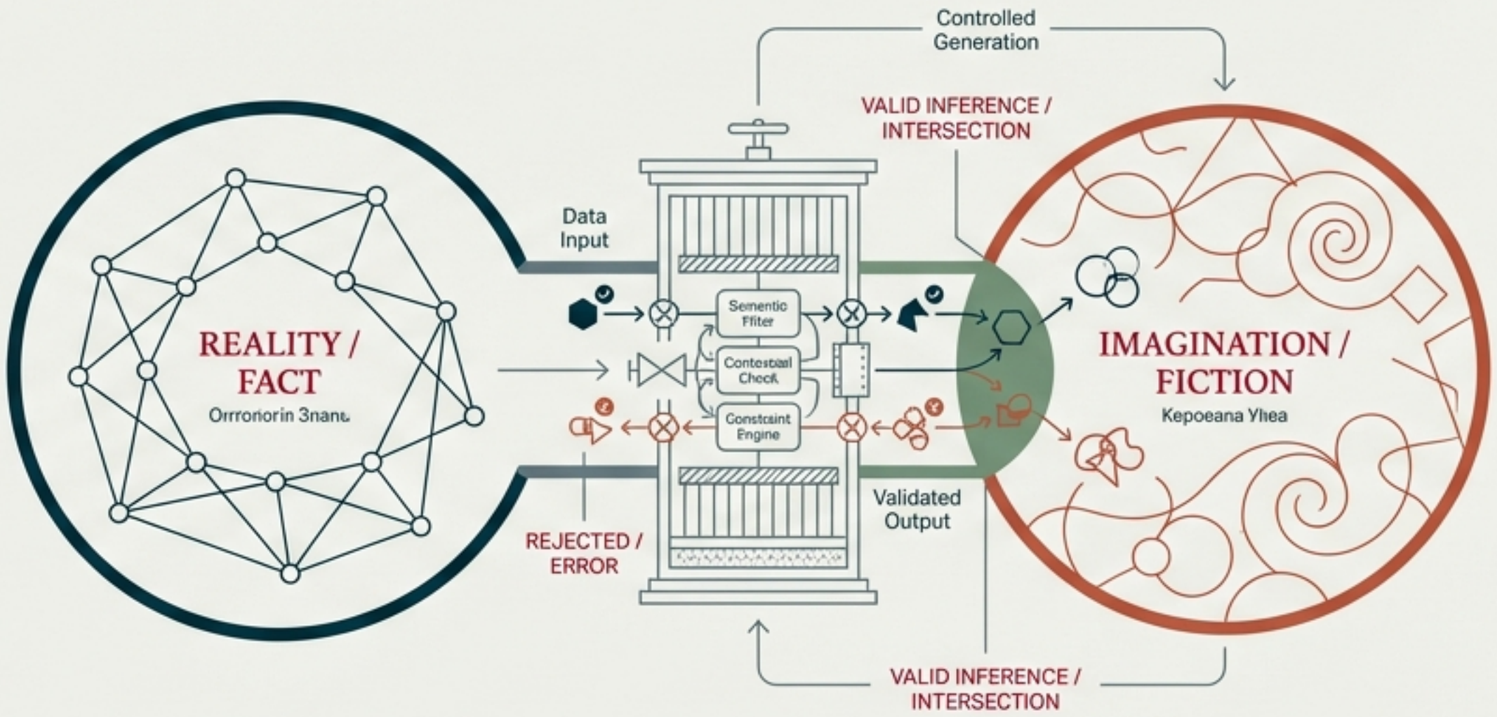
Осмислення галюцинацій: Семантичні бар’єри для ШІ

Від виправлення помилок до керованої уяви: новий підхід до онтологій та графів знань.



Vagan Terziyan, Svitlana Gryshko, Amit Shukla,
Oleksandr Terziyan, Oleksandra Vitko

