Leveraging Language Models for Clinical Data Management:
A Case Study on Text Cleaning and Mapping

Kian Norouzi
Novo Nordisk

Kian Norouzi
DM Process and Innovation Specialist
Novo Nordisk

Views and opinions expressed are those of the speaker and not Novo Nordisk

Introduction

The Challenge of Free Text Data

Free text data presents a unique set of challenges in data cleaning and mapping. Its unstructured nature, coupled with the potential for human error, misspellings, and variations in terminology, makes it a complex field to navigate. In the realm of healthcare, these challenges are amplified due to the critical importance of accuracy and consistency

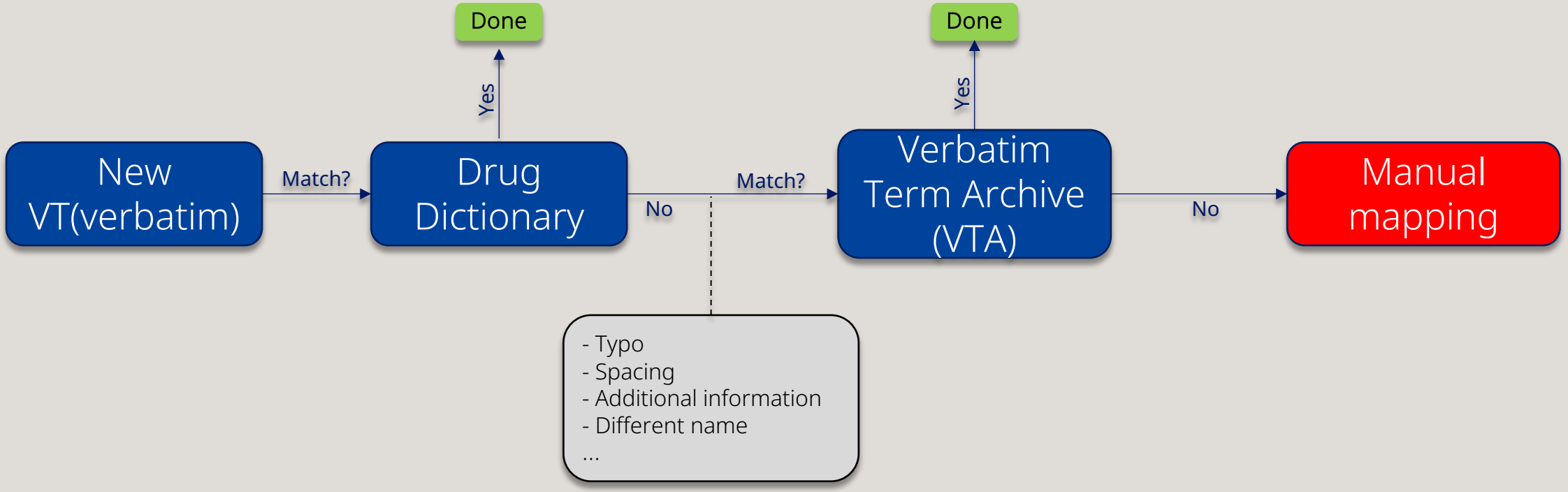
Concomitant Medication Mapping

Mapping concomitant medication verbatim to standard dictionary terms is a common approach to dealing with free text in healthcare. This process involves translating raw, unstructured medication data into a standardized format that can be easily understood and analyzed. However, traditional methods such as Regular Expression and Fuzzy matching have their limitations

