Bindung und Selektion

Eine Zwei-Mechanismen-Architektur für ein offenes Sprachsystem auf bescheidener Hardware: ein kleiner Transformer, der binden lernt, und ein selektiv vergessendes Gedächtnis, das trägt, was es weiß

Marcel Langjahr

Pantareon GmbH
https://pantareon.io/
langjahr@pantareon.io

Juni 2026

Zusammenfassung

Wir trennen zwei Fähigkeiten, die ein Sprachmodell vermengt und die ein zählbasiertes Substrat nicht beide bereitstellen kann. Die erste ist *sequenzielle Bindung*: Inhalt über Distanz verbinden—Subjekt mit Verb, Pronomen mit Referent, ein narrativer Faden über einen Absatz hinweg. Die zweite ist die offene *Wissensselektion*: behalten, was zu behalten sich lohnt, aus einem Strom, der nie endet, ohne unbegrenztes Wachstum. Eine Begleitstudie zeigt, indem sie die Architektur fixiert und nur das Gedächtnis-Regime variiert, dass ein Kontextgraph variabler Ordnung die lokale Statistik perfektioniert und dennoch eine Decke genau bei der Bindung trifft—bei 11.3M Zeichen erreicht er 1.83 held-out-Bits-pro-Zeichen (bpc) und gibt orthographisch perfekten, lokal flüssigen Text ohne langreichweitige Kohärenz aus. Wir argumentieren, dass diese Decke zwei orthogonale Seiten hat, dass das Gehirn sie mit zwei verschiedenen *Arten* von Mechanismus löst statt mit einer verfeinerten Repräsentation, und wir schlagen eine Architektur vor, die dasselbe tut: Ein kleiner Decoder-Transformer (der *Binder*, engl. „binder"), von Grund auf trainiert auf einem kuratierten, kapazitäts-angepassten Korpus, lernt die Bindungsmechanik in einem Fenster nach Art einer kritischen Periode; ein selektiv vergessender Zählgraph (der *Zähler*, engl. „counter") trägt offenes Wissen als Datastore mit den eigenen Repräsentationen des Binders als Schlüssel, per Retrieval zurückgelesen. Vergessen ist nicht Aufräumen, sondern Metabolismus: In einem System, dessen Repräsentationen driften, ist selektives Vergessen der einzige Weg, auf dem das Gedächtnis ausgerichtet und endlich bleibt, während alles andere sich bewegt. Dies ist ein Design- und Hypothesen-Paper. Wir sagen genau, was bereits etabliert ist (die Decke des Substrats; Vergessen als Regularisierer der Stichprobeneffizienz; ein gradientenverifizierter, von Grund auf gebauter Binder), was hypothetisch ist (der Wert der Komposition; Vergessen als Selbstheilung gegen Repräsentationsdrift) und was offen ist (die Driftrate; kontinuierlicher Kompetenzerwerb). Die empirischen Abschnitte tragen explizite Platzhalter; ihre Ergebnisse stehen auf stärkerer Hardware aus und werden in dieses Dokument eingearbeitet, sobald sie eintreffen.

Geltungsbereich dieses Papers: Dies ist ein Design-Paper, kein Ergebnis-Paper. Sein Zweck ist, eine Architektur zu modellieren und sie aus gemessenen Vorläufern und Vorarbeiten zu begründen, und im Voraus die Experimente festzulegen, die sie bestätigen oder falsifizieren würden. Ein Artefakt existiert und ist verifiziert: ein abhängigkeitsfreier, von Grund auf gebauter Decoder-Transformer, dessen handgeschriebener Backward-Pass eine Gradientenprüfung mit zentralen Differenzen besteht (maximaler relativer Fehler 4.8×10^{-5} über alle Parameter-Tensoren). Eine Kontrolle ist gemessen: die Zähl-Modell-Baseline auf demselben Korpus auf TinyStories (1.306 prequentielle Bits/Zeichen; lokal flüssige, narrativ inkohärente Generierung). Alles andere, das unten als *hypothetisch* oder *offen* markiert ist, ist genau das. Das angestrebte Regime ist bewusst bescheiden: eigene Gewichte, kein geerbtes vortrainiertes Modell, kein Rechenzentrum, ein kuratierter Korpus statt des offenen Webs und eine einzige Workstation. Absolute Zahlen sind relativ zu den angegebenen Baselines zu lesen, wie in der Begleitarbeit. Der epistemische Status jeder Behauptung ist in Tabelle 1 zusammengefasst, die zugleich als Gerüst für die spätere Überarbeitung dient: Wenn Läufe abgeschlossen sind, werden aus *offenen* Zeilen *etablierte* Zeilen.