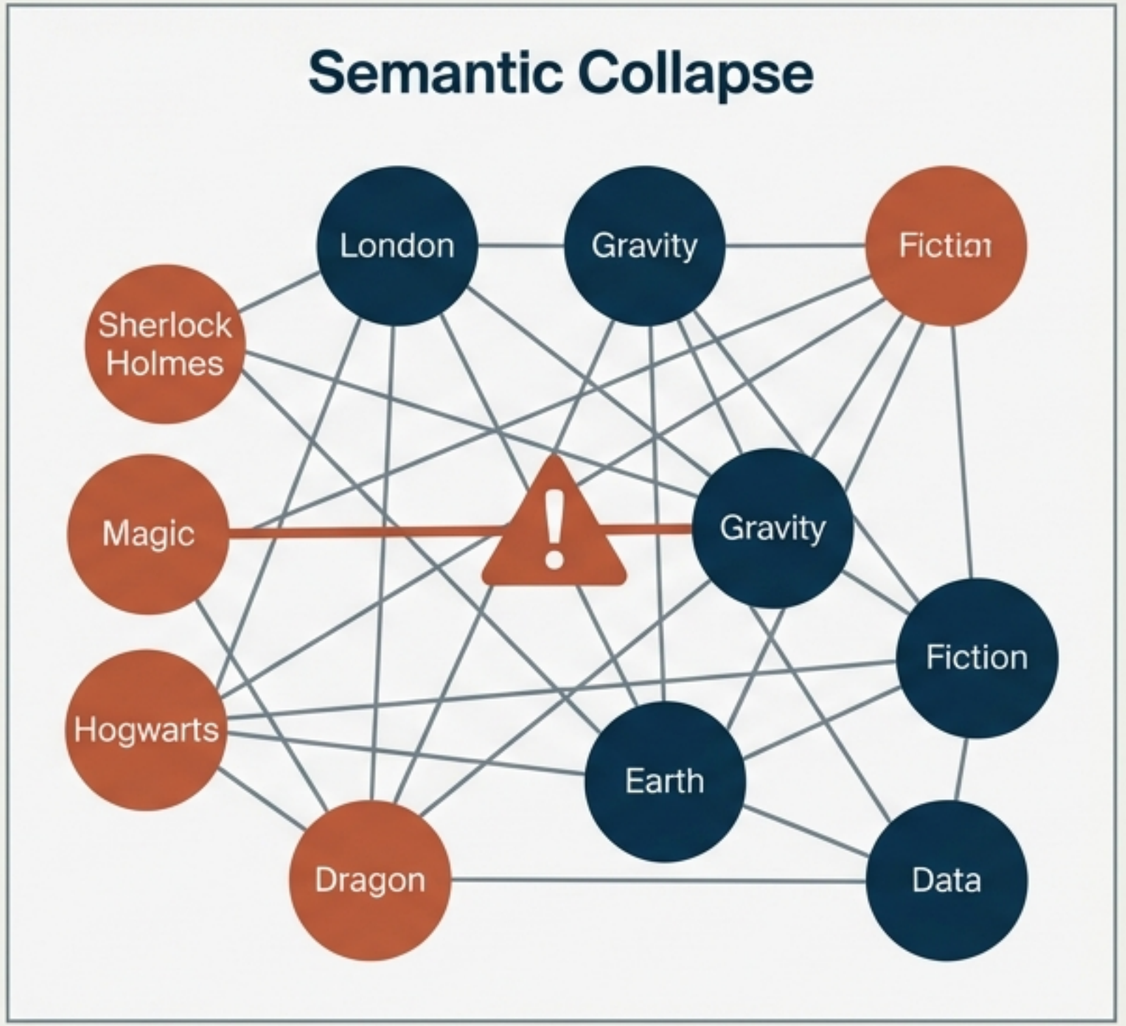
NotebookLM

Проблема сучасної семантики: Однополярний світ

Проблема 1:
Єдина реальність.
Сучасні онтології (RDF/OWL) розраховані на єдину, об’єктивну істину.

Проблема 2:
Обмеження контексту.
Мітки типу ‘True-in-Context’ лише додають теги, але не створюють окремих онтологічних режимів.

Проблема 3:
Семантичний колапс.
Змішування фактів і фікції в одному просторі призводить до суперечностей.



Плоский простір знань: Конфлікт між реальністю та уявою.

NotebookLM

Галюцинація чи некерована уява?

Сприйняття фактів

Побудова можливих світів (Mental Simulation)

Когнітивна наука

Люди мислять симуляціями. Ми постійно створюємо “можливі світи” для планування та навчання.

Зміна парадигми

Галюцинації в LLM — це функціональний аналог людської уяви, а не просто помилка.

Мета

Не придушувати уяву, а надати їй семантичні кордони.

NotebookLM

Що таке Семантичні бар’єри (Semantic Guardrails)?

Визначення: Формальні обмеження на рівні онтології, які регулюють логічне виведення між різними реальностями.

Першокласні об'єкти
Світи як об'єкти в графах.

Ізоляція
Відокремлення фактів від фікції.

Атрибуція
Збереження походження висновків.