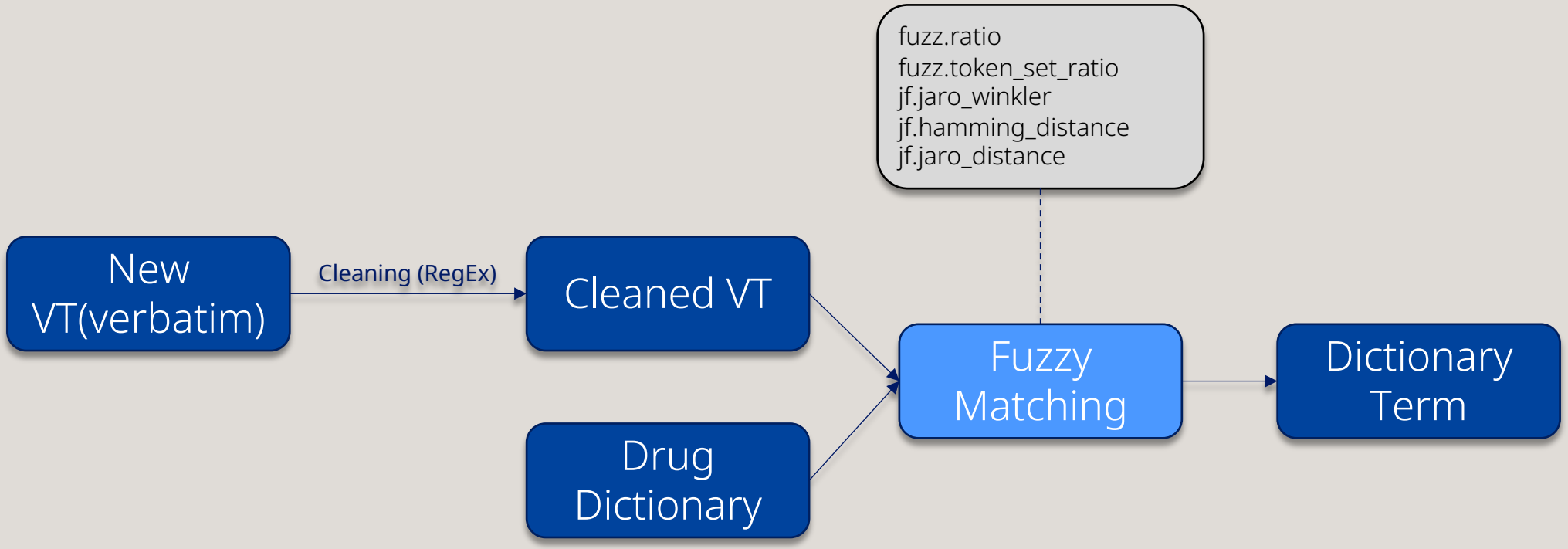
The Advent of LLMs: A New Era in Text Mapping

With the advent of Large Language Models (LLMs) and embedding models, we now have more sophisticated tools at our disposal. These models offer promising opportunities to improve the accuracy and efficiency of text cleaning and mapping

Current Process



First Approach



Regular Expression (RegEx)

Regular Expressions, often shortened to "regex", are like a secret code or special language used to find specific patterns in text. Regular expressions use special symbols to represent these patterns. Here are a few examples:

. (dot): This is like a wildcard. It can represent any single character. For example, **h.t** could match 'hat', 'hot', 'hit', etc.

*** (asterisk):** This means "zero or more of the previous thing". For example, **ho*t** could match 'ht', 'hot', 'hoot', 'hoooot', etc.

+ (plus): This means "one or more of the previous thing". For example, **ho+t** could match 'hot', 'hoot', 'hoooot', but not 'ht'.

^ (caret): This symbol is used to indicate the start of a line. If you put ^ before your pattern, it means that the pattern should be at the beginning of a line. For example, **^The** would match any line that starts with 'The'.

$ (dollar sign): This symbol is used to indicate the end of a line. If you put $ after your pattern, it means that the pattern should be at the end of a line. For example, **end$** would match any line that ends with 'end'.

[abc] (square brackets): This means "any one of the characters inside the brackets". For example, **h[aei]t** could match 'hat', 'het', 'hit', but not 'hot'.

