätzung basiert auf unseren Erarbeitungen im Rahmen des Seminars.

3.2.2 Ansatz 2: offen-cloudbasiert

Ansatz 2 (offen-cloudbasiert) zeichnet sich durch die Nutzung von offenen Modellen in einer cloud-basierten Rechenumgebung aus. Dadurch lassen sich die wissenschaftlichen Vorteile offener Modelle nutzen, während die technischen Limitationen des offen-lokalen Ansatzes (3.2.1) verringert werden. Im Gegensatz zu Ansatz 1 werden die Forschungsdaten jedoch nicht auf dem lokalen Computer verarbeitet, sondern an ein externes Rechenzentrum, etwa ein Hochschulrechenzentrum oder das Rechenzentrums eines privaten Anbieters weitergereicht. Insbesondere beim Vorliegen sensibler Forschungsdaten muss hier kritisch überprüft werden, ob der Ansatz anwendbar ist. Da dies im Rahmen dieser Arbeit jedoch nicht der Fall ist, wurde für diesen Ansatz die cloudcomputing Umgebung *Google Colaboratory* der Firma Alphabet genutzt. Hier standen *51GB*-RAM, ein leistungsstärkerer Prozessor (CPU) sowie eine Grafikkarte (GPU) zur Verfügung. Unter Abwägung der verfügbaren Modelle, der Modellgrößen und der technischen Ausstattung fand im Rahmen dieses Ansatzes das Modell *llama3.2-vision:11b-instruct-fp16* der Firma Meta verwendung. Die *Llama*-Modelle gelten im Bereich der offenen Vision-Modelle als leistungsstark und die Version *3.2-vision:11b-instruct-fp16* war mit *21GB* die Version mit der größten Modellgröße.

3.2.3 Ansatz 3: proprietär

Ansatz 3 (proprietär) erfüllt wissenschaftliche und datenschutztechnische Qualitätskriterien im Vergleich zu den anderen beiden Ansätzen im geringsten Maße. Konzept des Ansatzes ist, die Programmierschnittstelle (API) eines proprietären Modells zu nutzen. Somit lässt sich weder das verwendete Modell speichern, noch eine Output-Reproduzierbarkeit mittels Seed sicherstellen. Die Forschenden haben keine Kontrolle über die Modellverfügbarkeit, wodurch in Folge unternehmerischer Entscheidungen die Nachvollziehbarkeit des entstandenen Datenoutputs gefährdet sein kann. Die verwendeten Daten werden außerdem zwangsläufig an den privatwirtschaftlichen Modellanbieter weitergegeben, wodurch die Forschenden die Kontrolle über Verwendung der Daten verlieren. Somit kann dieser Ansatz lediglich bei nicht-sensiblen Forschungsdaten in Betracht gezogen werden. Die Vorteile eines proprietären Ansatzes können jedoch in einer erhöhten Performanz sowie einer leichten Nutzbarkeit liegen.

Die für diese Arbeit genutzte API ist die *gemini-api* der Firma Alphabet. Die Firma Alphabet zählt aktuell mit ihren *Gemini*-Modellen zu den großen kommerziellen Modellanbietern und bietet eine Auswahl an Vision-Modellen zu niedrigen Nutzungskosten an. Unter Vergleich der Modellkapazitäten und API-Kosten nutzte ich im Verlauf der Arbeit die Modelle *gemini-1.5-pro* und *gemini-2.0-flash* (Details siehe Abschnitt 4.1.3).

4 Ergebnisse

Im ersten Abschnitt dieses Kapitels werden die Ergebnisse der drei methodischen Ansätze evaluiert und die starken Unterschiede hinsichtlich der praktischen Nutzbarkeit und Anwendung erläutert. Im Anschluss finden sich die inhaltlichen Ergebnisse der kultursoziologischen Fragestellung.

4.1 Methodische Ergebnisse: Evaluierung der Ansätze und entstandene Herausforderungen

Da diese Arbeit im Rahmen eines Forschungspraktikums entstanden ist, möchte ich die folgenden drei Unterkapitel nutzen, um neben der fachlichen Evaluierung anhand der erarbeiteten Metriken auch gesammelte Erfahrungen und erlebte Herausforderungen zu schildern.

4.1.1 Ansatz 1: offen-lokal

Die Arbeit an diesem Ansatz war von viel 'Aus- und Herumprobieren' und ernüchternden Ergebnissen geprägt. Da mittlerweile auch im Bereich der MMs offene Modelle, beispielsweise über die Ollama-Plattform, angeboten werden, welche eine Modellgröße von wenigen Gigabyte (GB) aufweisen, war die erste Annäherung an die methodische Umsetzung dieses Ansatzes durch ein langes Ausprobieren mit diesen verschiedenen kleinen Modellen geprägt. Zu nennen sind hier die Modelle *qwen2.5vl:3b* und *qwen2.5vl:3b-q4_K_M* mit jeweils 3,2GB-Größe (Bai et al., 2025)[2], *llama3.2-vision:11b* mit 7,8GB-Größe (Meta Platforms, 2024), *llava:7b* mit 4,7GB-Größe (Liu et al., 2023) und *moondream* mit 1,7GB-Größe (Vikhyat, 2024).

Wie in Abschnitt 3.2.1 erläutert, bot das hierfür genutzte Gerät *16GB*-Arbeitsspeicher und einen *Intel i7-6600U × 4*-Prozessor, welcher aufgrund nicht vorhandener NVIDIA- oder AMD-GPU die Rechenprozesse übernahm. Trotz vielfältiger Versuche, den Workflow zu optimieren und an die Hardwareausstattung anzupassen (u.a. *Reduktion der Bildqualität, Graustufen, Einzelbildbearbeitung, gebündelte 'Batch'-Bildbearbeitung, Reduktion der Prompt-Komplexität* lieferten alle Modelle eine ungenügende Performanz [3]. Aufgetretene

[2] Hinweis zur Zitation von LLMs und MMs im Rahmen dieser Arbeit: Sofern auf der Ollama-Plattform, über welche ich die Modelle abgerufen habe ein *Technical Report* für das Modell verlinkt war, bezieht sich die Zitation auf diesen Bericht. Da ich jedoch häufig keinen solchen Bericht gefunden habe, bezieht sich die Zitation auf den Namen des Modells, die Entwickler_innen sowie das Jahr der Veröffentlichung.

[3] Zwischenzeitlich entwickelte ich daraufhin einen zweistufigen Ansatz, bei welchem mittels *Optical Character Recognition* Text aus den Werbeprospekten extrahiert wurde, anschließend mittels eines Text-LLMs (*qwen3:4b* von Cloud (2025)) eine Vorauswahl hinsichtlich Alkoholproduktpräsentation getroffen wurde, um somit die Anzahl der nötigen Bildannotationen zu verringern. Da jedoch durch die textbasierte Vorauswahl viele Vorteile und Ziele einer bildbasierten Analyse verloren gehen, verwarf ich diesen Ansatz nach einiger Zeit.