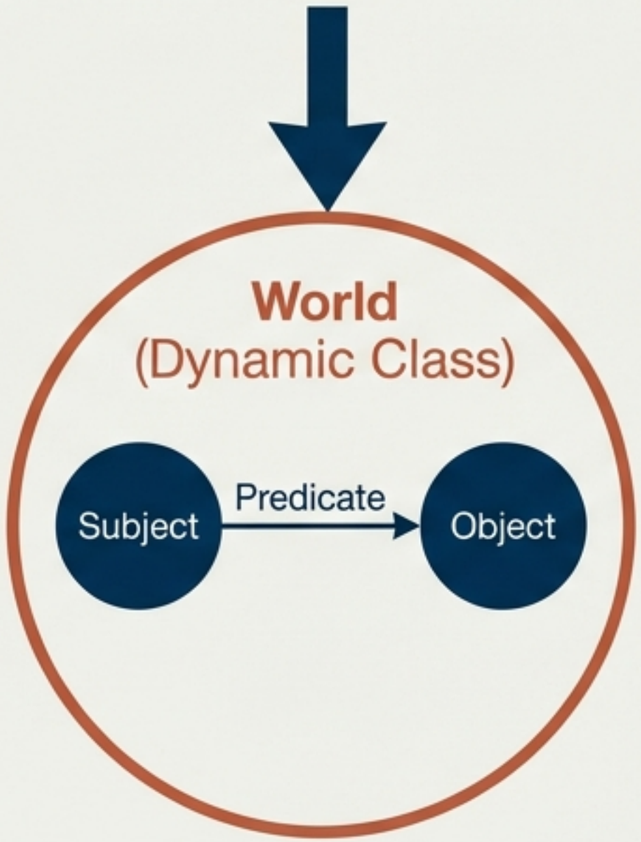
Метафора:
Це не стіна, що блокує рух, а правила дорожнього руху для логіки.

NotebookLM

Як це працює: Від атомарних триплетів до Світів

Guardrail 1: Atomic Induction

{Subject, Predicate, Object}



1. **Індукція:** Кожен RDF-триплет автоматично створює свій власний "атомарний світ".

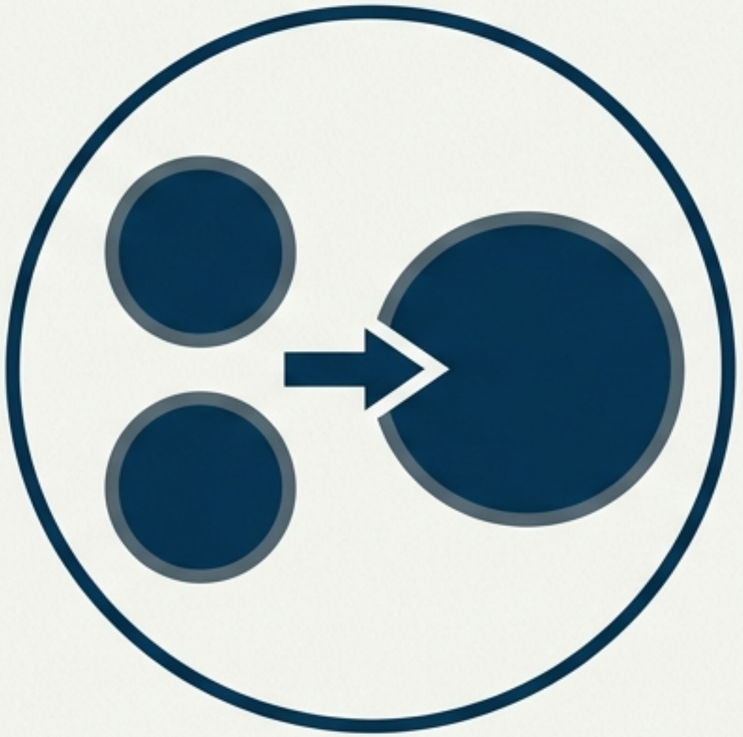
2. **Структура:** Світи — це не статичні коробки, а динамічні OWL-класи.

3. **Цілісність:** Якщо граф містить дані зі Світу А, весь граф належить до Світу А. Дані не можуть існувати в порожнечі.

