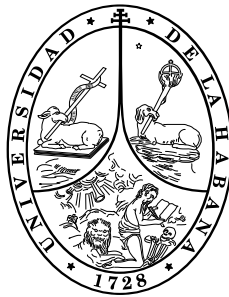


Universidad de La Habana  
Facultad de Matemática y Computación



# **Construcción Incremental de Bases de Conocimiento Semi-estructuradas con Modelos de Lenguaje**

Autor: Carlos Mauricio Reyes Escudero

Tutor: Dr.C Alejandro Piad Morffis

Trabajo de Diploma presentado en opción al título de  
**Licenciado en Ciencia de la Computación**

Junio de 2025

∀ aquellos que aún buscan saber

# Índice general

<b>1. Introducción</b>	<b>1</b>
1.1. Gestión del Conocimiento Personal . . . . .	1
1.1.1. Toma de notas . . . . .	2
1.1.2. Bases de Conocimiento Personal . . . . .	2
1.1.3. Sistemas Digitales de Toma de Notas . . . . .	3
1.2. Grandes Modelos de Lenguaje . . . . .	4
1.2.1. Técnicas de <i>Prompting</i> . . . . .	4
1.2.2. Optimización de Contexto . . . . .	5
1.2.3. Agentes Basados en <i>LLM</i> . . . . .	6
1.3. Extracción de Conocimiento y Enlaces . . . . .	7
1.3.1. Extracción de Conocimiento . . . . .	7
1.3.2. Creación de Enlaces . . . . .	8
1.4. <i>LLM</i> en Aplicaciones de Toma de Notas . . . . .	9
1.5. Fundamentos y Alcance de la Investigación . . . . .	11
<b>2. Importer</b>	<b>13</b>
2.1. Descripción del entorno . . . . .	14
2.2. Detalles de Implementación . . . . .	14
2.2.1. Agente reAct . . . . .	15
2.2.2. Interacción con la bóveda . . . . .	15
2.2.3. Sobre la versatilidad del sistema . . . . .	17
<b>3. Experimentos</b>	<b>18</b>
3.1. Base estructurada sobre Harry Potter . . . . .	18
3.2. Añadiendo artículos a una <i>Wiki</i> tecnológica . . . . .	19
3.3. Projects, Areas, Resources, Archived . . . . .	19
3.4. Note augmented <i>LLMs</i> are computationally universal . . . . .	20
<b>Bibliografía</b>	<b>23</b>

# Capítulo 1

## Introducción

La sobrecarga de información en la actualidad plantea un desafío significativo para la productividad y el aprendizaje efectivo. En este contexto, la gestión del conocimiento personal emerge como una disciplina fundamental, no solo para académicos e investigadores, sino para cualquier individuo que busque optimizar su capacidad de aprendizaje y generación de ideas. Como señala Ahrens, *How to Take Smart Notes: One Simple Technique to Boost Writing, Learning and Thinking – for Students, Academics and Nonfiction Book Writers*, una gestión eficaz de las notas y el conocimiento adquirido no solo facilita la escritura y el estudio, sino que transforma la manera en que se interactúa con la información, convirtiéndola en un activo dinámico y generador de nuevas perspectivas. El desarrollo de un *segundo cerebro* (Forte, *Building a Second Brain*), un sistema externo confiable para almacenar y conectar ideas, libera recursos cognitivos, permitiendo un enfoque más profundo en el pensamiento crítico y la creatividad. El potencial de los Grandes Modelos de Lenguaje (LLM) para revolucionar este campo es inmenso. Estas tecnologías ofrecen la posibilidad de automatizar y enriquecer la integración del conocimiento, asistiendo en la destilación de información, la identificación de conexiones y la generación de contenido relevante dentro de las bases de conocimiento personales. La automatización de estos procesos no solo promete un aumento en la eficiencia, sino también una democratización del acceso a metodologías avanzadas de gestión del conocimiento.

### 1.1. Gestión del Conocimiento Personal

La Gestión del Conocimiento Personal (PKM, por sus siglas en inglés) se define como el proceso mediante el cual un individuo recopila, clasifica, almacena, busca, recupera y comparte conocimiento en sus actividades diarias (Grundspenkis, «*Agent Based Approach for Organization and Personal Knowledge Modelling*»). Esta disciplina es particularmente relevante para los denominados *knowledge workers*, profesionales para quienes el conocimiento es su activo más valioso y que dedican una parte significativa de su tiempo a

gestionar grandes cantidades de información. La PKM busca empoderar a estos individuos para que sean más efectivos en sus entornos personales y sociales.

### 1.1.1. Toma de notas

La toma de notas, lejos de ser un mero acto de transcripción, es una estrategia fundamental para mejorar el aprendizaje y la retención de información (Jansen et al., «[An Integrative Review of the Cognitive Costs and Benefits of Note-Taking](#)»). Al tomar notas, el individuo no solo registra la esencia de la información, sino que también se involucra en un proceso activo de filtrado, organización y reestructuración del conocimiento.

Existen diversas metodologías para la toma de notas, que pueden clasificarse ampliamente en lineales y no lineales. Las metodologías lineales implican registrar la información en el orden en que se recibe. Un ejemplo común es denominado *outlining*, donde las notas y pensamientos se organizan de manera estructurada y lógica. Otra técnica lineal es el *sentence method*, que consiste en anotar cada tema como una oración corta y simple, ideal para lecciones de ritmo rápido.