Probleme waren Überschreitungen der maximalen Antwortzeit (Timeout von mir festgelegt auf maximal elf Minuten pro Katalogseite) sowie unzufriedenstellende Outputausgaben hinsichtlich Struktur und/oder Inhalt. Bereits einfache, explorative Anfragen an die Modelle in der Ollama-Terminalausführung deuteten darauf hin, dass die Modelle große Schwierigkeiten hatten, die Annotationsaufgaben sinnvoll zu bearbeiten.

Die Performanz der Modelle auf meinem Laptop erwies sich als derart ungenügend, dass es sich sogar als Herausforderung erwies, die Modellperformanz mit den Metriken zu evaluieren. Insbesondere die langen Rechenzeiten bis zum Erreichen des Timeouts deuteten für alle Modelle hinweg auf eine hardwareseitige Überforderung hin. In Abbildung 1 sind für die wenigen annotierten Seiten die Ergebnisse im Vergleich zum *Human Goldstandard* aufgeführt, die ich im Rahmen mehrtägigen Ausprobierens erzielen konnte. Die Aussagekraft dieser Tabelle ist jedoch insgesamt gering, da die Fallzahlen für beide Modelle (*Qwen* und *LLaVA*) mit jeweils weniger als zehn Vergleichsseiten keine systematische Bewertung erlauben. Die *Kappa*-Werte als auch die Werte des *F1-Score* des Modells *Qwen* deuten zwar auf eine leicht bessere Performanz hin, letztlich ist aber hinsichtlich der Metriken keine valide Aussage zu treffen. Somit lässt sich zumindest im Rahmen dieser Arbeit feststellen, dass der offen-lokale Ansatz an den Grenzen der praktischen Durchführbarkeit gescheitert ist.

Tabelle 1: Ansatz 1 (offen-lokal): Vergleich von Metriken zwischen Qwen und LLaVA

Variable	**κ (Cohen's Kappa)**		**F1-Macro**		**n**	
	Qwen	**LLaVA**	**Qwen**	**LLaVA**	**Qwen**	**LLaVA**
alc	0.000	−0.125	0.250	0.238	6	9
product	0.684	0.000	0.500	0.167	6	9
reduc	1.000	−0.154	1.000	0.416	6	9

Anmerkung: Bitte geringe Datenbasis (weniger als zehn annotierte Seiten je Modell) beachten, welche auf die erläuterten Herausforderungen zurückzuführen ist.
Verwendete Modelle: *Qwen* = qwen2.5vl:3b (Bai et al., 2025); *LLaVA* = LLaVA:7b (Liu et al., 2023).
Interpretation: *alc*: Sind alkoholische Produkte abgebildet, ja oder nein?; *product:* Wenn ja, welche Art von alkoholischen Produkte ist abgebildet (1 bis 4); *reduc:* Sind auf der Seite preisreduzierte Produkte abgebildet, ja oder nein? Weitere Details siehe Codebook (S. 27).

4.1.2 Ansatz 2: offen-cloudbasiert

Der offen-cloudbasierte Ansatz scheiterte nicht an der Durchführung, sondern an den Ergebnissen. Die Ergebnisse des verwendeten Modells *llama3.2-vision:11b-instruct-fp16* entsprechen, wie in den Abbildungen 2 und 3 nicht den gewünschten Gütekriterien. Ferner bestand eine weitere Herausforderung dieses Ansatzes in langen Rechenzeiten. Trotz der in Abschnitt 3.2.3 beschriebenen stark erhöhten Rechenleistung betrug die Antwortzeit des Modells noch circa sechs Minuten pro Katalogseite, wodurch die Verarbeitung

des Testdatensatzes (n=150) trotz *Colab Pro*-Lizenz annähernd 15 Stunden betrug. Im Falle einer Annotierung der eigentlichen 1562 Katalogseiten, wäre somit eine erhebliche Bearbeitungszeit zu erwarten gewesen.

Mit Blick auf die Ergebnisse in den Abbildungen 2 und 3 offenbaren sich erhebliche Schwächen. Bei vier der fünf kategorialen Variablen (*alc, product, warning, reduc*) liegt *Cohens Kappa* im negativen Bereich. Entsprechend der Interpretation nach Landis and Koch lassen sich Werte dieser Art als 'schlecht' beschreiben, da eine zufällige Zuordnung ein besseres Ergebnis erwarten ließen und das Modell somit systematische Fehlentscheidungen zu treffen scheint (Landis and Koch, 1977, S. 165). Im Kontrast dazu stehen die Werte des gewichteten *F1-Score*, welcher für die Variablen *product*, *reduc* und *child* gute Werte erzielt. Bei Kombination beider Metriken ist jedoch die Variable *child* die einzige, welche einen *moderaten* (Landis and Koch, 1977, S. 165) *Kappa*-Wert bei gleichzeitig gutem *F1*-Wert eine akzeptable Leistung erzielt.

Tabelle 2: Ansatz 2 (offen-cloudbasiert): Evaluationsergebnisse der kategorialen Variablen

Variable	**Cohen's Kappa**	**Weighted F1-Score**	**n**
alc	-0.6494	0.1443	149
product	-0.5964	0.0704	149
warning	-0.3904	0.3134	149
reduc	-0.0963	0.7169	149
child	0.4550	0.7364	149

Labelerklärungen: siehe Codebook (S. 27).

Tabelle 3: Ansatz 2 (offen-cloudbasiert): Evaluationsergebnisse der metrischen Variablen

Variable	**MAE**	**RMSE**	**n**
prod_pp	7.3592	21.3668	142
prod_alc	3.0134	3.2243	149

MAE: Mean Absolute Error; RMSE: Root Mean Squared Error.
Labelerklärungen: siehe Codebook (S. 27).

Die Ergebnisse der für ein MM erwartbar komplizierten 'Zählaufgabe' zur Erstellung der metrischen Variablen *prod_pp* und *prod_alc* offenbaren ebenfalls Schwächen. Das Modell hatte große Schwierigkeiten die Anzahl der Gesamtprodukte pro Seite zu ermitteln. Hier möchte ich auf die sehr klare Definition, was als *Produkt* zu zählen ist und was nicht (siehe Prompt im Anhang A.2) hinweisen, wodurch 'Missverständnisse' ausgeschlossen werden sollten. Viele Katalogseiten hatten weniger als zehn Produkte pro Seite abgebildet, wodurch der *MAE*-Wert von 7,4 nicht akzeptabel ist. Der fast dreimal so hohe *RMSE*-Wert deutet auf starke Ausreißer und somit signifikanter 'Verständnisprobleme' des Modells hin. Die solideren Werte für die Variable *prod_alc* sind nicht überraschend,

da hier in den meisten Fällen *keine alkoholischen Produkte vorhanden*, und somit der Wert = *0* war.