1 Einleitung

Das Substrat hinter dieser Arbeit stammt aus der Emergenten Graphentheorie: Funktion, die von einem diskreten Graphen getragen wird, der unter einer Erhaltungsgröße aus Daten wächst, statt von einem festen parametrischen Modell, das durch globale Optimierung angepasst wird. Realisiert als zeichenbasiertes Sprachmodell, ist dies ein Kontextgraph variabler Ordnung mit PPM-C-Backoff [12]: Ein Knoten ist ein Kontext, eine Kante trägt die Zählung eines nächsten Symbols, Vorhersage ist eine Relaxation die Suffix-Kette hinab, und es gibt keine gelernten Gewichte—die Topologie ist das Modell. Ein Begleitpaper untersucht dieses Substrat direkt und berichtet zwei Dinge [13]. Erstens ist beschränktes Vergessen ein Regularisierer: Ein Energieerhaltungs-Regime (zerfalls-finanzierte Einzahlungen, Floor-Pruning, Verdrängung (eviction) der energieärmsten Kontexte an einer Gedächtnis-Decke) senkt die held-out-bpc, wann immer die Kapazität die Daten übersteigt, und die Marge wächst, je knapper die Daten werden. Zweitens, und entscheidend, hebt kein Gedächtnis-Regime eine Decke, die das Substrat per Konstruktion trifft: perfektionierte lokale Statistik erzeugt Wörter, die richtig sind, und Sätze, die fast richtig sind, aber keine Bedeutung, die über das Fenster hinaus lebt.

Ein zweites Begleitpaper verortet die Sprachfähigkeit selbst durch eine andere kontrollierte Variation—die Architektur wird fixiert und die Lernregel variiert—und schließt, dass die Fähigkeit im gradientenbasierten Credit Assignment liegt, nicht in der nativen Dynamik irgendeines Substrats [14]. Das vorliegende Paper nimmt beide Ergebnisse als gegeben und stellt die konstruktive Frage, die sie offenlassen: Wenn lokale Statistik nicht binden kann und Bindung das ist, was ein gelerntes Modell liefert, was ist die kleinste, ehrlichste Architektur, die bindet *und* über die Zeit weiterlernt, ohne erneutes Training—auf Hardware, die eine einzelne Person besitzt?

Zwei Seiten der Decke. Die „Bedeutungs-Decke" ist nicht eine Wand, sondern zwei orthogonale, und sie zu vermengen hat das Feld Klarheit gekostet. Die erste ist *Ähnlichkeit*, oder verbindungslose Generalisierung: zu erkennen, dass zwei Kontexte, die kein Oberflächen-Suffix teilen, sich dennoch gleich verhalten, sodass Evidenz über sie hinweg gebündelt werden kann. Die zweite ist *sequenzielle Bindung*: eine spezifische Abhängigkeit über Distanz tragen, wo das relevante Token weit hinten im aktuellen Satz sitzen kann. Ein Zählgraph mit maximaler Ordnung D ist für die zweite strukturell blind: Sein Fenster *ist* D, und eine Abhängigkeit, die länger als D ist, kann nicht beobachtet, geschweige denn repräsentiert werden. Sparse Codes, Locality Hashing und spektrale Einbettungen—all die Maschinerie, die man anflanschen könnte—adressieren nur die erste Seite. Die zweite ist das, was Attention leistet, und nur das, was Attention leistet: inhaltsadressiertes Lesen über Distanz.

2 Zwei Mechanismen für zwei Achsen

Die Trennung ist kein Artefakt unseres Substrats; das Gehirn zeigt sie und löst die zwei Seiten mit zwei verschiedenen Arten von Mechanismus. Die *Ähnlichkeits*-Achse wird von verteilten, überlappenden Codes getragen: Ein Konzept ist ein dünnbesetztes Muster über viele Einheiten, ähnliche Konzepte überlappen, und Überlappung *ist* Ähnlichkeit. Dies ist das kortikale Bild der Repräsentationsgeometrie, und es ergibt sich aus Hebb'schem Lernen über überlappende Eingaben mehr oder weniger von selbst. Es ist im Kern die Idee der Sparse-Distributed-Representation, die das Feld lange umkreist hat [10].