Las metodologías no lineales, por otro lado, utilizan la organización espacial y los diagramas para ensamblar la información. Entre estas se encuentra el *charting*, útil para temas que pueden desglosarse en categorías. El *mapping*, como los mapas mentales, organiza las ideas en una estructura de árbol a partir de un punto central. Las Notas Cornell, desarrolladas por Walter Pauk (Pauk et al., [How to Study in College](#)), dividen la página en secciones para notas, pistas y un resumen. Las *guided notes* proporcionan un esquema predefinido con puntos clave faltantes que el estudiante completa. Más allá de estas técnicas tradicionales, han surgido sistemas más sofisticados. El método *Zettelkasten*, popularizado por Niklas Luhmann y descrito en detalle por Ahrens, [How to Take Smart Notes: One Simple Technique to Boost Writing, Learning and Thinking – for Students, Academics and Nonfiction Book Writers](#), se basa en la creación de notas atómicas interconectadas, formando una red de conocimiento que fomenta la generación de ideas y la escritura prolífica. Este sistema, aunque con raíces históricas, ha ganado nueva relevancia en la era digital. Por otro lado, el método PARA (*Projects, Areas, Resources, Archives*), propuesto por Forte, [Building a Second Brain](#), ofrece un marco para organizar la información digital en función de su accionabilidad, facilitando la gestión de proyectos y responsabilidades a largo plazo.

### 1.1.2. Bases de Conocimiento Personal

Una Base de Conocimiento Personal (PKB) es una herramienta electrónica utilizada por un individuo para expresar, capturar y recuperar conocimiento personal. Se diferencia de una base de datos tradicional al contener material subjetivo y específico del propietario.

El concepto de extender la memoria y las capacidades cognitivas del individuo mediante herramientas externas no es nuevo. Ya en 1945, Vannevar Bush imaginó el *Memex* (de *memory extension*), un dispositivo electromecánico en el que un individuo almacenaría todos sus libros, registros y comunicaciones, mecanizado de tal manera que pudiera ser consultado con gran velocidad y flexibilidad (Bush, *As We May Think*). Bush previó un futuro donde la sobrecarga de información requeriría nuevas formas de acceder y conectar el conocimiento acumulado.

La forma en que una PKB organiza el conocimiento se define por su modelo de datos. Estudios como los de Davies y colegas (Davies et al., «[Building the Memex Sixty Years Later: Trends and Directions in Personal Knowledge Bases](#)»; Davies, «[Still Building the Memex](#)») han analizado estos modelos en función de su marco estructural (cómo se interrelacionan los elementos), los elementos de conocimiento (las unidades básicas de información) y su esquema (el nivel de semántica formal). Un aspecto crucial destacado es la transclusión, la capacidad de ver el mismo elemento de conocimiento en múltiples contextos sin duplicación.

*Grafos de Conocimiento Personal (PKG).* Dentro de las PKB, los Grafos de Conocimiento Personal han ganado prominencia. Un PKG representa el conocimiento como una red de nodos interconectados, donde cada nodo es una pieza de información y las aristas representan las relaciones entre ellas, a menudo con visualizaciones gráficas que facilitan la exploración y el descubrimiento de conexiones (Pyne et al., «[Meta-Work](#)»).

### 1.1.3. Sistemas Digitales de Toma de Notas

La visión del *Memex* de Bush encuentra un eco contemporáneo en la plétora de sistemas digitales de toma de notas. Estos sistemas varían ampliamente en sus características y enfoques. Algunos, como *Roam Research* y *Logseq*, enfatizan los enlaces bidireccionales y la visualización gráfica, alineándose estrechamente con la idea de un PKG. Otros, como *Notion*, ofrecen una gran flexibilidad para crear bases de datos y vistas personalizadas, mientras que *Evernote* y *Google Keep* son populares por su simplicidad y accesibilidad multiplataforma. *Microsoft OneNote* tiene casi todos los *checks* en el artículo de la *Wikipedia*. Alternativas de código abierto como *Zettlr* y *Joplin* también ofrecen robustas funcionalidades, a menudo con un enfoque en la privacidad y el control local de los datos. Herramientas más tradicionales como *VimWiki* (para usuarios de *Vim*) y *OrgMode* (para *Emacs*) ofrecen sistemas de toma de notas altamente personalizables y potentes para usuarios con conocimientos técnicos. *Markor* y *SimpleNote* se centran en la simplicidad y la edición *Markdown*.

Entre estas herramientas, *Obsidian.md* ha ganado una considerable popularidad para la escritura académica y la gestión del conocimiento personal. Siguiendo el modelo de

datos de las *PKB*, *Obsidian* almacena las notas como archivos locales de texto plano en formato *Markdown*, lo que garantiza la portabilidad y la longevidad de los datos. Su marco estructural se basa en un grafo de conocimiento, donde cada nota es un nodo y los enlaces bidireccionales permiten crear una red interconectada de información. Esto facilita la *transclusión*, ya que una misma idea o nota puede ser referenciada y contextualizada desde múltiples puntos del grafo sin necesidad de duplicación. Los elementos de conocimiento son flexibles, desde conceptos simples hasta notas extensas. *Obsidian* también soporta un esquema enriquecido mediante el uso de metadatos (*frontmatter*), etiquetas y la posibilidad de extender su funcionalidad mediante *plugins*, como *Dataview*, que permite realizar consultas complejas sobre las notas. Su vista de grafo visualiza las conexiones, ayudando a identificar relaciones y patrones emergentes.