NotebookLM

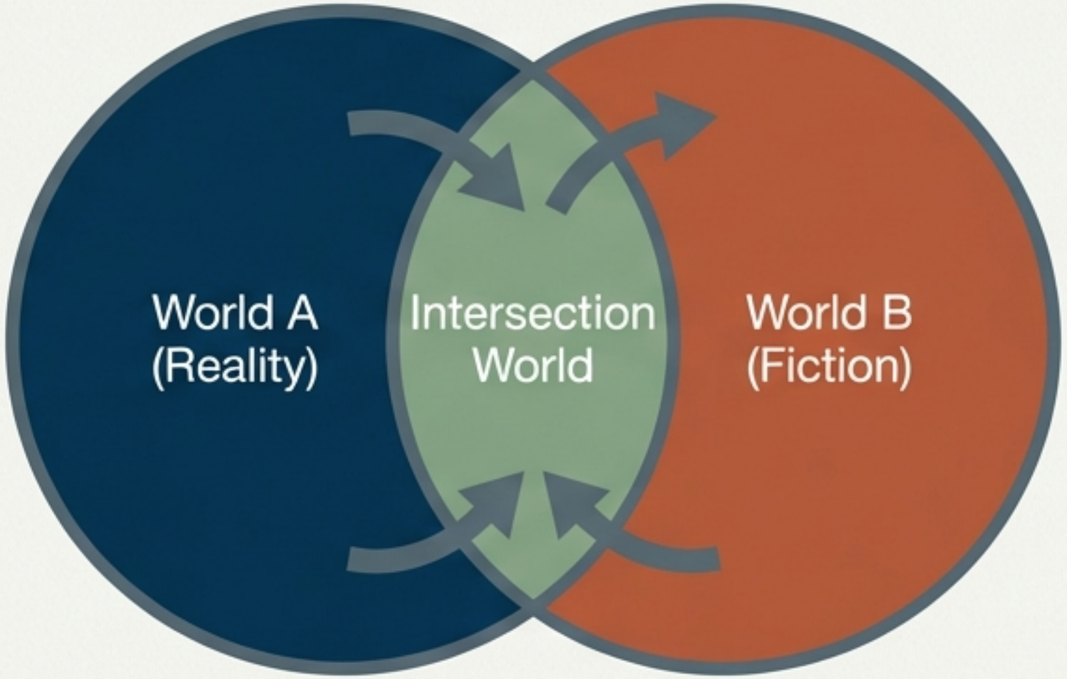
Об’єднання світів: Механіка злиття

A. Guardrail 2: Intra-world Merge



Злиття фактів однієї реальності
= Розширення знань.

B. Guardrail 3: Cross-world Intersection



Злиття різних реальностей =
Створення нового світу-перетину.

Оригінальні світи залишаються чистими. Гібридна реальність існує окремо.

NotebookLM

Логіка в ізоляції: Правила виведення

Guardrail 4: Logic in Isolation

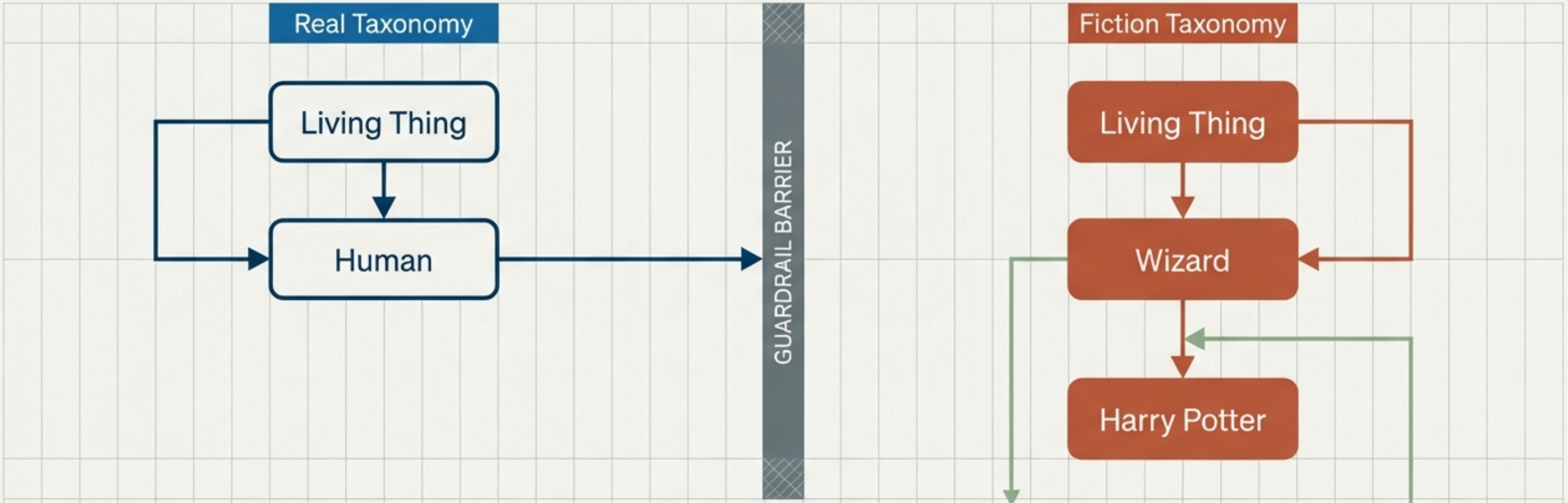
Локалізація правил: Правило "Левітація" діє лише в контексті, де магія можлива.