Die *Bindungs*-Achse ist keine reichere statische Repräsentation, sondern ein *zeitlich dynamischer* Mechanismus. Die Korrelationstheorie der Hirnfunktion [9]—Bindung durch Synchronie—schlägt vor, dass Merkmale eines Objekts durch Feuern in zeitlicher Koinzidenz gebunden werden, wobei die Synchronie selbst das Etikett „diese gehören zusammen" ist und die kombinatorische Explosion einer dedizierten Bindungseinheit vermeidet. Für Sequenz speziell hält ein Theta–Gamma-Phasencode geordnete Elemente durch relatives Timing innerhalb eines oszillatorischen Zyklus, und Phasenpräzession komprimiert eine erlebte Sequenz in ein Fenster, das kurz genug ist, dass lokale Plastizität sie erfassen kann. Und beliebige relationale Bindung wird an eine schnelle, dedizierte Struktur delegiert—das hippocampal-entorhinale System—, deren Berechnung als formal nahe an der Attention des Transformers argumentiert wurde [8]. Die Konvergenz ist der Punkt: Bindung, wie auch immer implementiert, konvergiert auf inhaltsadressiertes Lesen über Distanz, und sie ist *separate Maschinerie* gegenüber dem kortikalen Ähnlichkeitscode, keine Verfeinerung davon. Wir nehmen dies als die Lizenz für eine Zwei-Mechanismen-Architektur.

Eine Analogie zur kritischen Periode, mit Vorsicht gebraucht. Menschen erwerben die Bindungsmechanik der Sprache—Grammatik, Kongruenz, langreichweitige Struktur—in einem frühen Fenster, nach dem muttersprachliche Syntax schwer zu erreichen ist, während *Vokabular* und *Fakten* lebenslang erwerbbar bleiben [7]. Dies ist dieselbe Kompetenz/Wissen-Asymmetrie, die unsere Substrat-Studie zieht, nun mit einer zeitlichen Signatur: Die *Mechanik* hat ein Fenster; der *Inhalt* nicht. Das Gehirn hält das Syntax-Fenster nicht offen; es schließt es und friert die Mechanik ein und nimmt in Kauf, dass Neu-Primen danach kostspielig ist. Die Analogie ist suggestiv, nicht beweisend—ein Gehirn ist kein Transformer, und adulte Plastizität ist real, wenn auch begrenzt—, aber sie rahmt unser schwierigstes offenes Problem (Abschnitt 5) um, von „den Binder für immer plastisch halten, ohne zu vergessen" zu „die Mechanik einmal primen, sie einfrieren und die Reichweite durch externes Gedächtnis statt durch Neuverdrahtung erweitern".

3 Die Architektur

Tabelle 1: Epistemischer Status der Behauptungen des Papers. Diese Tabelle ist das Gerüst für die Überarbeitung: Wenn die Experimente aus Abschnitt 6 abgeschlossen sind, sollen *offene* und *hypothetische* Zeilen zu *etablierten* wandern, mit der Beleg-Spalte aktualisiert auf das gemessene Ergebnis.

Behauptung	**Status**	**Beleg / wo**
Das Zähl-Substrat perfektioniert die lokale Statistik, kann aber nicht über Distanz binden	Etabliert	Begleitpaper [13]: 1.83 held-out-bpc, keine langreichweitige Kohärenz bei 11.3M Zeichen
Beschränktes Vergessen ist ein Regularisierer der Stichprobeneffizienz (Kapazität > Daten)	Etabliert	Begleitpaper [13]: niedrigere held-out-bpc bei jeder Größe, Marge wächst mit knapperen Daten
Auf einem kuratierten, regelmäßigen Korpus erreicht der Zähler eine niedrigere bpc und seine Inkohärenz wird *offengelegt* statt verborgen	Etabliert	Diese Arbeit, §6: TinyStories 1.306 prequentielle bpc; flüssige Oberfläche, kein narrativer Faden
Ein von Grund auf gebauter Decoder-Transformer mit handgeschriebener, gradienten-verifizierter Backpropagation lässt sich abhängigkeitsfrei bauen	Etabliert	Diese Arbeit: Gradientenprüfung max. rel. Fehler 4.8×10^{-5}; trainiert und generiert
Ein kleiner Transformer bindet auf kapazitätsangepassten kuratierten Daten	Hypothese (Lit.)	TinyStories [4]; unser Bindungstest ausstehend (§6, [**pending — Lauf**])
Ein Zählgraph mit den Repräsentationen des Binders als Schlüssel ist ein nützlicher, online arbeitender, selektiv vergessender Datastore	Hypothese	Retrieval-Augmentation-Linie [2, 3]; unsere Komposition ungetestet
Selektives Vergessen hält den Datastore unter Repräsentationsdrift ausgerichtet (Selbstheilung)	Hypothese	Argumentiert in §3.1; Driftrate ungemessen
Die eingefrorene Mechanik greift auf neues Wissen, das aus vertrauten Repräsentationen gebaut ist, ohne erneutes Training	Hypothese	In-Context-Learning als Präzedenz; Reichweite noch nicht abgegrenzt
Repräsentationsdrift ist langsamer als die Wiederkehr von Inhalt (sodass die Migration Schritt hält)	Offen	Messbar; §5, [**pending — Driftkurve**]
Der Zähler kann den Binder gegen katastrophales Vergessen für die Erweiterung von *Kompetenz* leiten	Offen	Forschungsfrage; §5, [**pending — A/B-Experiment**]