## 1.2. Grandes Modelos de Lenguaje

Los Grandes Modelos de Lenguaje (LLM) representan un avance transformador en el campo de la inteligencia artificial, con la capacidad de comprender, generar y manipular el lenguaje natural a niveles sin precedentes. Modelos como *GPT-4* de *OpenAI* (OpenAI et al., [GPT-4 Technical Report](#)) y *Gemini* de *Google* (Team et al., [Gemini](#)) han demostrado habilidades notables en una amplia gama de tareas lingüísticas. En paralelo, la comunidad de código abierto ha respondido con modelos competitivos como *DeepSeek* (DeepSeek-AI et al., [DeepSeek-V3 Technical Report](#)), *Qwen* (Bai et al., [Qwen Technical Report](#)) y *LLaMa* (Grattafiori et al., [The Llama 3 Herd of Models](#)), impulsando la innovación y la accesibilidad en este campo.

### 1.2.1. Técnicas de Prompting

Para interactuar eficazmente con los LLM, se han desarrollado un sinfín de técnicas de *prompting*. Inicialmente, se descubrió que proporcionar unos pocos ejemplos (*few-shot learning*) mejoraba significativamente los resultados, apelando a la capacidad de generalización de los modelos sin necesidad de reentrenamiento (Brown et al., [«Language Models Are Few-Shot Learners»](#)). Investigaciones posteriores han explorado qué partes de estos ejemplos son cruciales, sugiriendo que la estructura del texto o el formato pueden ser más importantes que la etiqueta correcta para que el modelo aprenda la tarea (Min et al., [«Rethinking the Role of Demonstrations»](#)).

Diversas estrategias se centran en guiar el comportamiento del modelo. El *role-prompting* (Kong et al., [Better Zero-Shot Reasoning with Role-Play Prompting](#)), que asigna un rol específico al LLM (e.g., actuar como un experto), y el *style prompting* (Lu et al., [Bounding the Capabilities of Large Language Models in Open Text Generation with Prompt Constraints](#)), que especifica el estilo deseado, permiten adaptar la respuesta a necesidades concretas. Para

mejorar la calidad en tareas complejas, se ha aprovechado la capacidad de los LLM para inducirlos a generar pasos intermedios de razonamiento, un enfoque conocido como *cadena de pensamiento* (*Chain-of-Thought*) (Nye et al., *Show Your Work*; Wei et al., *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models*). Este paradigma se ha extendido incluso a escenarios sin ejemplos (*zero-shot CoT*), donde se instruye al modelo a *pensar paso a paso* (Kojima et al., *Large Language Models Are Zero-Shot Reasoners*; Wang et al., *Plan-and-Solve Prompting*).

Otras técnicas se enfocan en la descomposición y la consistencia. El *complexity-based prompting* aprovecha que respuestas más complejas a menudo correlacionan con una mayor probabilidad de acierto (Fu et al., *Complexity-Based Prompting for Multi-Step Reasoning*), mientras que la descomposición de problemas (*least-to-most prompting*) resuelve sub-tareas más simples para integrar sus respuestas en la solución final (Zhou et al., *Least-to-Most Prompting Enables Complex Reasoning in Large Language Models*). La consistencia también se ha utilizado como criterio, generando múltiples respuestas y eligiendo la más frecuente (*self-consistency*) (Wang et al., *Self-Consistency Improves Chain of Thought Reasoning in Language Models*). Para problemas que requieren planificación, estructuras como el *Tree of Thoughts* y el *Graph of Thoughts* permiten al LLM explorar diferentes caminos de razonamiento (Yao et al., *Tree of Thoughts*; Besta et al., «*Graph of Thoughts*»).

Finalmente, se han desarrollado enfoques de auto-mejora y meta-razonamiento. Técnicas como *Self-Ask* incitan al LLM a generar y responder preguntas de seguimiento para clarificar la tarea (Press et al., *Measuring and Narrowing the Compositionality Gap in Language Models*), mientras que *Self-Refine* establece un marco iterativo donde el modelo critica y mejora sus propias respuestas (Madaan et al., *Self-Refine*). El *step-back prompting* instruye al modelo a resolver primero una versión más abstracta del problema para simplificar la tarea (Zheng et al., *Take a Step Back*). Otros enfoques exploran la inducción de un análisis previo del modo de proceder (*meta-reasoning*) (Gao et al., *Meta Reasoning for Large Language Models*), el uso de representaciones cognitivas como el *Sketch of Thought* (Aytes et al., *Sketch-of-Thought*) o la generación y evaluación iterativa de pasos de razonamiento (*Cumulative Reasoning*) (Zhang et al., *Cumulative Reasoning with Large Language Models*).

### 1.2.2. Optimización de Contexto

A pesar de que los LLM modernos poseen ventanas de contexto nominalmente muy grandes, su capacidad para utilizar eficazmente toda esa información es limitada, un fenómeno conocido como *lost in the middle* (Liu et al., «*Lost in the Middle: How Language Models Use Long Contexts*»). La optimización de contexto busca mitigar este problema presentando al modelo únicamente la información más relevante y densa para una tarea específica.