4.1.3 Ansatz 3: proprietär

Nicht nur im Vergleich zu den Ergebnissen der Ansätze 1 und 2, sondern auch ohne eine derartige Relativierung sind die Ergebnisse des proprietären Ansatzes als äußerst bemerkenswert, wenn nicht sogar herausragend zu bewerten. Für alle kategorialen Variablen sind exzellente *Kappa*- und *F1*-Werte zu verzeichnen (siehe Tabelle 4). Ebenso war das Modell in der Lage, die Produktzählungen höchst zufriedenstellend zu erfüllen. Für beide metrischen Variablen *prod_pp* und *prod_alc* liegen *MAE*- und *RMSE*-Werte vor, die als absolut vertretbar einzustufen sind, da im Falle mehrerer menschlicher Kodierer ebenfalls mit leichten Abweichung zu rechnen gewesen wäre (Stichwort: *Inter-Coder-Reliabilität*). Selbst für die Kategorie *child*, welche im Vergleich zu anderen Labeln vergleichsweise offen definiert war (siehe Anhang A.2), liegen hohe Übereinstimmungsraten vor.

Tabelle 4: Ansatz 3 (proprietär): Evaluationsergebnisse der kategorialen Variablen

Variable	**Cohen's Kappa**	**Weighted F1-Score**	**n**
alc	0.9676	0.9934	149
product	0.9379	0.9859	149
warning	1.0000	1.0000	149
reduc	0.6831	0.8641	149
child	0.6160	0.8255	149

Labelerklärungen: siehe Codebook (S. 27).

Tabelle 5: Ansatz 3 (proprietär): Evaluationsergebnisse der metrischen Variablen

Variable	**MAE**	**RMSE**	**n**
prod_pp	1.3778	2.0111	135
prod_alc	0.0604	0.3378	149

MAE: Mean Absolute Error; RMSE: Root Mean Squared Error.
Labelerklärungen: siehe Codebook (S. 27).

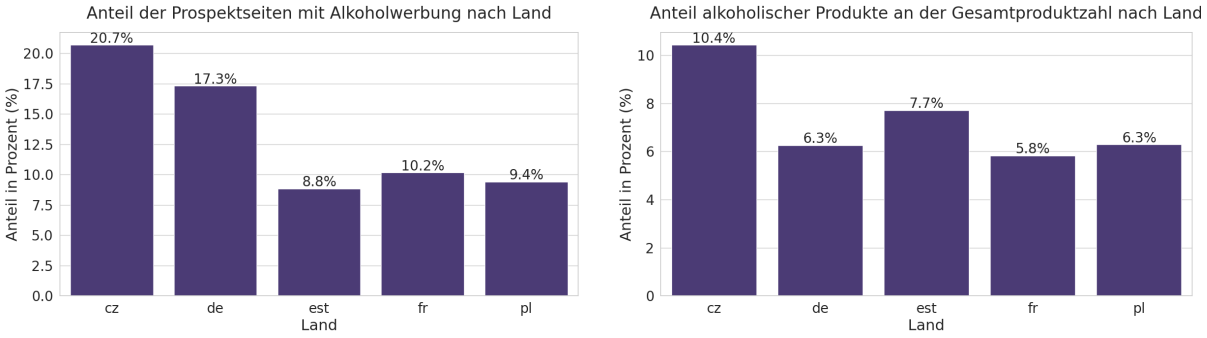
Ebenso bemerkenswert ist die herausragend schnelle Bearbeitung der Aufgaben, sowie die überaus geringen Nutzungskosten. Die entstandenen Kosten durch die API-Nutzung, welche wenige Euro betrugen[4], ermöglichen selbst bei geringen Forschungsbudgets preisgünstige Skalierung in fast beliebige Höhe. Die Rechenzeit für alle 1562 Katalogseiten lag bei circa einer Stunde, was im Vergleich zu Ansatz 2 eine etwa hundertfünzigfach schnellere Bearbeitungsgeschwindigkeit bei vielfach besseren Ergebnissen zu bedeuten hat.

[4]Insgesamt wurden mindestens 2000 PDFs annotiert. Dabei sind lediglich Kosten in Höhe von circa 2,30 Euro angefallen.

Gleichzeitig sind, wie bereits erläutert, Abstriche hinsichtlich wissenschaftlicher Gütekriterien, Datenautonomie und Kontrolle über den Forschungsprozess zu beachten. Die Abhängigkeit von privatwirtschaftlichen Firmenentscheidungen war im Prozess dieser Arbeit direkt festzustellen. So hatte ich erfolgreiche Pre-Tests mit dem Modell *gemini-1.5-pro* abgeschlossen, nur um zehn Tage später, als ich den Gesamtdatensatz annotieren lassen wollte, festzustellen, dass das Modell nicht mehr zur Verfügung steht. Daraufhin musste ich mir Informationen zu alternativen Modellen einholen und neue Pre-Tests durchführen, woraufhin ich das Modell *gemini-2.0-flash* genutzt habe. Somit ist die Nutzung, falls es die Forschungsdaten überhaupt zu lassen, aufgrund der Abhängigkeit von Konzernentscheidungen gut zu überlegen. Meine Detailergebnisse der Evaluierung des Modells *gemini-2.0-flash* sind in Tabelle 8 im Anhang zu finden.

4.2 Ergebnisse der Fallstudie: Alkoholwerbung in europäischen Supermarktkatalogen

Die vorliegenden Daten deuten darauf hin, dass es klare Unterschiede in der alkoholischen Produktpräsentation zwischen den untersuchten europäischen Staaten gibt. Wie in **Abbildung 2a** abgebildet, weisen tschechische Supermärkte mit 20,7 Prozent und deutsche Supermärkte mit 17,3 Prozent deutlich höhere Anteile an Prospektseiten auf, die Alkoholprodukte präsentieren, als Frankreich (10,2 Prozent), Polen (9,4 Prozent) und Estland (8,8 Prozent). Die Detailanalyse, welche die Anzahl alkoholischer Produkte in Relation zur Gesamtproduktzahl setzt, weist jedoch weniger starke Länderunterschiede auf. So enthalten in allen fünf untersuchten Ländern zwischen 6,3 Prozent und 10,4 Prozent der Gesamtprodukte Alkohol (siehe **Abbildung 2b**).



(a) Anteil der Prospektseiten mit Alkoholwerbung.

(b) Anteil alkoholischer Produkte an der Gesamtproduktzahl.

Abbildung 2: Quantitativer Anteil der Alkoholwerbung in den Prospekten nach Land. Für Ländercodes siehe S. 6.

Alkoholprodukte nehmen innerhalb der Ländergruppe außerdem einen unterschiedlichen Stellenwert ein. Supermärkte aus Polen, Frankreich, Tschechien und Deutschland

verteilen ihre Alkoholproduktpräsentation eher auf die mittleren Katalogseiten (siehe **Abbildung 4**). Hier sticht Tschechien heraus, als das Land in dem im Ländergruppenvergleich die meisten Alkoholprodukte schon auf der Katalogtitelseite oder dem ersten Zehntel der Werbebroschüre (19,6 Prozent) präsentiert werden. Im Gegensatz dazu wählten estnische Supermärkte für die untersuchten Kataloge für 58,7 Prozent der Alkoholproduktpräsentationen das hintere Fünftel der Katalogseiten (siehe **Abbildung 3**).