Die Architektur ist zwei Mechanismen, komponiert als Geschwister, nicht als gestapelte Layer.

Der Binder. Ein kleiner Decoder-only-Transformer—kausale Multi-Head-Attention, Pre-Norm-Residual-Blöcke, Next-Token-Vorhersage [1]—von Grund auf trainiert auf einem kuratierten, kapazitäts-angepassten Korpus. Der Existenzbeweis, dass dies genügt, ist das TinyStories-Ergebnis: Modelle mit wenigen Millionen Parametern erzeugen kohärentes, grammatisches Englisch, wenn die Komplexität der Daten an ihre Kapazität angepasst ist, während dieselben Modelle auf offenem Web-Text nur Brei erzeugen [4]. Wir behandeln Bindung als *ablatierbare Mechanik*: Die Frage ist nicht „bindet ein großes Modell" (es tut es), sondern „tut es ein kleines, von Grund auf, auf kuratierten Daten, auf bescheidener Hardware"—und der Test ist die gelesene Generierung, nicht die Bits-pro-Zeichen, denn

das Begleitpaper zeigt bpc nahe 1.8–2.2, *während* die Bedeutung zerfällt. Unsere Implementierung ist vollständig und ihr Gradient ist verifiziert; was bleibt, ist, sie laufen zu lassen (Abschnitt 6).

Das Gedächtnis. Der Zähler, umgewidmet. Entscheidend ist, dass er nicht semantisch *sein* muss; er borgt sich den Schlüssel des Binders. Wie in kNN-Sprachmodellen [2] nutzt der Datastore die verborgene Repräsentation des Modells als Schlüssel: Der Binder liefert einen bedeutungsvollen Schlüssel gratis (er läuft ohnehin), und der Speicher steuert bei, was er messbar gut kann—schnelles, exaktes, inkrementelles Schreiben und Nachschlagen, mit einer selektiv vergessenden Retention-Policy. Dies löst das Ähnlichkeitsproblem auf, das wir viele Seiten lang umkreist haben: Die Metrik wird nicht *gebaut*, sie wird vom Binder *geerbt*. Der Zähler behält seinen Geist (exakt, online, stichprobeneffizient, vergessend), gibt aber seine Oberflächen-Suffix-*Adressierung* auf—er wird ein repräsentations-geschlüsselter Speicher, kein Zeichen-Suffix-Baum.

Die Komposition. Der Binder läuft auf Rohtext und *fragt* das Gedächtnis *ab*: entweder indem er seine Ausgabeverteilung mit einem Nächste-Nachbarn-Lookup interpoliert [2], oder indem er auf abgerufene Chunks attendiert [3]. Das Kortex→Hippocampus-Bild bricht hier auf lehrreiche Weise: Der Binder empfängt rohe Token, nicht die Ausgabe des Zählers, als Eingabe. Der Zähler ist ein *Seiten*-Speicher, aus dem der Binder liest, kein Fundament, auf dem er sitzt. Deshalb ist „Zähler zuerst" die falsche Reihenfolge: Der Binder wird zuerst geprimt; das Gedächtnis ist das, was hinzukommt; und in der repräsentationsgeschlüsselten Variante muss der Binder überhaupt vorhanden sein, um die Schlüssel zu erzeugen.