RegEx - Example

^([a-zA-Z0-9_-.]+)@([a-zA-Z0-9_-.]+)\.([a-zA-Z]{2,5})$

Basic Email Address Validation Regular Expression

Cleaning VTs Using RegEx

NORMOSOL 400ML

COAPROVEL 300/25 ONE TABLET DAILY

ACETOMINOPHN 500MG 1-2 TABLETS EVERY 6 HOURS PRN

500MG OF METFORMINN TWICE A DAY

1% LIDOKAIN LOCAL ANESTHESIA FOR LUMBAR PUNCTURE

1 LITRE OF HARTMANS SOLUTION IV INFUSION VIA ELECTRONIC PUMP @125 MLS PER HOUR.

```
(?<![A-Z]\d?)\b\d+\s*(GRAMS|GRAM|MG|UG|ML|MMOL|G|PINT|PERCENT|LITER|LITRE|LTR|UNITS OF|UNIT OF|UNITS|UNIT|L|IU|IE|%)\b
(\b(ONE|TWO|THREE|FOUR|FIVE|SIX|SEVEN|EIGHT|NINE|TEN|\d+)\b[-/]*)?\b(TABLETS?|ORAL|SOLUTION IN WATER|EYE DROPS?|CAPSULES?)\b
\b((EVERY|PER) \d+ (HOURS?|DAYS?)?|ONCE A DAY|TWICE A DAY|THREE TIMES A DAY|FOUR TIMES A DAY|DAILY|WEEKLY|MONTHLY|YEARLY|AS NEEDED)\b
```

.

.

.

Fuzzy Matching

- A technique used in computer-based information analysis and retrieval that identifies non-exact matches of your pattern or search criteria. It's useful when dealing with typos, spelling variations, or other minor discrepancies in data
- Fuzzy matching algorithms, like the **Levenshtein distance**, quantify the degree of similarity between two strings, allowing for more flexible and comprehensive search and analysis

Levenshtein Distance

The Levenshtein distance between two strings is the minimum number of single-character edits (insertions, deletions, or substitutions) required to change one string into the other.

VT	DT	Distance	Similarity
ACETAMINOFEN	ACETAMINOPHEN	2	0.85
ACETAMINOPHN	ACETAMINOPHEN	1	0.92
TYLENOL	ACETAMINOPHEN	11	0.15

Cleaning VTs using GPT models

1. NORMOSOL 400ML
2. COAPROVEL 300/25 ONE TABLET DAILY
3. ACETOMINOPHN 500MG 1-2 TABLETS EVERY 6 HOURS PRN
4. 1% LIDOKAIN LOCAL ANESTHESIA FOR LUMBAR PUNCTURE
5. 1 LITRE OF HARTMANS SOLUTION IV INFUSION VIA ELECTRONIC PUMP @125 MLS PER HOUR.
6. 500MG OF METFORMINN TWICE A DAY

Prompt

You are a medical coding professional at a pharmaceutical company. Your job is to map concomitant medications to standard drug terms. The verbatim is collected as free text, which may contain additional information such as drug dosage, administration method, form, as well as potential typos and misspellings.
Please clean the provided text and return only the drug name.

1. NORMOSOL
2. COAPROVEL
3. ACETAMINOPHEN
4. LIDOCAINE
5. HARTMANS SOLUTION
6. METFORMIN

Word Embeddings

Word embeddings are a way to represent words and whole sentences in a numerical manner