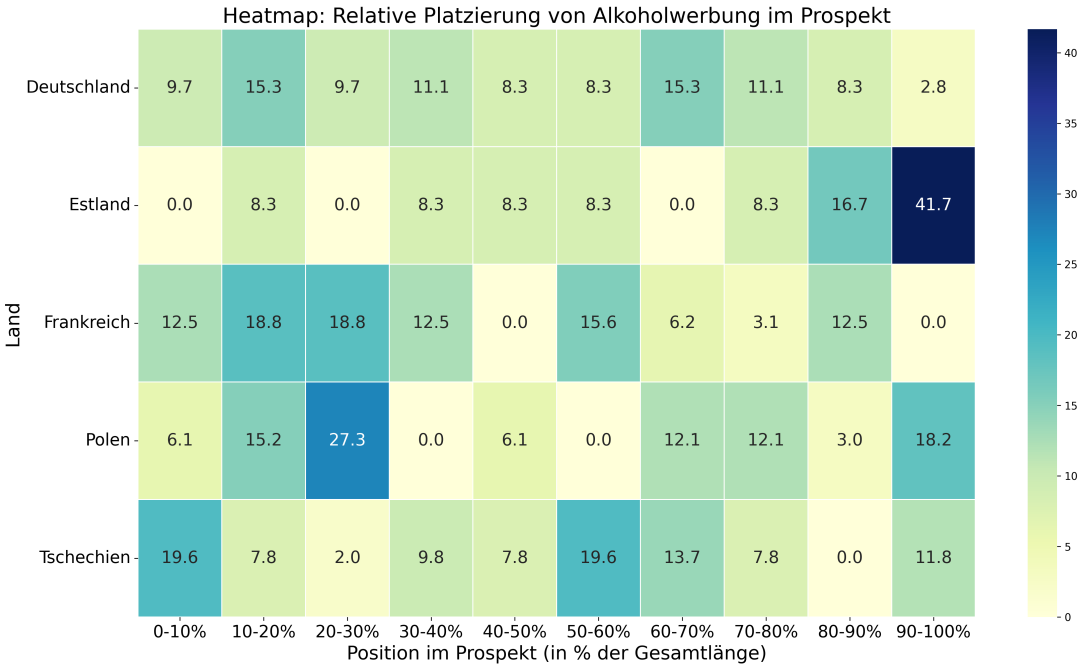


Abbildung 3: Heatmap der relativen Platzierung von Alkoholwerbung im Prospekt nach Land. Die x-Achse zeigt die relative Position im Prospekt von Anfang (0%) bis Ende (100%).

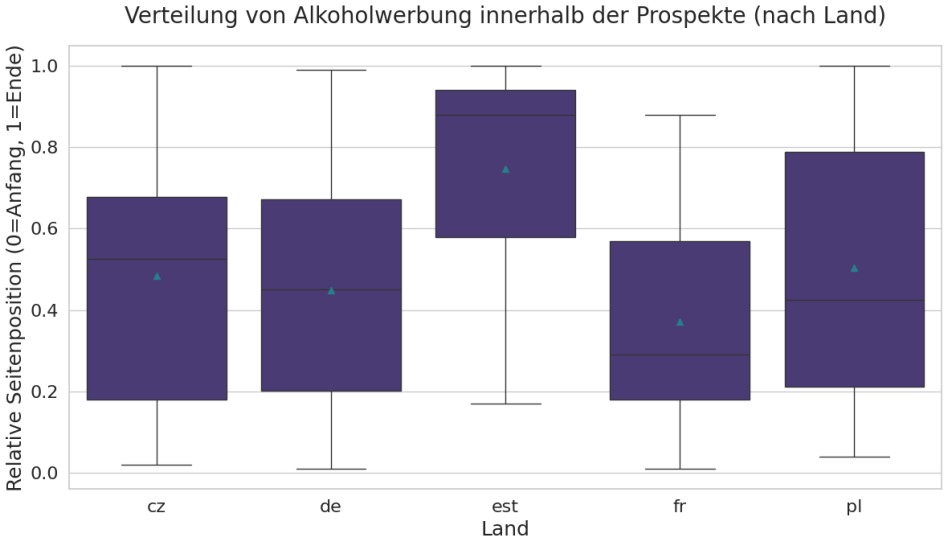


Abbildung 4: Verteilung der Alkoholwerbung innerhalb der Prospekte (0=Anfang, 1=Ende) als Boxplot dargestellt. Für Ländercodes siehe S. 6.

Spezielle Kennzeichungen alkoholischer Produkte sind insbesondere in Polen, Frankreich und Estland verbreitet. In polnischen Prospekten identifizierte das MM in 97,1 Prozent der Alkoholpräsentationen einen darauf bezogenen Warnhinweis. Für Frankreich identifizierte das MM für 87,9 Prozent der Alkoholprodukte einen entsprechenden Warnhinweis. Auch in Estland scheinen Warnhinweise üblich zu sein. Hier konnten für 53,8 Prozent der Alkoholwerbungen Warnhinweise festgestellt werden. In deutschen und tschechischen Werbeprospekten sind trotz überdurchschnittlich vielen Seiten mit alkoholischen Produkten (Abbildung 2) keine oder kaum Warnhinweise zu identifizieren (siehe **Abbildung 5**).

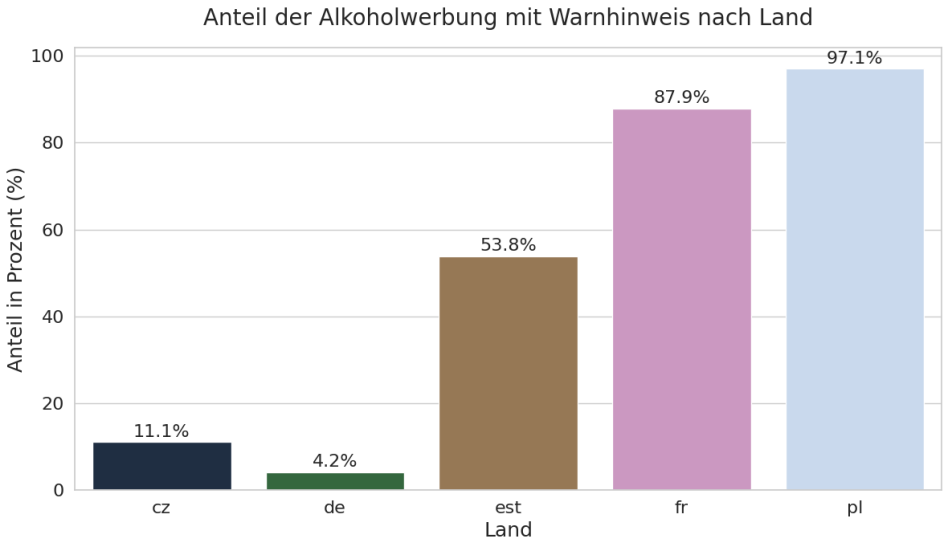


Abbildung 5: Anteil der Alkoholwerbung mit Warnhinweis nach Land. Für Ländercodes siehe S. 6.