3.1 Vergessen als Metabolismus

Ein endlicher Speicher, der über die Zeit schreibt, muss löschen, sonst läuft er über—aber das ist die triviale Hälfte, erfüllbar durch jedes FIFO. Die Hälfte, die zählt, ist spezifisch für ein *dynamisches* System, eines, dessen Schlüssel sich bewegen, während der Binder lernt. Dort ist Vergessen Kohärenzpflege, keine Platzrückgewinnung. Ohne Löschen würde sich der Speicher nicht bloß mit Knoten füllen, sondern mit *veralteten* Einträgen, geschrieben unter Repräsentationen, die der Binder nicht mehr erzeugt; das System würde nicht an Erschöpfung scheitern, sondern an Inkohärenz—einer wachsenden Schicht toter Schlüssel, durch die die lebenden immer schwerer zu finden sind. Selektives Vergessen, das die kalten und nicht-reaktivierten verdrängt, entfernt genau diese drift-veralteten Einträge und hält den Speicher an den aktuellen Repräsentationsraum des Binders ausgerichtet. Ein dummes FIFO tut das nicht; „was nicht mehr getroffen wird, raus" tut es.

Zwei Eigenschaften folgen, und sie sind der Grund, warum das Design einen sich bewegenden Schlüssel toleriert, wo die Standard-Retrieval-Modelle ihren Encoder einfrieren, um ihn zu vermeiden. Erstens ist ein drift-veralteter Eintrag doppelt harmlos: Er wird nicht *abgerufen*, weil Anfragen den aktuellen Schlüssel des Binders verwenden, während der veraltete Eintrag unter dem alten sitzt; und er wird nicht *behalten*, weil er, unreaktiviert, kalt wird und verdrängt wird. Zweitens migriert der Speicher für wiederkehrenden Inhalt: Während der Binder besser wird, kehrt derselbe Inhalt wieder, wird unter dem besseren Schlüssel neu geschrieben, und der alte Eintrag stirbt—Vergessen ist das, was die Migration sauber macht, statt alte und neue Kodierungen den Raum gemeinsam zumüllen zu lassen. Dies hebt den empirischen Befund des Begleitpapers—Vergessen als Regularisierer—zu einem architektonischen Prinzip: In einem dynamischen, lernenden System mit sich bewegenden Schlüsseln ist selektives Vergessen der einzige Weg, auf dem das Gedächtnis zugleich ausgerichtet und endlich bleibt, während alles andere sich bewegt. Es ist der Metabolismus des Systems, nicht dessen Hausmeister.

4 Warum dies zur Beschränkung passt

Das Design ist von einer harten Grenze geformt: eigene Gewichte, kein geerbtes vortrainiertes Modell, kein Rechenzentrum, eine einzige Workstation. Die Beschränkung ist nur deshalb überlebbar, weil zwei Kosten, die die „Moonshot"-Rhetorik vermengt, tatsächlich um Größenordnungen getrennt sind. Allgemeine, offene Kompetenz in Gewichte vorzutrainieren ist der Mondpreis—hier unbezahlbar, und kein Hybrid umgeht ihn. Aber *diese Architektur zahlt ihn nicht*: Sie erwirbt nur die Bindungsmechanik, auf einem kuratierten Korpus, der auf ein kleines Modell abgestimmt ist, was billig ist. Wissen ist der Cent-Preis—ein Online-Durchlauf in einen Index, kein Trainingslauf—und genau das, was der Zähler am besten kann.

Offen als Prozess, nicht als Zustand. Die entscheidende Umdeutung ist, dass „offen" keine

eingefrorene Breite ist, die in Gewichten gehalten wird (der Mondpreis), sondern ein Prozess: schmal beginnen und die Grenze über die Zeit durch das selektiv vergessende Gedächtnis wandern lassen. Der Binder allein ist *fester Zustand*: Um irgendetwas Neues zu lernen, muss er neu trainiert werden, teuer und mit katastrophalem Vergessen. Der Hybrid hält die Kompetenz des Binders fest und lässt Wissen durch das Gedächtnis fließen. Deshalb ist der Binder allein keine billigere Version derselben Sache—er ist ein eingefrorenes Modell mit fester Domäne, genau das, was das Projekt zu vermeiden sucht.