La principal estrategia para ello es la generación aumentada por recuperación (*Retrieval-*



*Augmented Generation* o RAG), que combina la memoria paramétrica del LLM con una memoria no paramétrica externa (Lewis et al., [Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks](#)). En lugar de procesar un documento extenso, un sistema RAG primero recupera los fragmentos de texto más relevantes de una base de conocimientos (e.g., *Wikipedia*) y luego los proporciona como contexto al LLM para que genere la respuesta final. La expansión de la ventana de contexto en modelos como *Gemini 1.5* y *Qwen2.5-1M* ha hecho que el aprendizaje con muchos ejemplos (*many-shot in-context learning*) se vuelva una estrategia viable y potente, donde se pueden incluir numerosos ejemplos relevantes directamente en el *prompt* (Team et al., [Gemini 1.5](#); Yang et al., [Qwen2.5-1M Technical Report](#); Agarwal et al., [Many-Shot In-Context Learning](#)).

La efectividad de RAG depende críticamente de la granularidad de la información recuperada. En lugar de recuperar pasajes de longitud fija, se ha demostrado que el uso de unidades más finas como las *proposiciones* —unidades de texto atómicas y autocontenidas— mejora la densidad de la información relevante, reduce el ruido y aumenta el rendimiento en tareas posteriores (Chen et al., [«Dense X Retrieval»](#)).

Más allá de la recuperación, han surgido marcos de compresión y razonamiento agéntico. **QwenLong-CPRS** es un sistema que, mediante instrucciones en lenguaje natural, comprime dinámicamente un contexto extenso en un resumen optimizado y específico para la consulta, actuando como un intermediario inteligente antes de pasar la información al LLM final (Shen et al., [QwenLong-CPRS](#)). Por otro lado, la *Cadena de Clarificaciones* (*Chain-of-Clarifications*) propone un flujo de trabajo en el que el modelo se enseña a sí mismo generando preguntas de clarificación sobre la consulta original, recuperando evidencia para responderlas y refinando iterativamente su comprensión antes de dar la respuesta definitiva (Zhuang et al., [Self-Taught Agentic Long Context Understanding](#)). Finalmente, los *Recitation-Augmented LLMs* se centran en mejorar la capacidad del modelo para recordar y citar fielmente la información del contexto proporcionado, ajustándolo para que aprenda a copiar explícitamente los segmentos relevantes, lo que mejora la fiabilidad de las respuestas basadas en el texto (Sun et al., [Recitation-Augmented Language Models](#)).

### 1.2.3. Agentes Basados en LLM

El verdadero potencial de los LLMs reside en su capacidad para actuar de forma autónoma y utilizar herramientas externas para superar sus limitaciones inherentes, como en cálculos matemáticos, razonamiento complejo o la verificación de hechos. A medida que los LLM han mejorado, investigadores y empresas han explorado cómo permitirles interactuar con sistemas externos.

El sistema MRKL (*Modular Reasoning, Knowledge, and Language*) (Karpas et al., [MRKL Systems](#)) es una de las formulaciones más simples de un agente, utilizando un LLM como *router* para acceder a múltiples herramientas (e.g., obtener el clima o la fecha actual) y com-

binar la información para generar una respuesta final. Modelos como *PAL (Program-aided Language Model)* (Gao et al., [PAL](#)) traducen problemas directamente a código ejecutable, mientras que *ToRA (Tool-Integrated Reasoning Agent)* (Gou et al., [ToRA](#)) intercala pasos de código y razonamiento. El paradigma *ReAct (Reasoning and Acting)* (Yao et al., [ReAct](#)) permite a los agentes generar un pensamiento, tomar una acción y recibir una observación, manteniendo un historial de estos pasos para informar decisiones futuras. *Reflexion* (Shinn et al., [Reflexion](#)) extiende *ReAct* incorporando retroalimentación lingüística para refinar el comportamiento del agente. (RAG) se considera un sistema de agente cuando la propia recuperación se trata como una herramienta externa.

Para organizar la creciente diversidad de agentes y establecer un marco conceptual coherente, se han propuesto las *Arquitecturas Cognitivas para Agentes de Lenguaje (CoALA)*, por sus siglas en inglés (Sumers et al., [Cognitive Architectures for Language Agents](#)). Inspirado en la ciencia cognitiva y la inteligencia artificial simbólica, *CoALA* describe a los agentes de lenguaje en función de tres componentes clave: una memoria modular (de trabajo, episódica, semántica y procedimental), un espacio de acciones estructurado (acciones internas como razonar o aprender, y acciones externas para interactuar con el entorno) y un proceso de toma de decisiones generalizado. Este marco no solo permite clasificar y comparar los sistemas existentes, como *ReAct* o *Reflexion*, sino que también proporciona un plano para diseñar agentes futuros más capaces y estructurados, contextualizando los desarrollos actuales dentro de la historia más amplia de la IA.

### 1.3. Extracción de Conocimiento y Enlaces

La extracción de conocimiento y la creación de enlaces entre piezas de información son tareas fundamentales en la construcción de bases de conocimiento robustas y útiles. Tradicionalmente, estos procesos han sido manuales y laboriosos, pero los avances recientes en *LLMs* han abierto nuevas vías para su automatización y enriquecimiento.

#### 1.3.1. Extracción de Conocimiento

La extracción de conocimiento, en el contexto de la gestión del conocimiento personal, implica identificar y capturar las ideas clave, entidades y relaciones presentes en diversas fuentes de información. Los *LLM* han demostrado una notable capacidad para esta tarea, abordándola desde diversas perspectivas metodológicas.