Trotz augenscheinlicher Bemühungen in Polen, Frankreich und Estland alkoholische Produkte durch Warnhinweise inhaltlich von anderen Produkten abzugrenzen, ergibt sich bei der Trennung von alkoholischen Produkten und Produkten, die für Kinder und Jugendliche von Interesse sein könnten, eine recht hohe Nähe. So wurden alkoholische Produkte über die fünf untersuchten Länder hinweg regelmäßig neben Produkten wie Chips, Süßigkeiten, Säften, Softdrinks und Produkten mit 'kindlicher Aufmachung' (für Details hinsichtlich der Annotierung siehe Absatz 3, sowie Anhang A.2) präsentiert. Vermischungen von Produkten mit potenziell unterschiedlicher Zielaltersgruppe finden in Deutschland am häufigsten statt. In annähernd 40 Prozent der Fälle, in denen eine deutsche Katalogseite ein oder mehrere alkoholische Produkte enthielt, befand sich auf der selben Katalogseite ein oder mehrere Produkte, welche für die Zielgruppe *Kinder/Jugendliche* von Interesse sein könnte. Bezieht man direkt benachbarte Seiten in die Betrachtung mit ein, zeigt sich, dass in mehr als 90 Prozent der Fälle Produkte mit *Kinder-* oder *Jugend*bezug in direktem Kontext von Alkoholwerbung stehen. Für die Supermärkte der restlichen Länder, insbesondere in Frankreich, Estland und Tschechien sind ebensolche

Vermischungen ebenfalls festzustellen, wenn auch in leicht geringerer Ausprägung. Lediglich für polnische Supermärkte scheint eine stärkere räumliche Trennung vorzuliegen. Hier gibt es in circa 10 Prozent der Alkoholproduktseiten einen *Kinder-* oder *Jugend*bezug auf der selben Seite und in insgesamt circa 50 Prozent der Fälle auf der selben oder direkt benachbarten Seite (siehe **Abbildung 6**).

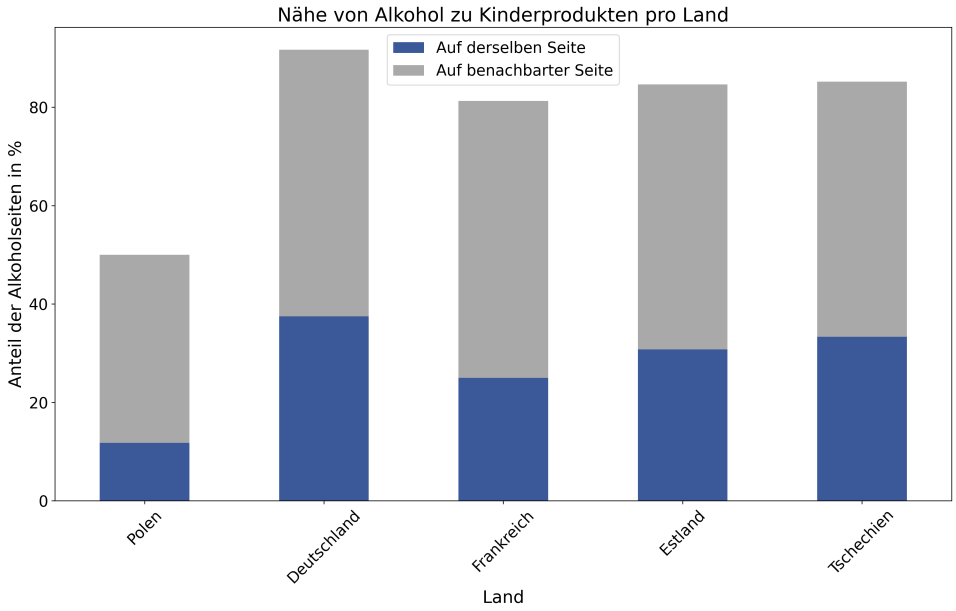


Abbildung 6: Anteil der Alkoholwerbung, die sich auf derselben oder einer benachbarten Seite wie Kinderprodukte befindet.

Durch die internationale Verbreitung einiger Supermärkte wie beispielsweise Lidl, bietet sich unter Annahme einer international-konsistenten Unternehmenskultur durch einen Länder-Fall-Vergleich die Möglichkeit, Länderunterschiede auf unterschiedliche länderbezogene Marketingstrategien zurückzuführen. Da leider (Kommentar Reinhardt: *ausgerechnet*) in Deutschland die Lidl-Werbekataloge nicht zum Download zur Verfügung standen, konnte der Ländervergleich nur für die Länder Estland, Frankreich, Polen und Tschechien durchgeführt werden. Im Vergleich mit Abbildung 2 scheint Lidl Alkohol insgesamt auf unterdurschnittlich vielen Katalogseiten zu bewerben. Die Länderrelationen zueinander bleiben jedoch erhalten: Estland weist den geringsten Anteil an Alkoholwerbeseiten auf. Polen und Frankreich liegen zwar leicht über Estland, jedoch wie Estland im $<$6 Prozent-Bereich. Der tschechische Lidl-Katalog enthält mit einem Anteil von circa 15 Prozent Katalogseiten mit alkoholsichen Produkten den im Ländervergleich höchsten Anteil (siehe **Abbildung 7**).

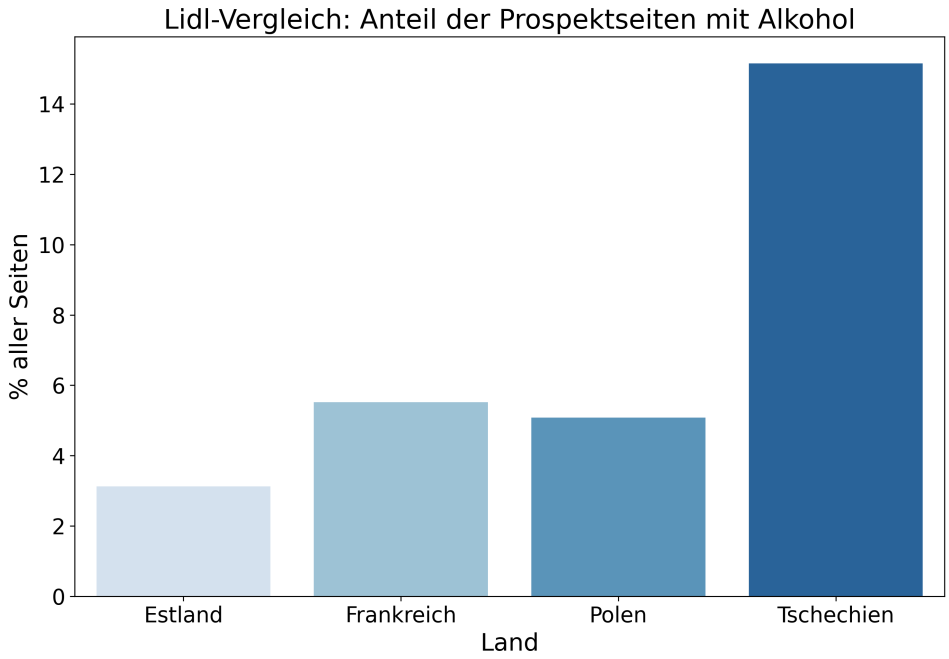


Abbildung 7: Lidl-Vergleich: Anteil der Prospektseiten mit Alkoholwerbung im Ländervergleich.