Word Embeddings: Turning words into numerical representations

Vector Space: The mathematical space where these numerical representations exist

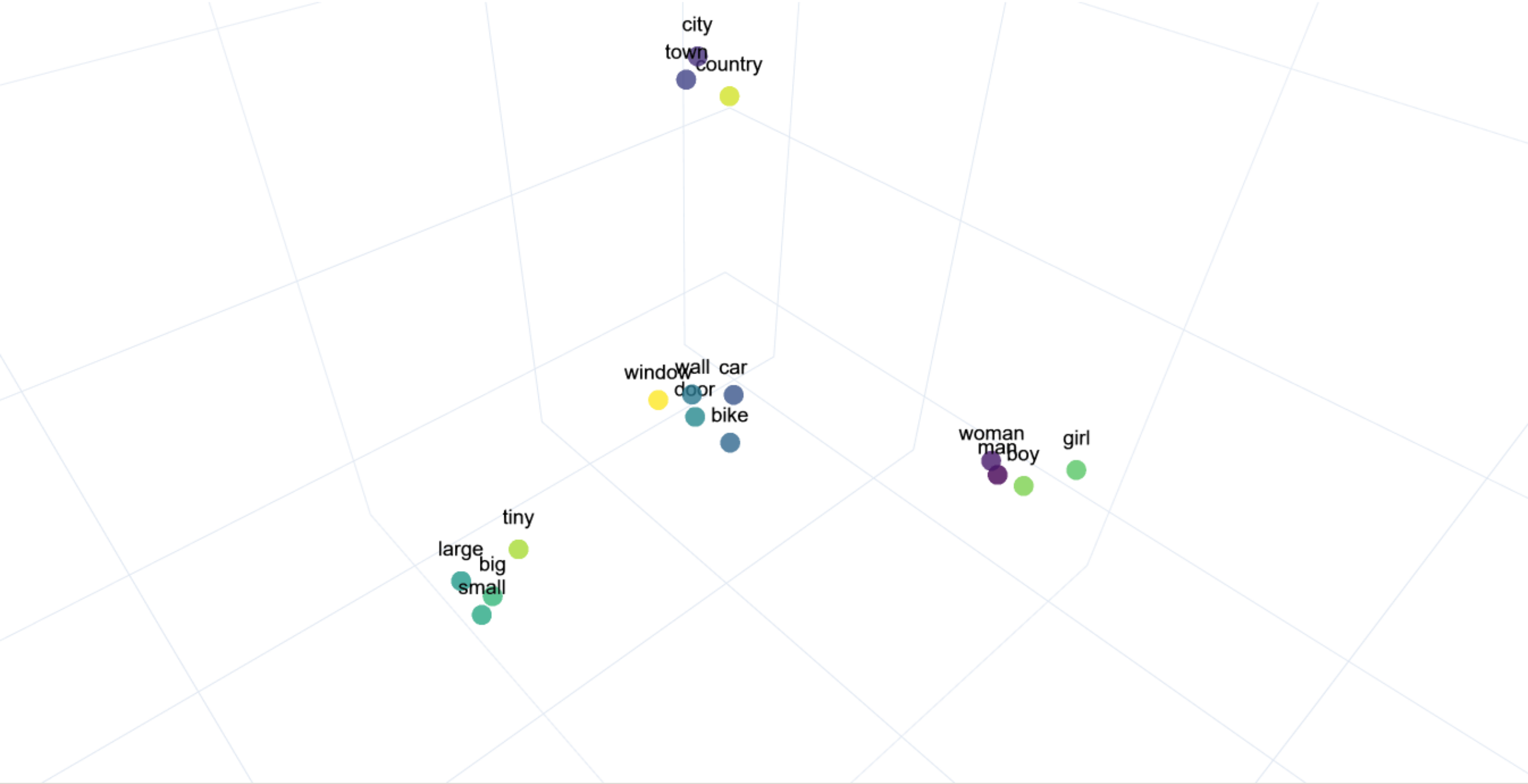
Contextual Learning: The process of learning word meanings from their usage in text