Un enfoque fundamental es la *extracción de conocimiento abierto*, que busca generar representaciones lógicas a partir de texto sin restricciones de dominio. Investigaciones pioneras como la de (Van Durme et al., [«Open Knowledge Extraction through Compositional Language Processing»](#)) utilizan el procesamiento composicional del lenguaje para derivar proposiciones estructuradas, mientras que trabajos más recientes como (Song

et al., «OpenFact») se centran en la veracidad y expresividad de los tripletes extraídos, utilizando marcos semánticos y enlaces a *Wikidata* para garantizar la calidad. A escala industrial, (Qian et al., *Open Domain Knowledge Extraction for Knowledge Graphs*) presenta un *framework* para extraer conocimiento de la web abierta, abordando desafíos de volumen, variedad y veracidad.

Para estructurar este proceso, se han propuesto diversos *frameworks* y *pipelines* basados en *LLM*. El trabajo de (Zhang et al., *Extract, Define, Canonicalize*) introduce un marco de tres fases (Extraer-Definir-Canonizar) que realiza una extracción abierta seguida de una definición y canonización de esquemas *post-hoc*, permitiendo una gran flexibilidad. De manera similar, (Kommineni et al., *From Human Experts to Machines*) propone un *pipeline* semi-automatizado que abarca todo el proceso de construcción de grafos de conocimiento, desde la generación de preguntas de competencia (CQs) hasta la creación de la ontología y su poblamiento, utilizando un '*LLM juez*' para la evaluación. Un paradigma innovador es el de (Li et al., *KnowCoder*), que representa los esquemas de conocimiento como clases de Python, un formato que los *LLM* comprenden de forma nativa, y emplea un aprendizaje en dos fases para la comprensión y seguimiento del esquema. Por otro lado, (Luo et al., *OneKE*) diseña un sistema multi-agente (*Schema, Extraction, Reflection*) para guiar la extracción de conocimiento en diversos dominios y formatos de datos.

La aplicación de estas técnicas a dominios específicos y la extracción guiada es otra área clave. Para el dominio clínico, (Li et al., *Automated Clinical Data Extraction with Knowledge Conditioned LLMs*) utiliza *LLM* condicionados con conocimiento externo mediante aprendizaje en contexto (*in-context learning*) para extraer información de informes médicos. Para garantizar la coherencia con bases de conocimiento existentes, enfoques como los de (Feng et al., *Ontology-Grounded Automatic Knowledge Graph Construction by LLM under Wikidata Schema*) y (McCusker, *LOKE*) utilizan ontologías preexistentes (como la de *Wikidata*) para guiar y fundamentar el proceso de extracción del *LLM*. Finalmente, la interacción con el usuario se explora en (Abolhasani et al., *Leveraging LLM for Automated Ontology Extraction and Knowledge Graph Generation*), que presenta un *pipeline* interactivo guiado por un algoritmo adaptativo de Cadena de Pensamiento (*CoT*) para alinear la extracción de ontologías con los requisitos del usuario.

### 1.3.2. Creación de Enlaces

La creación de enlaces se refiere al establecimiento de conexiones significativas entre las piezas de conocimiento extraídas. Los *LLM* no solo extraen, sino que también pueden inferir y predecir estas relaciones.

Un paradigma dominante consiste en utilizar los *LLM* para la predicción de enlaces como una tarea de inferencia. En este sentido, (Shu et al., *Knowledge Graph Large Language Model (KG-LLM) for Link Prediction*) convierte la estructura del grafo en *prompts* de lenguaje

natural para afinar *LLM* en la tarea de predicción de enlaces multi-salto. De forma similar, (He et al., *LinkGPT*) entrena un *LLM* de extremo a extremo que integra información estructural pareada mediante un ajuste de instrucciones en dos etapas. Para abordar la escalabilidad en grafos de gran tamaño, (Bi et al., *LPNL*) propone un *pipeline* de muestreo en dos etapas y una estrategia de 'divide y vencerás' para gestionar la sobrecarga de información en los *prompts*.

Una estrategia diferente consiste en la creación de enlaces sin supervisión o con pocos ejemplos. El trabajo de (Carta et al., *Iterative Zero-Shot LLM Prompting for Knowledge Graph Construction*) destaca en este aspecto, utilizando un *pipeline* de *prompting* iterativo en modo *zero-shot*, que no requiere ejemplos ni recursos externos para construir el grafo de conocimiento.

Finalmente, otra línea de investigación utiliza los *LLM* como enriquecedores de características para modelos existentes. En lugar de predecir el enlace directamente, (Park et al., *Enhancing Future Link Prediction in Quantum Computing Semantic Networks through LLM-Initiated Node Features*) emplea un *LLM* para generar descripciones ricas de los nodos, que luego se utilizan como características iniciales para mejorar el rendimiento de los modelos de Redes Neuronales de Grafo (*GNN*) en la predicción de enlaces, especialmente en escenarios de arranque en frío (*cold-start*).

La combinación de estas dos capacidades —extracción y enlace— tiene el potencial de transformar la construcción de bases de conocimiento. Trabajos como los de (Zhu et al., «*LLMs for Knowledge Graph Construction and Reasoning*») y (Machado et al., «*{LLM Store}: Leveraging Large Language Models as Sources of {Wikidata}-Structured Knowledge*») exploran las capacidades generales de los *LLM* para estas tareas, el primero concluyendo que son mejores asistentes de inferencia que extractores, y el segundo implementando un *LLM* como un 'almacén' de conocimiento dinámico que sintetiza respuestas estructuradas al momento de la consulta. La investigación continúa abordando desafíos como la interpretabilidad, la escalabilidad y la superación de las limitaciones inherentes de los *LLM* mediante arquitecturas de agentes y la integración de herramientas externas.