5 Fazit und Diskussion

Das Ziel dieser Arbeit war es, die forschungspraktischen Erfahrungswerte zur Nutzung von multimodalen Modellen für Bildannotationsaufgaben zu erweitern und die Implikationen verschiedener praktischer Ansätze zu beleuchten. Dies geschah anhand einer soziologischen exemplarischen Fallstudie, bei welcher Alkoholwerbung in europäischen Supermarktkatalogen untersucht wurde. Dieses abschließende Kapitel fasst die zentralen Ergebnisse zusammen, diskutiert diese im Kontext der Forschung und Limitationen der Arbeit und gibt einen Ausblick auf zukünftige Forschungspotentiale.

Die erste, zentrale Forschungsfrage dieser Arbeit untersuchte Implikationen, die sich aus dem Vergleich der Nutzung von offenen und proprietären Modellen mit unterschiedlichen technischen Ausführungen ergeben. Hier zeigte sich ein Spannungsfeld zwischen wissenschaftlichen Ansprüchen und praktischer Umsetzbarkeit auf. Die Ansätze 1 mit offenen Modellen, Ansatz 1 (offen-lokal) und Ansatz 2 (offen-cloudbasiert), gewährleisten zwar ein Höchstmaß an Reproduzierbarkeit und bei passender Cloud-Infrastruktur auch ein Höchstmaß an Datenautonomie. In der praktischen Anwendung scheiterten beide Ansätze im Rahmen dieser Arbeit an technischen Hürden. Der lokale Ansatz war aufgrund mangelnder Performanz nicht durchführbar, während der cloudbasierte Ansatz trotz besserer Hardware unzureichende und teils systematisch fehlerhafte Ergebnisse lieferte. Der proprietäre Ansatz (Ansatz 3) dagegen lieferte herausragende Ergebnisse mit exzellenten Evaluationswerten bei hoher Geschwindigkeit und geringen Kosten. Damit einher gehen

jedoch der Verlust der Datenhoheit, die Abhängigkeit von kommerziellen Anbietern und eine eingeschränkte Reproduzierbarkeit. Passend illustriert wurde dies durch die Abschaltung eines Modells während des Forschungsprozesses. Für die Forschungspraxis deuten die Ergebnisse dieser Arbeit auf ein Dilemma hin. Für komplexe multimodale Analysen scheinen proprietäre Modelle derzeit oft die einzig sinnvoll-nutzbare Lösung zu sein.

Die zweite, kultursoziologische Forschungsfrage befasste sich mit quantitativen und qualitativen Unterschieden in der alkoholischen Produktpräsentation verschiedener europäischer Länder. In den Forschungsdaten, welche aus Werbekatalogen bestanden, zeigten sich klare Länderunterschiede. Die Betrachtung der Anzahl an Prospektseiten mit Alkohol, des Anteils alkoholischer Produkte an der Gesamtzahl sowie der Vergleich einer in verschiedenen Ländern vertretenen Supermarktkette (*Lidl*) zeigt, dass Alkoholwerbung in Tschechien deutlich präsenter ist als in Frankreich, Polen und Estland. Für Deutschland zeigt sich ein differenzierteres Bild, da es bei der Anzahl der alkoholischen Prospektseiten den zweithöchsten Wert, beim Anteil an alkoholischen Produkten zusammen mit Polen jedoch den zweitniedrigsten Wert aufweist. Zudem zeigen sich klare länderspezifische Platzierungsstrategien und eine höchst unterschiedliche Verwendung von gesetzlichen Warnhinweisen, die in Polen und Frankreich Standard sind, in Deutschland und Tschechien jedoch fast gänzlich fehlen. Besonders auffällig ist die mangelnde räumliche Trennung zwischen Alkoholwerbung und Produkten, die Kinder und Jugendliche ansprechen beziehungsweise ansprechen könnten. Diese Vermischung ist in deutschen Katalogen am stärksten ausgeprägt, während polnische Supermärkte eine deutlich striktere Trennung vorzunehmen scheinen.

Während im Bereich der Textannotation klare Vorteile der Nutzung offener Modelle bei gleichzeitiger Praktikabilität nachgewiesen wurden (bspw. Weber and Reichardt, 2023, S. 11), zeigt diese Studie, dass sich diese Erkenntnisse nicht ohne Weiteres auf die Bildannotation übertragen lassen. Die technischen Anforderungen durch die höhere Datenmenge pro Analyseobjekt stellen Hürden dar, die die Nutzung offener Modelle für Forschende ohne Zugang zu spezialisierter Hardware nur schwer überwindbar machen.

Zugleich muss kritisch angemerkt werden, dass viele Ergebnisse der Fallstudie auch mit alternativen methodischen Ansätzen hätten durchgeführt werden können. Mittels *Optical Character Recognition* und anschließender Analyse durch ein Text-LLM hätten Alkoholprodukte, Warnhinweise und Rabattaktionen auf Werbekatalogseiten ebenfalls identifiziert werden können (siehe Fußnote 3). Komplexere und differenziertere Annotationsaufgaben, insbesondere die Zählaufgaben oder die Identifikation von Produkten mit 'kindlicher Aufmachung', wären jedoch mit diesem Ansatz nicht durchführbar gewesen.

Die vorliegende Arbeit unterliegt zudem Limitationen. Die Analyse basiert auf einer relativ kleinen Anzahl von Werbekatalogen aus einem kurzen Zeitraum. Um robustere und verallgemeinerbare Aussagen über nationale Werbestrategien zu treffen oder Veränderungen im Zeitverlauf zu erfassen, müsste die Datenbasis erheblich vergrößert wer-

den. Der Umfang der manuellen Annotationen (n=150) war ausreichend für generelle Modellevaluationen, erlaubte jedoch keine differenzierte Bewertung einer möglicherweise unterschiedlichen Modellperformanz hinsichtlich verschiedener Länder- und Sprachräume. Außerdem war der verwendete Prompt, aufgrund der Annotationsansprüche recht umfassend. Hier könnte die Reduzierung der Komplexität durch die Unterteilung in Einzelaufgaben bessere Ergebnisse erzielen. Ebenso wurde nur ein Ausschnitt der verfügbaren offenen Modelle in den Vergleich miteinbezogen. So stehen auch neuere und größere offene Modelle, etwa *Llama 4* (Meta Platforms, 2025) zur Verfügung, welche im Rahmen dieser Arbeit aufgrund der technischen Ausstattung nicht berücksichtigt werden konnten. Ferner stellt die Arbeit angesichts rasanter technischer Entwicklungen im Bereich der generativen Sprachmodelle allgemein, jedoch auch im Bereich der offenen Modelle eine aktuelle Bestandsaufnahme dar, welche periodisch zu evaluieren ist. Ebenfalls bieten multimodale Modelle Potenziale zur Analyse anderer Datenformate, etwa Audio- und Videodaten, welche, unter Berücksichtigung ethischer Gesichtspunkte, für den Bereich der Methodenforschung in den Computational Social Science weitere Forschungsfelder eröffnen.