Конструктивне виведення: Результат виконання правила потрапляє в перетин світів (Rule World ∩ Graph World).

NotebookLM

Цілісність онтології: Класи та Властивості

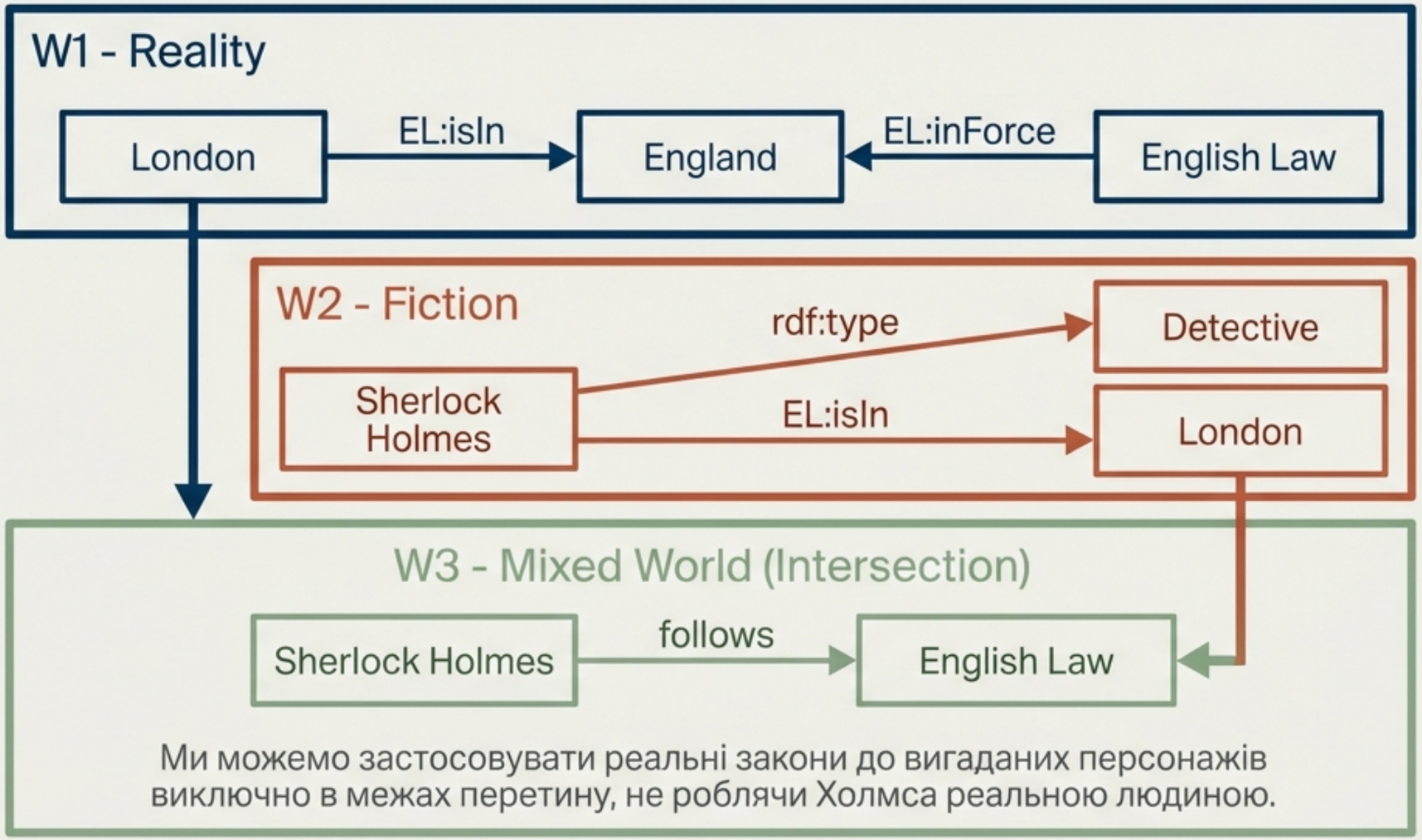
Guardrails 5 & 6: Запобігання витоку схеми (Schema Leakage)



- Ізоляція класів: Якщо вигаданий світ визначає клас "Wizard", це визначення не псує таксономію "Human" у реальному світі. **Ізоляція властивостей:** Домени та **діагастивостей:** Домени та діапазони властивостей (наприклад, 'castsSpell') залишаються локальними для свого світу.
- **Результат:** Чиста онтологія навіть при використанні суперечливих визначень.