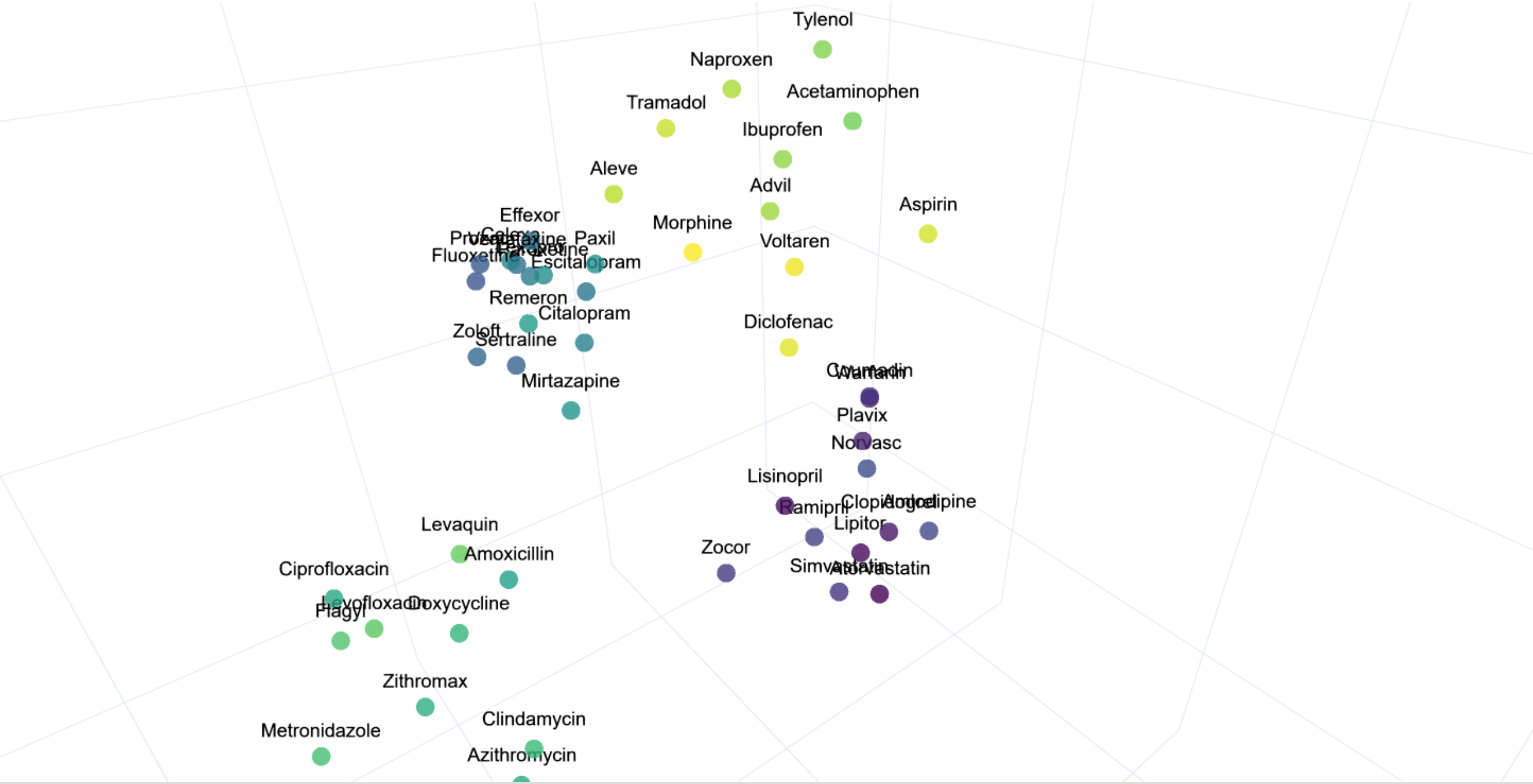
Dimensionality: The number of features (numbers) used to represent each word

Semantic Relationships: The relationships between words captured by their proximity in the vector space

Acetaminophen (OpenAI text-embedding-ada-002)

- [0.015241741202771664, 0.010303111746907234, 0.007805195171386035, -0.04838204011321068, -0.014911227859556675, 0.021114705130457878, -0.026898685842752457, -0.011301007121801376, 0.003495812183246016, [illegible]


city
town
country
window
wall
car
door
bike
woman
man
boy
girl
tiny
large
big
small