Literatur

Alizadeh, M., Kubli, M., Samei, Z., Dehghani, S., Bermeo, J. D., Korobeynikova, M., and Gilardi, F. (2023). Open-source large language models outperform crowd workers and approach ChatGPT in text-annotation tasks. *arXiv preprint arXiv:2307.02179*, 101.

Alizadeh, M., Kubli, M., Samei, Z., Dehghani, S., Zahedivafa, M., Bermeo, J. D., Korobeynikova, M., and Gilardi, F. (2025). Open-source LLMs for text annotation: a practical guide for model setting and fine-tuning. *Journal of Computational Social Science*, 8(1):17.

Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al. (2025). Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Cloud, A. (2025). Qwen3:4b. Großes Sprachmodell. Abgerufen von Ollama.

Cluley, R. and Nixon, E. (2019). What is an advert? A sociological perspective on marketing media. *Marketing Theory*, 19(4):405–423.

Gilardi, F., Alizadeh, M., and Kubli, M. (2023). Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.

Grimmer, J., Roberts, M. E., and Stewart, B. M. (2022). *Text as data: A new framework for machine learning and the social sciences*. Princeton University Press.

Gunes, E. and Florczak, C. K. (2023). Multiclass classification of policy documents with large language models. *arXiv preprint arXiv:2310.08167*.

Hall, G. and Kappel, R. (2018). Gender, alcohol, and the media: The portrayal of men and women in alcohol commercials. *The Sociological Quarterly*, 59(4):571–583.

Hwang, J. and Naik, N. (2023). Systematic social observation at scale: Using crowdsourcing and computer vision to measure visible neighborhood conditions. *Sociological Methodology*, 53(2):183–216.

Isański, J. and Leszkowicz, M. (2011). "Keeping up with the Joneses." A sociological content analysis of advertising catalogues with the eye-tracking method. *Qualitative Sociology Review*, 7(2):85–100.

Johnston, R., Stafford, J., Pierce, H., and Daube, M. (2017). Alcohol promotions in Australian supermarket catalogues. *Drug and alcohol review*, 36(4):456–463.

Kim, J. H., Ki, D., Osutei, N., Lee, S., and Hipp, J. R. (2024). Beyond visual inspection: capturing neighborhood dynamics with historical Google Street View and deep learning-based semantic segmentation. *Journal of Geographical Systems*, 26(4):541–564.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Law, T. and Roberto, E. (2025). Generative Multimodal Models for Social Science: An Application with Satellite and Streetscape Imagery. *Sociological Methods & Research*, page 00491241251339673.

Liu, H., Li, C., Wu, Q., and Lee, Y. J. (2023). Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

Livingston, M. (2013). To reduce alcohol-related harm we need to look beyond pubs and nightclubs. *Drug & Alcohol Review*, 32(2).

Meta Platforms, I. (2024). Llama 3.2-vision. Großes Sprachmodell. Abgerufen von Ollama.

Meta Platforms, I. (2025). Introducing llama 4: Advancing multimodal intelligence.

Ollion, E., Shen, R., Macanovic, A., and Chatelain, A. (2023). ChatGPT for text annotation? Mind the hype. *SocArXiv preprint*, page 32.

OpenAI (2024). GPT-4o. Großes Sprachmodell.

Sarmadi, H., Hall, O., Rögnvaldsson, T., and Ohlsson, M. (2025). Leveraging ChatGPT's Multimodal Vision Capabilities to Rank Satellite Images by Poverty Level: Advancing Tools for Social Science Research. *arXiv preprint arXiv:2501.14546*.

Schaller, K., Kahnert, S., Garcia-Verdugo, R., Treede, I., Graen, L., and Ouédraogo, N. (2022). *Alkoholatlas Deutschland 2022*. Pabst Science Publishers, Lengerich.

Sheng, H., Yao, K., and Goel, S. (2021). Surveilling surveillance: Estimating the prevalence of surveillance cameras with street view data. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 221–230.

Spirling, A. (2023). Why open-source generative AI models are an ethical way forward for science. *Nature*, 616(7957):413–413.

Statistisches Bundesamt (2008). *Klassifikation der Wirtschaftszweige mit Erläuterungen*. Statistisches Bundesamt, Wiesbaden. Erschienen im Dezember 2008.

Stevenson, T. H. (2002). The portrayal of african-americans in business-to-business catalog advertising. *Journal of Current Issues & Research in Advertising*, 24(2):41–49.

Tan, P. J., Tanusondjaja, A., Corsi, A., Lockshin, L., Villani, C., and Bogomolova, S. (2023). Audit and benchmarking of supermarket catalog composition in five countries. *International Journal of Advertising*, 42(3):589–616.

Torres, M. and Cantú, F. (2022). Learning to see: Convolutional neural networks for the analysis of social science data. *Political Analysis*, 30(1):113–131.

Vikhyat (2024). Moondream2: A tiny vision language model. `https://huggingface.co/vikhyatk/moondream2`.

Weber, M. and Reichardt, M. (2023). Evaluation is all you need. prompting generative large language models for annotation tasks in the social sciences. a primer using open models. *arXiv preprint arXiv:2401.00284*.

White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., and Schmidt, D. C. (2023). A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:2302.11382*.

Williams, N. W., Casas, A., and Wilkerson, J. D. (2020). *Images as data for social science research: An introduction to convolutional neural nets for image classification*. Cambridge University Press.

World Health Organization (2024). *Global status report on alcohol and health and treatment of substance use disorders*. World Health Organization.

Yue, X., Ni, Y., Zhang, K., Zheng, T., Liu, R., Zhang, G., Stevens, S., Jiang, D., Ren, W., Sun, Y., et al. (2024). Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567.