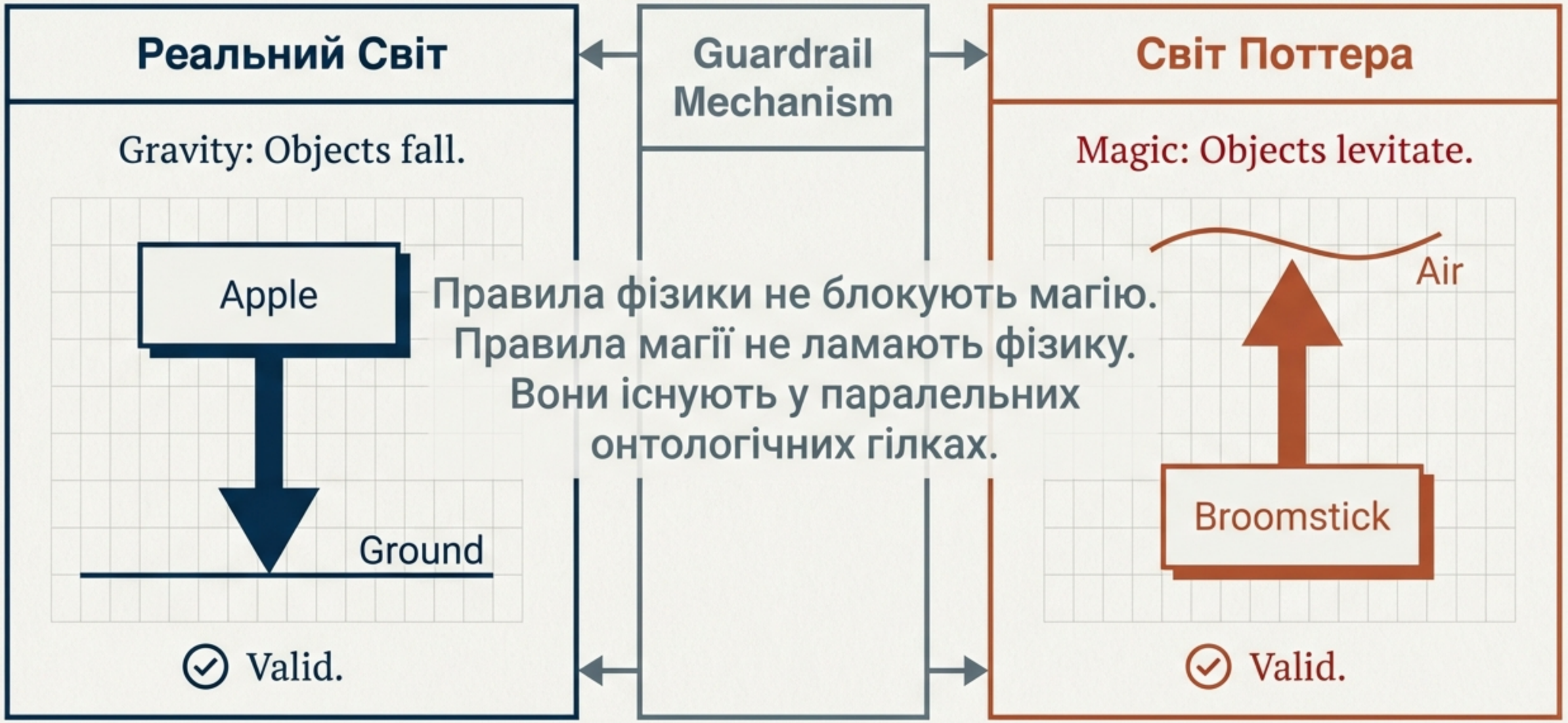
NotebookLM

Кейс 1: Шерлок Холмс і "справжній" Лондон



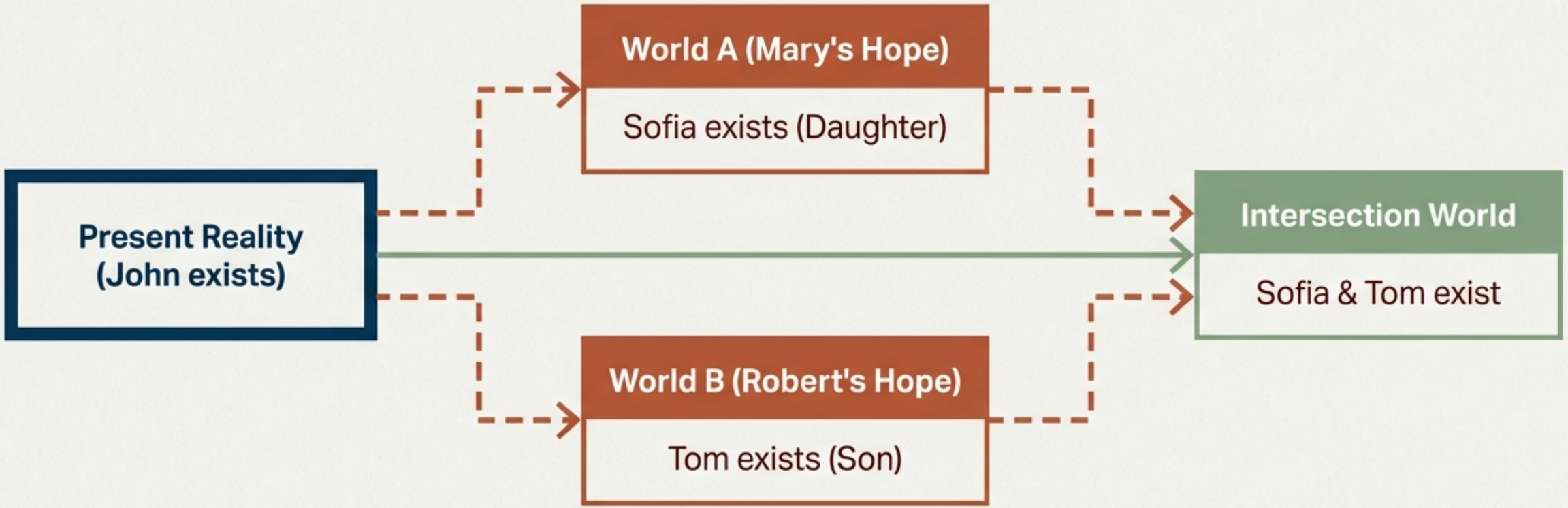
NotebookLM

Кейс 2: Гаррі Поттер і конфлікт законів фізики



NotebookLM

Кейс 3: Планування майбутнього (Гіпотези)



Система відстежує взаємовиключні або гіпотетичні сценарії як окремі світи. Жоден варіант не оголошується "хибним" — вони є валідними планами.

NotebookLM

Технічна реалізація та екосистема

Inputs

LLMs DBpedia WikiData

Semantic Guardrails Engine

- Guardrails Ontology (OWL)
- Rules (N3 Logic)