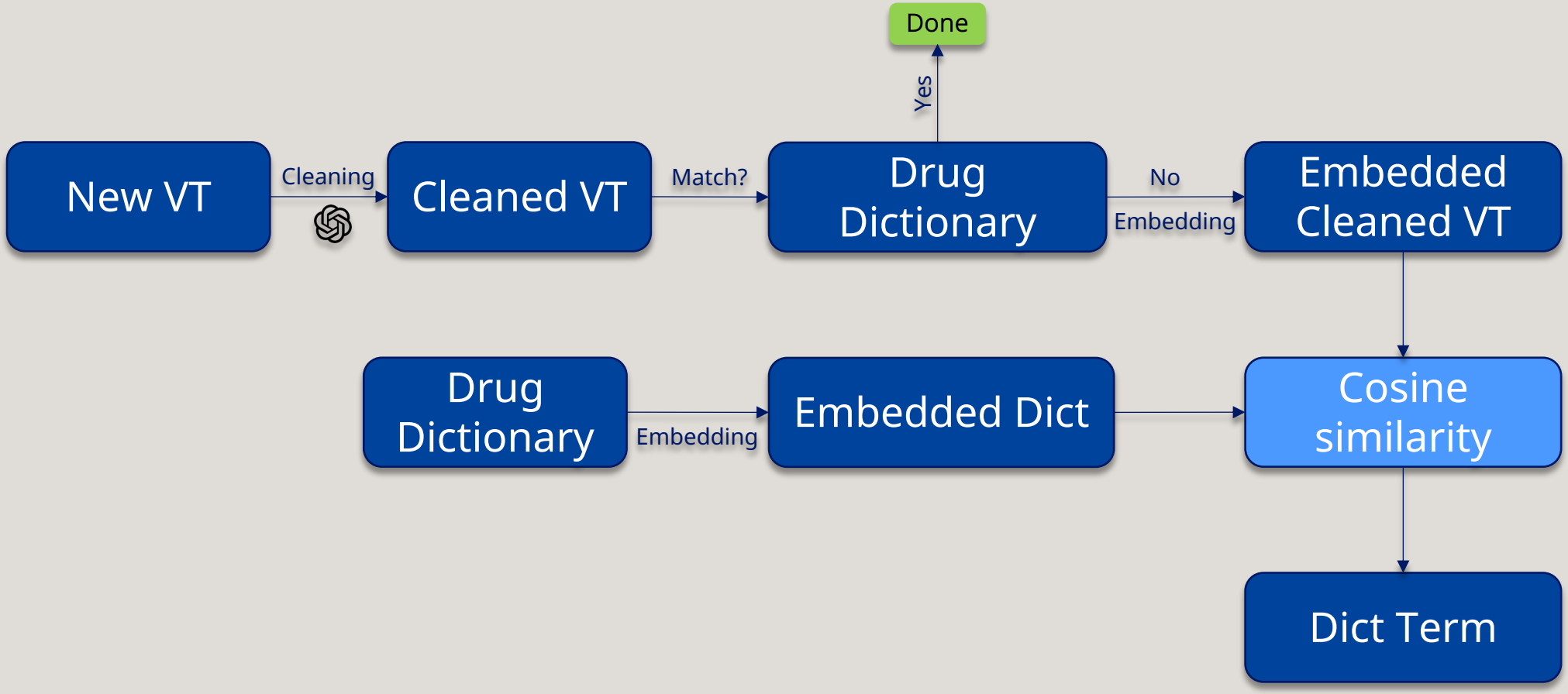

Tylenol
Naproxen
Acetaminophen
Tramadol
Ibuprofen
Aleve
Advil
Aspirin
Effexor
Morphine
Paxil
Voltaren
Fluoxetine
Escitalopram
Remeron
Citalopram
Diclofenac
Zoloft
Sertraline
Mirtazapine
Plavix
Norvasc
Lisinopril
Ramipril
Lipitor
Zocor
Levaquin
Amoxicillin
Ciprofloxacin
Doxycycline
Flagyl
Zithromax
Clindamycin
Metronidazole
Azithromycin

Cosine Similarity

A measure that calculates the cosine of the angle between two vectors, providing a similarity score between 0 and 1, and is often used to compare word embeddings in natural language processing.

		Similarity
ACETAMINOFEN	ACETAMINOPHEN	0.93
ACETAMINOPHN	ACETAMINOPHEN	0.95
TYLENOL	ACETAMINOPHEN	0.93

The Alternative Approach



Conclusion and considerations

- Improve the quality by Prompt Engineering
- Be aware of the nature of Semantic Search
- We still need the human in the loop
- This was just one example, what else can be done?!

Thank You!

Kian Norouzi
vxmn@novonordisk.com
/in/knorouzi