A Anhang

A.1 Codebuch

Tabelle 6: Codebook

Aspekt	**Variable**	**Codierung**
Alkoholische Getränke abgebildet?	alc	0 = nein 1 = ja
Art?	product	0 = nicht zutreffend 1 = Bier & Biermischgetränke 2 = Wein & Weinmischgetränke 3 = Spirituosen & Spirituosenmischgetränke 4 = mehrere Klassen oder andere alkoholische Produkte
Warnhinweis?	warning	0 = kein Warnhinweis in Bezug auf Alkohol 1 = Warnhinweis
Rabatt?	reduc	0 = kein Rabatt 1 = Rabattaktion 99 = unklar
Produkte mit Zielgruppe Kind/Jugendlich?	child	0 = nein 1 = ja 99 = unklar
Produkte pro Seite?	prod_pp	int (Anzahl Produkte); 99 = keine Angabe / sonstiges
Anzahl prod_pp mit Alkohol?	prod_alc	int (Anzahl Produkte); 99 = keine Angabe / sonstiges
Prospekt fehlerhaft?	jede Variable	= 98

A.2 Verwendeter, finaler Prompt

You are a scientific research assistant. Your task is to meticulously examine the provided image of a supermarket brochure page and classify its content based on a strict codebook. IMPORTANT: The images can be in any European language (e.g., German, French, English, Polish).

Follow these steps precisely to determine the values for the final JSON object.

1. Faulty Brochure (‘fehlerhaft‘):

- ***First, assess if the brochure page is faulty, illegible, or completely unrelated (e.g., a blank page, a picture of a car).***

- *If YES, assign the value '98' to* **every** *variable and stop here.*

2. **Alcohol Presence ('alc'):**

   - *Determine if any alcoholic beverages (beer, wine, spirits, etc.) are visible.*
   - *If NO, the value is '0'.*
   - *If YES, the value is '1'.*

3. **Product Type ('product'):**

   - *If 'alc' is '0', the value for 'product' is '0'.*
   - *If 'alc' is '1', identify the category of alcoholic beverages shown:*
     - *Value '1': Only beer and/or beer-mixes.*
     - *Value '2': Only wine and/or sparkling wine/wine-mixes.*
     - *Value '3': Only spirits and/or spirit-mixes.*
     - *Value '4': Multiple categories from above or other alcoholic products (e.g., cider, alcopops).*

4. **Alcohol Warning ('warning'):**

   - *If 'alc' is '1', check for any legal or health warnings related to alcohol consumption (e.g., "l'abus d'alcool est dangereux pour la santé", "Drink Responsibly", "Kein Verkauf von Alkohol an Jugendliche").*
   - *If NO alcohol-related warning is present, the value is '0'.*
   - *If a warning is present, the value is '1'.*
   - *If 'alc' is '0', this value must also be '0'.*

5. **Price Reduction ('reduc'):**

   - *Examine all products. Look for any indication of a price reduction (e.g., 25%", SSale", "Rabatt", "Aktion").*
   - *If NO discounts are visible, the value is '0'.*
   - *If ANY price reduction is visible, the value is '1'.*
   - *If you cannot clearly determine whether a discount is present or not (e.g., due to image quality), the value is '99'.*

6. **Products Targeting Children/Adolescents ('child'):**

   - *Examine the products to see if any are specifically aimed at children or adolescents, or could attract their attention. This includes items like sweets, icecream, sugary cereals, juices, lemonade, toys, ice tea, specific snack products with cartoon characters, etc.*

- *If NO such products are visible, the value is ‘0‘.*
- *If YES, one or more such products are visible, the value is ‘1‘.*
- *If it is unclear whether a product targets this group, the value is ‘99‘.*

7. **Total Products per Page (‘prod_pp‘):**
   - *Count every distinct product advertised on the page. You can use the number of price quotations as a guide. A ”product” is one or several items for sale (e.g., two bottles of Cola (one of them Cola Zero and one of them Cola Light), a specific type of cheese, a bag of apples) while connected to a price. Different flavors or variations of the same item count only as distinct products if priced separately.*
   - *The value is the final integer count.*
   - *If you cannot count the products for any reason, or there are no distinct prices on the page the value is ‘99‘.*
8. **Alcoholic Products per Page (‘prod_alc‘):**
   - *Count every distinct alcoholic product advertised on the page, following the same counting rule as for ‘prod_pp‘.*
   - *If ‘alc‘ is ‘0‘, this value must be ‘0‘.*
   - *The value is the final integer count of alcoholic products.*
   - *If you cannot count the alcoholic products (but they are present), the value is ‘99‘.*

FINAL INSTRUCTION:
You MUST provide your response as a single, compact JSON object. Do not add any explanations or markdown. The JSON object must match this exact structure and contain only these keys with integer values:

```
{
  "alc": <integer>,
  "product": <integer>,
  "warning": <integer>,
  "reduc": <integer>,
  "child": <integer>,
  "prod_pp": <integer>,
  "prod_alc": <integer>
}
```

A.3 Dataset

Tabelle 7: Übersicht der Supermarktkataloge mit Seitenanzahl

Supermarkt	Land	Kalenderwoche	Seitenanzahl
albert	Tschechien	38	47
albert	Tschechien	37	25
coop	Tschechien	37	10
coop	Tschechien	35	10
globus	Tschechien	38	39
lidl	Tschechien	38	53
lidl	Tschechien	38	51
tesco	Tschechien	38	33
aldi nord	Deutschland	33	47
aldi nord	Deutschland	34	45
aldi sued	Deutschland	33	43
aldi sued	Deutschland	34	44
aldi sued	Deutschland	35	34
edeka	Deutschland	33	35
netto	Deutschland	34	67
netto	Deutschland	33	24
norma	Deutschland	34	18
norma	Deutschland	35	18
rewe	Deutschland	33	30
rewe	Deutschland	34	26
a1000	Estland	38	12
konsum	Estland	38	25
lidl	Estland	38	66
maxima	Estland	38	32
promo	Estland	33	16
auchan	Frankreich	34	36
auchan	Frankreich	34	46
intermarche	Frankreich	32	32
intermarche	Frankreich	33	40
leclerc	Frankreich	33	0
lidl	Frankreich	37	76
lidl	Frankreich	38	76
superu	Frankreich	33	24
aldi	Polen	33	40
aldi	Polen	33	15
groszek	Polen	38	61
lidl	Polen	38	66
lidl	Polen	38	61
stokrotka	Polen	38	49
zabka	Polen	37	78

A.4 Detailevaluation

Tabelle 8: Ansatz 3 (proprietär): Detaillierte Evaluationsergebnisse der kategorialen Variablen

Variable	Klasse	Precision	Recall	F1-Score	Support (n)
alc	0	1.00	0.99	1.00	132
	1	0.94	1.00	0.97	17
product	0	1.00	0.99	1.00	132
	1	0.83	1.00	0.91	5
	2	1.00	0.50	0.67	2
	3	1.00	1.00	1.00	2
	4	0.89	1.00	0.94	8
warning	0	1.00	1.00	1.00	143
	1	1.00	1.00	1.00	6
reduc	0	0.97	0.63	0.77	49
	1	0.85	0.99	0.91	100
child	0	0.87	0.87	0.87	97
	1	0.75	0.75	0.75	52

Labelerklärungen: siehe Codebook (S. 27).