- Reasoner (e.g., EYE)

Multi-World Knowledge Space

Structured Imagination

NotebookLM

Значення для Generative AI та LLM

- Current State: LLM ‘сплющує’ всі знання в один простір.
- Future State: Guardrails діють як проміжне ПЗ, що сортує фантазії у відповідні світи.
- Result: ШІ залишається креативним, але стає епістемічно підзвітним.

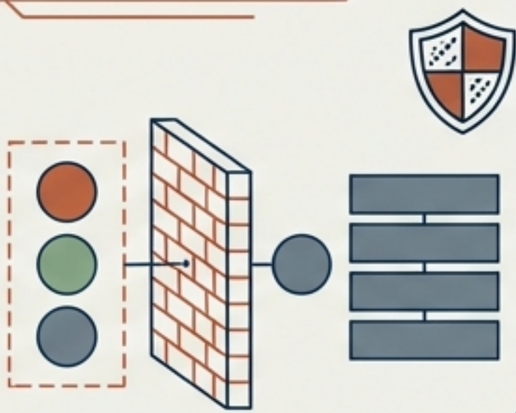
NotebookLM

Переваги підходу "Multi-Reality"

01.

Безпека

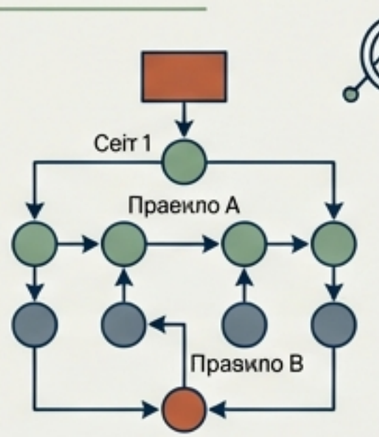
Ізоляція фіктивного та гіпотетичного контенту від фактичних даних.



02.