#### **1.4. *LLM* en Aplicaciones de Toma de Notas**

Los avances en los *LLMs* han catalizado una diversidad de herramientas para automatizar y enriquecer la toma de notas y la gestión del conocimiento personal (*PKM*). Estas varían desde aplicaciones web hasta *plugins* para plataformas consolidadas, enfocándose en distintos casos de uso.

Un área prominente es la *asistencia conversacional* y la *respuesta a preguntas contextualizadas*. En *Logseq*, *plugins* como *Logseq Copilot* permiten interactuar con la IA, indexando las notas del usuario para ofrecer respuestas basadas en el propio conocimiento. En *Obsi-*

*dian.md*, herramientas como *BMO Chatbot for Obsidian* y *ChatGPT MD* ofrecen interfaces de chat versátiles, conectándose a servicios en la nube y LLMs locales, y permitiendo referenciar la nota actual o notas enlazadas en las conversaciones. *Caret Obsidian Plugin* extiende esta funcionalidad al *Canvas* para un chat no lineal. *Notion AI* también integra de forma nativa la capacidad de responder preguntas sobre el espacio de trabajo.

La *generación y mejora de contenido* es otro caso de uso fundamental. Herramientas como *LLM Summary* en *Obsidian* automatizan la creación de resúmenes de archivos *PDF* y la extracción de conceptos clave. Para la creación de material de estudio, *plugins* como *Obsidian Flashcards LLM* y *Quiz Generator* facilitan la generación de tarjetas de estudio y cuestionarios. En *Logseq*, *ollama-logseq* y *logseq-rag* permiten generar resúmenes y tarjetas de estudio. La mejora de la escritura y la reestructuración de notas se abordan con *Zettelkasten LLM Tools* y *Simple Prompt* en *Obsidian*, mientras que *LaTeX Generator Plugin for Obsidian* convierte lenguaje natural a ecuaciones *LaTeX*. Para la generación de contenido dentro de lienzos visuales, *Canvas LLM Extender* en *Obsidian* añade nodos de texto generados por IA.

En cuanto a la *organización, estructuración y enlace del conocimiento*, *InfraNodus* destaca como una aplicación web que visualiza texto como redes para identificar términos influyentes y lagunas conceptuales. En *Obsidian*, *ExMemo Tools* se enfoca en la gestión de documentos y la generación de metadatos como etiquetas, una tarea que *Obsidian LLM Tagger Plugin* también realiza usando LLMs locales. *InsightA Obsidian Plugin* transforma artículos largos en notas atómicas interconectadas y genera Mapas de Contenido (MOCs). *Obsidian Cloud Atlas Plugin* utiliza reconocimiento de entidades para crear *wikilinks* automáticamente, mejorando la interconexión.

La integración con LLMs locales para mayor privacidad y control es una tendencia creciente. En *Logseq*, *ollama-logseq* y *logseq-rag* se integran con *Ollama*. Para *Obsidian*, *Local LLM Helper - Obsidian Plugin* y *Obsidian AI plugin* conectan con servidores locales como *Ollama* o *LM Studio*. Muchas de las herramientas de chat y generación de contenido mencionadas también ofrecen la opción de operar con LLMs locales.

Finalmente, la *automatización de flujos de trabajo* y el RAG permiten interacciones más sofisticadas. *Logseq Composer* conecta notas con cualquier LLM mediante RAG a través de *LiteLLM*. En *Obsidian*, *Cannoli* permite construir *scripts* de LLM sin código en el editor *Canvas*, y *LLM Workspace plugin for Obsidian* permite crear conjuntos de fuentes curadas para fundamentar las conversaciones con IA mediante RAG. *Obsidian Cloud Atlas Plugin* también introduce flujos de trabajo en *Canvas* o *Markdown*.

Esta evolución subraya un esfuerzo por automatizar tareas, mejorar la interconexión de ideas y potenciar la utilidad de los sistemas de PKM, con una diversidad de enfoques que buscan adaptarse a las necesidades específicas de los usuarios.

## 1.5. Fundamentos y Alcance de la Investigación

El presente trabajo de tesis se enmarca en la intersección de la Gestión del Conocimiento Personal (PKM) y los avances en Inteligencia Artificial, específicamente los *LLM*. La investigación busca abordar los desafíos inherentes a la integración eficiente y significativa de nuevo conocimiento en bases de conocimiento personales semiestructuradas.

El problema científico que se aborda es la optimización del proceso de integración de conocimiento en dichas bases. Tradicionalmente, este proceso es manual y consume tiempo, especialmente al incorporar información de diversas fuentes y formatos, y al establecer conexiones relevantes. Se busca explorar cómo los *LLM* pueden automatizar y enriquecer esta tarea, facilitando la asimilación de información y la adaptación a diferentes paradigmas de toma de notas (e.g., *Zettelkasten*, notas conectadas, resúmenes progresivos). El objeto de estudio es, por tanto, el proceso de construcción y enriquecimiento automatizado de bases de conocimiento personal mediante *LLM*, centrándose el campo de acción en sistemas basados en lenguajes de marcado como *Markdown* y la representación del conocimiento mediante grafos, con especial atención a los *trabajadores del conocimiento*.

