

World Models

Nicht das nächste Wort, sondern der nächste Zustand der Welt

Ein Sprachmodell schreibt einen fehlerfreien Text übers Einparken und würde am Steuer trotzdem scheitern. Ihm fehlt ein inneres Abbild von Raum, Schwerkraft und Ursache. World Models bauen genau das: Sie sagen nicht das wahrscheinlichste Wort voraus, sondern den nächsten Zustand einer ganzen Umgebung.

WORAN HEUTIGE MODELLE RÄUMLICH SCHEITERN

- **Beziehungen.** Entfernungen, Größen und Ausrichtungen werden geraten. Was rechts liegt, wird links verortet.
- **Perspektive.** Ein neuer Blickwinkel gelingt nicht, weil es keine echte 3D-Darstellung gibt.
- **Physik.** Gegenstände verschwinden, durchdringen einander, bewegen sich unnatürlich.
- **Beständigkeit.** Nach wenigen Sekunden ist die Szene vergessen, Figuren wechseln Position und Aussehen.

WAS RÄUMLICHE INTELLIGENZ LEISTET

- **Erfassen.** Eine Umgebung als zusammenhängender Raum, nicht als flaches Bild.
- **Physik mitrechnen.** Schwerkraft und Zusammenstöße gelten automatisch.
- **Folgen abschätzen.** Wenn ich dieses Ding bewege, passiert als Nächstes das.
- **Handeln.** Schritte planen und mit Menschen zusammen ausführen.

DREI EIGENSCHAFTEN, DIE EIN WORLD MODEL AUSMACHEN

GENERATIV
Erzeugt aus einer Eingabe eine vollständige 3D-Umgebung, die nach physikalischen Regeln funktioniert und über die Zeit zusammenhängend bleibt.

MULTIMODAL
Verbindet Text, Fotos, Video und Tiefendaten zu einem gemeinsamen Weltverständnis, wie ein Mensch seine Sinne verbindet.

INTERAKTIV
Reagiert. Greift ein Roboter nach etwas, wird der neue Zustand berechnet, samt der nötigen Zwischenschritte.

DER UNTERSCHIED IN EINER ZEILE

	Sprachmodell	World Model
Sagt voraus	Das nächste Wort	Den nächsten Zustand der Umgebung
Optimiert auf	Wahrscheinlichkeit	Räumliche und physikalische Stimmigkeit
Sieht die Welt als	Flache Eindrücke	Zusammenhängenden Raum mit Regeln

Auch multimodale Modelle, die Bild und Text verbinden, bleiben bei flachen Eindrücken.

WOZU DAS GUT IST

Feld	Was möglich wird	Warum es ohne Weltmodell nicht geht
Architektur	Aus einer Beschreibung ein begehbare 3D-Entwurf, Varianten wie Licht oder barrierefreie Wege prüfen	Ein Renderer liefert Bilder, keinen Raum mit Regeln
Robotik	Millionen Durchläufe in einer digitalen Halle, bevor Hardware anläuft	Reale Fehlversuche kosten Material und gefährden Menschen
Medizin und Lehre	Eingriffe und Notfälle in einer reagierenden Umgebung üben	Am Patienten gibt es keinen zweiten Versuch

WAS NOCH OFFEN IST

- **Datenbasis.** Text und Bilder reichen nicht. Es braucht Video, Tiefendaten und Sensorik, ohne dass sich das Modell in Widersprüche verstrickt.
- **Verarbeitung.** Weltzustände sind riesig. Sie über längere Zeit stabil zu halten, ist ungelöst.
- **Genauigkeit.** Nicht nur überzeugend aussehen, sondern echte Ursache-Wirkung treffen.

MERKSÄTZE

Die Prüffrage bei jedem Anbieter: Arbeitet das System nur mit Sprache und Pixeln oder mit einem Modell der physischen Welt?

Ein größeres Sprachmodell wird nicht von selbst räumlich.

Erste Arbeitsabläufe laufen in Film, Design und Gaming. Ein Standardprodukt ist es nicht.