Senkt der Hybrid die Trainingskosten? Ehrlich: nicht für die Bindung. Retrieval lagert *Wissen* aus den Gewichten aus, nicht *Kompetenz*; der Bindungs-Boden—die Parameter und Schritte, bei denen Grammatik und Kongruenz erscheinen—ist irreduzibel und muss einmal bezahlt werden. Was der Hybrid eliminiert, ist die *wiederkehrenden* Kosten des Einarbeitens neuen Wissens in ein Modell: neues Wissen in den Zähler ist ein Durchlauf; in den Binder ist es erneutes Training. In der kNN-LM-Variante wird der Binder genau so trainiert, wie er es allein würde, und der Datastore wird mit nahezu null zusätzlichem Training angehängt, sodass offenes Wissen auf der Trainingsseite fast nichts kostet [2]. Der Hybrid ist nicht billiger zu *starten*; er ist billiger *aktuell zu halten*, für immer.

5 Offene Probleme

Wir benennen die offenen Probleme, statt sie ins Design aufzunehmen, weil sie Forschungsfragen sind, keine Ingenieursarbeit.

Driftrate versus Wiederkehr-Intervall. Die Selbstheilung aus Abschnitt 3.1 hält nur Schritt, wenn Repräsentationen langsamer driften, als Inhalt wiederkehrt: Wissen überlebt die Migration, wenn es unter einem neuen Schlüssel neu geschrieben wird, bevor der alte Eintrag zerfällt; eine Domäne, die verstummt, während der Binder stark driftet, kann verdunsten, bevor sie aufgefrischt wird. Dies ist biologisches Vergessen—man verliert, was man nicht wieder aufsucht—und ist die Selektion, die das Projekt will, außer im Fall von seltenem, aber einmal wichtigem Wissen, dem inhärenten harten Fall jedes endlichen vergessenden Gedächtnisses. Die maßgebliche Größe ist daher nicht „wird es verdrängt" (das wird es), sondern das Verhältnis von Driftrate zu Wiederkehr-Intervall, das messbar ist (§6).

Kontinuierliche Kompetenz und katastrophales Vergessen. Wissen, das aus *vertrauten* Repräsentationen gebaut ist, ist der gelöste, mühelose Fall: Die eingefrorene Mechanik greift in-context auf neue Namen, Fakten und Kombinationen, die sie im Training nie sah, so wie es ein eingefrorenes großes Modell tut. Eine wirklich *neue* Domäne—eine, die Repräsentationen verlangt, die der Binder nie zu bilden gelernt hat—ist es nicht: Ihre Einbettung ist uninformiert, und einen uninformierten Schlüssel zu speichern erfasst nichts Nützliches. Kompetenz dort zu erweitern verlangt, dass der Binder weiterlernt, und da sitzt das katastrophale Vergessen—ein robustes, jahrzehntealtes Phänomen [5], kein offenes Rätsel. Der *naive* Weg (einfach weitertrainieren) ist als gescheitert bekannt. Offen ist, ob nicht-naive Methoden—Replay, elastische Konsolidierung [6]—es zähmen, und hier hat unser Stack einen Kandidaten, keine Lösung: Der selektive Zähler weiß bereits, *welches* alte Wissen salient und wiederkehrend ist, was genau das Signal ist, das solche Methoden zum Schützen brauchen. Wir bieten dies als Hypothese an. Die ehrliche Rahmung, nach der Analogie zur kritischen Periode, ist eine Frage des *Grades*: Wie weit erstreckt sich die mühelose Reichweite (eingefrorene Mechanik auf einem vertrauten Repräsentationsraum, erweitert durch externes Gedächtnis), bevor Neu-Primen nötig wird—und das selektive Gedächtnis ist die Maschinerie, um diese Reichweite durch Retrieval statt durch Neuverdrahtung auszudehnen.

6 Versuchsplan

Die Experimente werden hier im Voraus festgelegt; ihre Ergebnisse werden in diesen Abschnitt eingearbeitet, sobald sie abgeschlossen sind.

1. Der Bindungstest (die Kontrolle steht). Der Zähler und der Binder werden auf demselben kuratierten Korpus (TinyStories [4]) betrieben, sodass die Differenz zwischen ihren Generierungen die Bindung isoliert, während nichts anderes variiert. Die Zähler-Baseline ist bereits gemessen: 1.306 prequentielle Bits/Zeichen, mit einer Generierung, die an der Oberfläche flüssig ist (echte Namen, Märchen-Register, saubere kurze Fragmente) und als Narrativ inkohärent (Figuren verwandeln sich mitten im Satz, kein Faden). Der Binder muss 1.306 auf der langweiligen Achse schlagen

und, auf der Achse, die zählt, beim Lesen einen narrativen Faden halten, wo der Zähler eine texturreiche Wolke erzeugt. Erfolg wird gelesen, nicht gemessen. [**pending — bpc des Binders und Generierungs-Beispiele; Seite-an-Seite-Lesung gegen die Zähler-Baseline.**]