La investigación se guía por la pregunta: ¿En qué medida la aplicación de *LLMs*, a través de un *framework* de agentes personalizables, puede automatizar y optimizar la integración de conocimiento proveniente de diversas fuentes (predominantemente no estructuradas) en bases de conocimiento personales semiestructuradas (como las basadas en *Markdown* y grafos), y cómo se adapta esta automatización a diferentes paradigmas de toma de notas?

Para responder a esta pregunta, el *Objetivo General* es avanzar en la automatización de la construcción incremental y progresiva de bases de conocimiento personal, con un enfoque en la integración contextualizada de información mediante *LLM*. Los *Objetivos Específicos* son:

- a. Desarrollar un *framework* flexible capaz de procesar datos en múltiples formatos (e.g., texto, *PDF*, web), predominantemente no estructurados, para su integración en bases de conocimiento personales existentes (e.g., *Obsidian.md*).
- b. Diseñar e implementar un conjunto de agentes basados en *LLM* con distintas capacidades (resumen, extracción de entidades, generación de enlaces, creación de notas atómicas) para la integración de conocimiento.
- c. Evaluar la eficacia y adaptabilidad del sistema para soportar diversos paradigmas de toma de notas, analizando la coherencia, relevancia y utilidad del conocimiento integrado.

Este esfuerzo continúa la línea de investigación de trabajos como Fraga et al., «*On the Automatic Generation of Knowledge Connections*», pero se enfoca en el aprovechamiento de *LLM* para generar y enriquecer información directamente en el contexto de una base

de conocimiento personal existente o en desarrollo, particularmente para la creación y expansión de notas semiestructuradas en *Markdown* que forman un grafo de conocimiento dinámico. Hasta donde alcanza el conocimiento del autor, no existe una herramienta que combine integralmente *LLM* para la construcción progresiva y contextualizada de una base de conocimiento personal basada en notas interconectadas, considerando la diversidad de fuentes y paradigmas. Para abordar esta brecha, se ha desarrollado un prototipo que procesa archivos, interactúa con un *LLM* para extraer, sintetizar e integrar conocimiento en una base de notas (existente o nueva), generando nuevas notas, resúmenes, conexiones y metadatos. Este trabajo explorará la viabilidad, eficacia y desafíos de dicha automatización, buscando una contribución significativa a la *PKM* y a las aplicaciones prácticas de los *LLM*.



## Capítulo 2

# IImporter

Este capítulo presenta la propuesta de *IImporter*, un plugin para Obsidian diseñado para automatizar la integración de nuevo conocimiento en una bóveda existente. El sistema recibe un conjunto de archivos y, guiado por instrucciones del usuario, genera y vincula nuevas notas de manera inteligente.

El problema fundamental que se aborda puede definirse formalmente. Se construye un Agente,  $A$ , que opera sobre una bóveda de Obsidian. La función del agente es transformar la bóveda a un nuevo estado basándose en un conjunto de archivos de entrada y unas instrucciones específicas.

Sea  $V$  el estado de la bóveda, definido como un conjunto de notas en formato Markdown. Sea  $F = \{f_1, f_2, \dots, f_n\}$  un conjunto de archivos de entrada y sea  $I$  un conjunto de instrucciones en lenguaje natural que guían el proceso. La operación del agente se puede describir como una función:

$$A(F, I, V) \rightarrow V'$$

donde  $V'$  es la bóveda actualizada, que contiene nuevas notas, modificaciones a notas existentes y nuevos vínculos entre ellas, como resultado de la evaluación del agente.

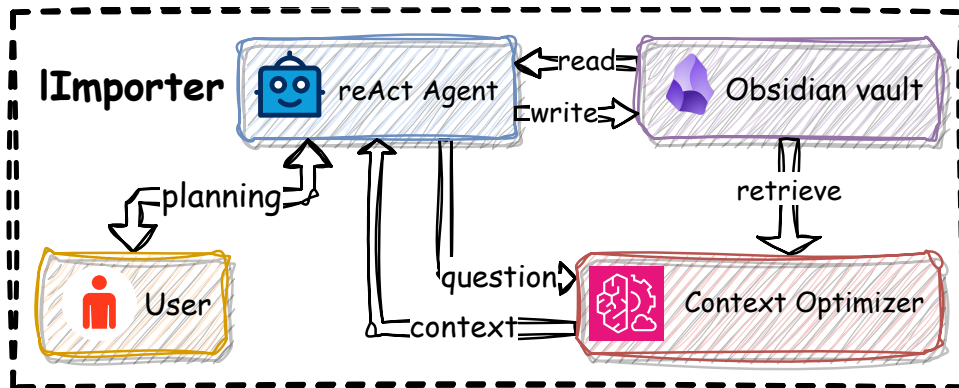


Figura 2.1. Esquema general del flujo de trabajo del agente *IImporter*.



## 2.1. Descripción del entorno

El entorno de trabajo para este proyecto es Obsidian. A diferencia de otras herramientas basadas en la nube, el modelo de datos de Obsidian es simple y robusto: una carpeta local en el sistema de archivos del usuario que contiene todos los datos. Las notas se almacenan como archivos de texto plano en formato Markdown (‘.md’), y pueden organizarse en una jerarquía de carpetas tradicional. Este enfoque garantiza la portabilidad, la longevidad y el control total del usuario sobre su información.

Una de las características más distintivas de Obsidian es su capacidad para visualizar las conexiones entre notas como un grafo de conocimiento. Cada nota es un nodo en el grafo, y los vínculos entre notas se representan como aristas. Esto permite al usuario descubrir relaciones emergentes y navegar por su conocimiento de una manera no lineal.