Пояснюваність

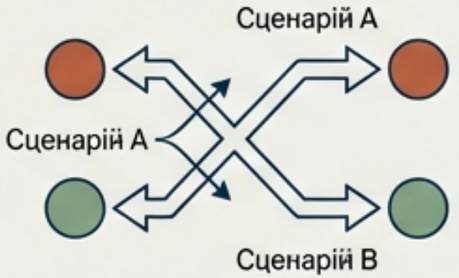
Кожен висновок можна відслідкувати до конкретного "Світу" та набору правил.



03.

Гнучкість

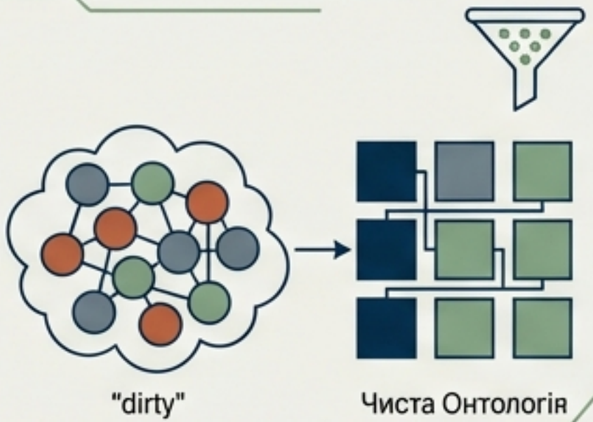
Можливість міркувати над суперечливими сценаріями (контрфактуальне мислення).



04.

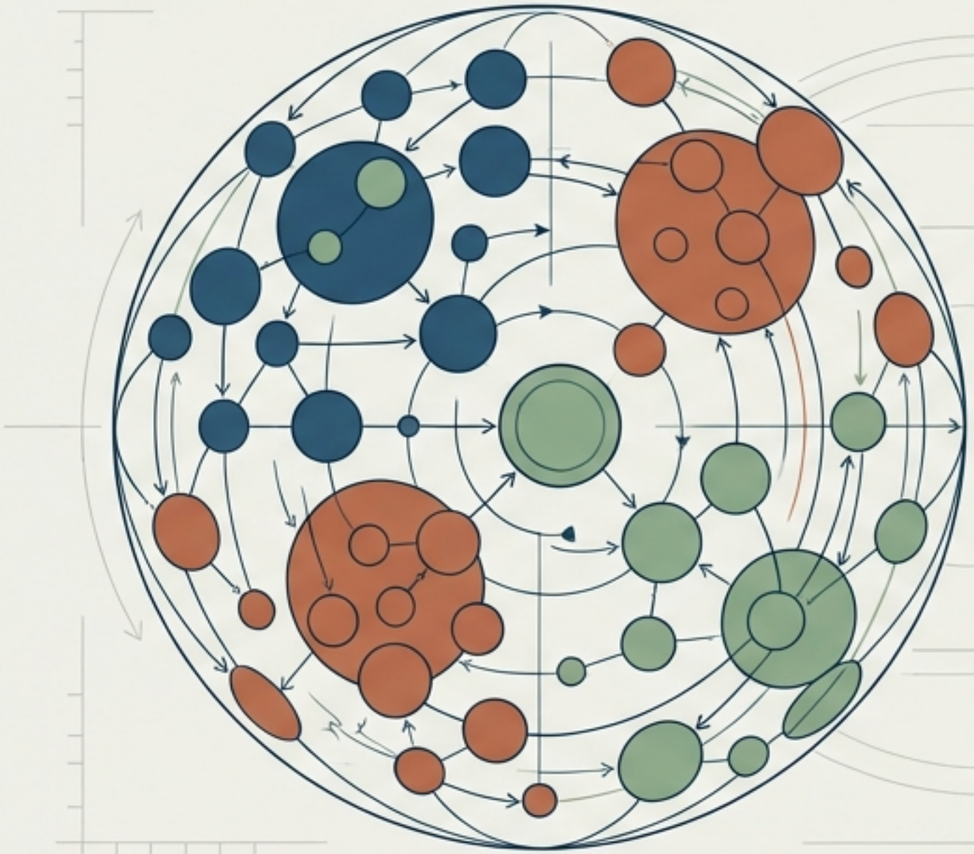
Цілісність

Підтримка чистих онтологій навіть при роботі з "брудними" даними LLM.



NotebookLM

Майбутнє: Від боротьби з галюцинаціями до їх використання



Замість того, щоб намагатися “виправити” галюцинації, ми повинні їх моделювати.

Semantic Guardrails — це фундамент для гібридних нейро-символічних систем.

Структурована Уява.

NotebookLM