2. Driftrate. Ein kleiner Binder wird trainiert; an festen Checkpoints wird ein festes Probe-Set von Kontexten eingebettet und die Kosinus-Ähnlichkeit zur Einbettung desselben Kontexts am vorherigen Checkpoint aufgezeichnet, was eine Driftkurve ergibt. Eine Kurve, die schnell fällt, rechtfertigt das einfache Design (aufwärmen, dann Schlüssel einfrieren, klassisches Retrieval mit festem Encoder); eine Kurve, die anhält, erzwingt kontinuierliches Neu-Indexieren oder die Vergessen–Drift-Kopplung aus §3.1. Dies erfordert keinen *guten* Binder, nur einen *trainierenden*, und kann daher jetzt auf bescheidener Hardware laufen. [**pending — Driftkurve über Checkpoints; charakteristische Zeit der Stabilisierung.**]

3. Katastrophales Vergessen. Trainiere auf Domäne A, dann auf Domäne B, und miss die Degradation von A. Vergleiche die naive Baseline mit zähler-geleitetem Schutz (Replay, gewichtet durch das Salienz-/Wiederkehr-Signal des Zählers). Relative Degradation, nicht absolute Qualität, ist die Ablesung. [**pending — A-Degradation unter naivem vs. zähler-geleitetem kontinuierlichem Training.**]

Zur Hardware. Die Referenzimplementierung ist reines `f64`, single-threaded und abhängigkeitsfrei, was korrekt, aber nicht schnell ist; der Bindungstest bei einer bindungsfähigen Größe dauert Stunden auf einer bescheidenen mobilen CPU. Die verifizierte Implementierung ist daher die *Spezifikation*, gegen die eine schnellere Variante (thread-parallele Matrixmultiplikationen oder der Autodiff eines GPU-Frameworks) validiert wird: Die Gradientenprüfung und identische Bits-pro-Zeichen sind der Abnahmetest, sodass Geschwindigkeit erkauft wird, ohne die Korrektheit preiszugeben, die der von Grund auf gebaute, handgeprüfte Backward-Pass garantieren sollte.

7 Verwandte Arbeiten

Das Retrieval-Augmentation-Grundgerüst ist etabliert und wir beanspruchen nichts davon: Nächste-Nachbarn-Sprachmodelle interpolieren ein parametrisches Modell mit einem Lookup über einen Datastore und verbessern den langen Schwanz und frisch hinzugefügtes Wissen ohne erneutes Training [2]; RETRO lässt einen vergleichsweise kleinen Transformer auf Chunks attendieren, die aus einem sehr großen Speicher abgerufen werden, und erreicht, was sonst ein weit größeres Modell bräuchte [3]. Unser Beitrag ist enger und spezifisch: (i) ein *selektiv vergessender* Datastore als Retention-Policy anstelle eines statischen, vorgebauten Index—die gemessene Stärke des Zählers als Speicher-Disziplin; (ii) das Lesen von Vergessen als *Kohärenzpflege unter Repräsentationsdrift*, was es dem Schlüssel erlaubt, sich zu bewegen, wo die Standardmodelle den Encoder einfrieren; und (iii) ein Binder nach dem Prinzip *erst primen, dann einfrieren*, mit dem Gedächtnis in der Rolle der adulten Plastizität, das die Reichweite durch Retrieval statt durch Neu-Primen erweitert. Die Prämisse der Bindung im kleinen Maßstab ruht auf TinyStories [4]; die Zwei-Mechanismen-Lesart auf der hippocampal-kortikalen Literatur [9, 8] und der Erklärung der Syntax über eine kritische Periode [7]; katastrophales Vergessen und seine Gegenmaßnahmen auf der konnektionistischen Literatur [5, 6]; und verteilte/holographische Repräsentation auf ihren Klassikern [10, 11]. Das Substrat, sein Vergessens-Regime und seine gemessene Decke sind die Begleitpaper des vorliegenden Autors [13, 14].

8 Fazit

Wir haben, als falsifizierbares Design, ein Zwei-Mechanismen-Sprachsystem dargelegt: einen kleinen, von Grund auf gebauten Transformer, der die Bindungsmechanik in einem Fenster nach Art einer kritischen Periode lernt, und einen selektiv vergessenden Zählgraphen, der offenes Wissen per Retrieval trägt, geschlüsselt auf den Repräsentationen des Binders, mit Vergessen als dem Metabolismus, der den Speicher unter Drift ausgerichtet und endlich hält. Die Architektur ist in gemessenen Vorläufern verankert—einem Substrat, dessen Decke die Bindung ist, einem Vergessens-Regime, das ein Regularisierer ist, und einem gradienten-verifizierten Binder—und in einer Zwei-Mechanismen-Lesart, die die Neurowissenschaft stützt. Ihre zentrale Tugend unter der Beschränkung des Projekts ist, dass sie die Bindungskosten einmal und die Wissenskosten pro Durchlauf zahlt, was „offen" zu einem Prozess statt zu einem eingefrorenen Zustand macht, auf Hardware, die eine einzelne Person besitzt. Ihre zentrale Ehrlichkeit ist, dass der schwierigste

Schritt offen bleibt: Kontinuierlicher Kompetenzerwerb ist ein echtes Forschungsproblem, der naive Weg ist als gescheitert bekannt, und unser selektives Gedächtnis ist darin ein möglicher Hebel, kein Beweis. Aber der Wert hängt nicht an diesem Schritt. Ein Binder, der in seiner Domäne bindet, plus ein Gedächtnis, das frisches Wissen aufnimmt, das aus vertrauten Repräsentationen gebaut ist, ohne erneutes Training, ist bereits „offen als Prozess" für die große Klasse von Wissen, das aus vertrauten Bausteinen besteht. Die Reihenfolge ist daher sauber: zuerst die feststehende Decke durchbrechen (zeigen, dass der Binder bindet), den brauchbaren Boden bauen (Binder plus Gedächtnis im beherrschten Raum), dann die offene Wette eingehen (der Zähler gegen katastrophales Vergessen). Jeder Schritt steht für sich, und keiner hängt davon ab, dass der nächste gelingt.

Literatur

[1] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., Polosukhin, I., "Attention Is All You Need", *Advances in Neural Information Processing Systems* **30** (2017).

[2] Khandelwal, U., Levy, O., Jurafsky, D., Zettlemoyer, L., Lewis, M., "Generalization through Memorization: Nearest Neighbor Language Models", *Proc. International Conference on Learning Representations (ICLR)* (2020).

[3] Borgeaud, S., Mensch, A., Hoffmann, J., et al., "Improving Language Models by Retrieving from Trillions of Tokens" (RETRO), *Proc. International Conference on Machine Learning (ICML)* (2022).

[4] Eldan, R., Li, Y., "TinyStories: How Small Can Language Models Be and Still Speak Coherent English?", arXiv:2305.07759 (2023).

[5] McCloskey, M., Cohen, N. J., "Catastrophic Interference in Connectionist Networks: The Sequential Learning Problem", *Psychology of Learning and Motivation* **24**, 109–165 (1989).

[6] Kirkpatrick, J., Pascanu, R., Rabinowitz, N., et al., "Overcoming Catastrophic Forgetting in Neural Networks", *Proceedings of the National Academy of Sciences* **114**(13), 3521–3526 (2017).

[7] Lenneberg, E. H., *Biological Foundations of Language*, Wiley, New York (1967).

[8] Whittington, J. C. R., Muller, T. H., Mark, S., Chen, G., Barry, C., Burgess, N., Behrens, T. E. J., "The Tolman–Eichenbaum Machine: Unifying Space and Relational Memory through Generalization in the Hippocampal Formation", *Cell* **183**(5), 1249–1263 (2020).

[9] von der Malsburg, C., "The Correlation Theory of Brain Function", in *Models of Neural Networks II*, E. Domany, J. L. van Hemmen, K. Schulten (eds.), Springer, New York, pp. 95–119 (1994); orig. MPI für Biophysikalische Chemie, Internal Report 81-2 (1981).

[10] Kanerva, P., *Sparse Distributed Memory*, MIT Press, Cambridge, MA (1988).

[11] Plate, T. A., "Holographic Reduced Representations", *IEEE Transactions on Neural Networks* **6**(3), 623–641 (1995).

[12] Cleary, J. G., Witten, I. H., "Data Compression Using Adaptive Coding and Partial String Matching", *IEEE Transactions on Communications* **32**(4), 396–402 (1984).

[13] Langjahr, M., "Forgetting as Regularisation in an Energy-Conserving Context Graph", Pantareon Foundations (2026).

[14] Langjahr, M., "The Source of Language Capability in an Energy-Based Associative Substrate", Pantareon Foundations (2026).