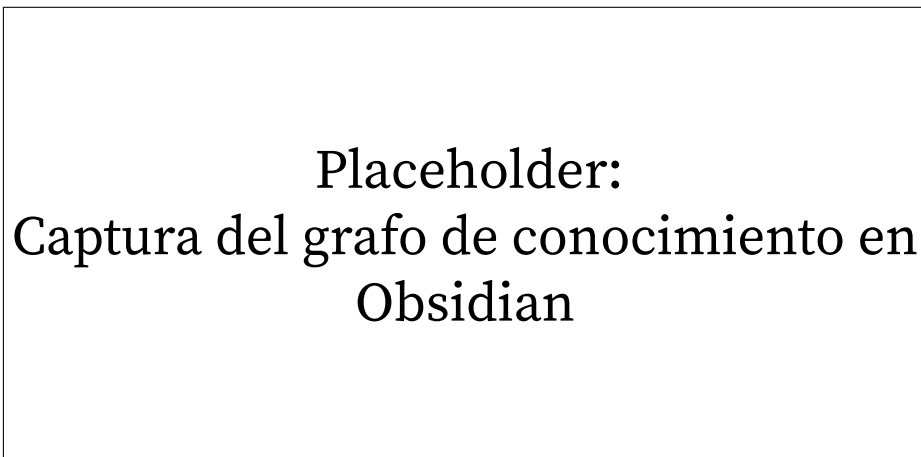


Figura 2.2. Ejemplo de la vista de grafo en Obsidian, mostrando nodos y sus interconexiones.

La creación de vínculos es fundamental para construir el grafo. En Obsidian, esto se logra mediante la sintaxis de *wikilinks*, simplemente encerrando el nombre de otra nota entre dobles corchetes, por ejemplo, ‘[[Nota Destino]]’. Además de los vínculos simples, Obsidian soporta el concepto de **transclusión**, que permite incrustar el contenido de una nota (o una sección de ella) dentro de otra usando la sintaxis ‘![[Nota a Incrustar]]’. La transclusión crea un vínculo compuesto y fuerte, ya que el contenido de una nota se convierte en parte integral de otra, fomentando la atomicidad y la reutilización de la información.

## 2.2. Detalles de Implementación

Para la implementación de *Importer*, se propone un agente autónomo basado en el paradigma ReAct (Reasoning and Acting). Este agente está dotado de un conjunto de

herramientas especializadas que le permiten interactuar directamente con la bóveda de Obsidian, leyendo su estructura, obteniendo el contenido de las notas y escribiendo nueva información.

La explicación de la arquitectura se realizará de manera *top-down*. Se comenzará describiendo el funcionamiento general del agente, para luego profundizar en cada uno de sus componentes clave: el mecanismo de interacción con la bóveda y las estrategias para la obtención de contexto relevante.

### 2.2.1. Agente reAct

En el núcleo del sistema se encuentra un agente basado en el framework ReAct (Yao et al., [ReAct](#)). Este paradigma permite a los modelos de lenguaje combinar el razonamiento (thought) con la acción (action). El agente opera en un ciclo donde, a partir de una tarea y una observación del entorno, razona sobre el siguiente paso a seguir, elige una herramienta de su repertorio, la ejecuta, y observa el resultado para informar su siguiente ciclo de razonamiento.

Siguiendo las mejores prácticas observadas en agentes avanzados como los utilizados para investigación o codificación (e.g. *Deep Research*), se ha adoptado la estrategia de *Plan-and-Solve* (PS) (Wang et al., [Plan-and-Solve Prompting](#)). En lugar de actuar de forma impulsiva, el agente primero genera un plan detallado y paso a paso para resolver la tarea. Una característica distintiva de nuestra implementación es la introducción de un punto de control humano: el usuario tiene la opción de revisar, confirmar y dar retroalimentación sobre el plan generado antes de su ejecución. Este enfoque de *human-in-the-loop* combina la eficiencia del planeamiento autónomo con la supervisión y el direccionamiento estratégico del usuario.

Para la implementación del agente se utiliza la familia de modelos Gemini de Google, debido a su alta capacidad, accesibilidad y, de manera crucial, su soporte nativo para *function calling*, un mecanismo que se detallará a continuación.

### 2.2.2. Interacción con la bóveda

La capacidad del agente para leer y escribir archivos en la bóveda es posible gracias al uso de la API de Obsidian, orquestada a través de *Function Calling*. Esta técnica puede ser entendida como un caso particular de *style prompting* que emplea decodificación restringida (*constrained decoding*) para forzar al modelo de lenguaje a generar una salida que se adhiere estrictamente a una gramática o formato predefinido Geng et al., [Grammar-Constrained Decoding for Structured NLP Tasks without Finetuning](#).

En el caso de la API de Gemini, esta restricción obliga al modelo a producir una salida en formato JSON que se corresponde con la firma de una de las funciones (herramientas)

disponibles. Se aprovecha esta capacidad para mejorar la fiabilidad del agente. Por ejemplo, al definir los parámetros de una función, se puede utilizar el tipo de dato 'ENUM' para restringir las entradas a un conjunto de valores válidos (e.g., una lista de carpetas o archivos existentes), evitando así que el modelo intente operar sobre rutas inválidas o "alucinadas".

## **Lectura**

Para la lectura de información de la bóveda, el agente dispone de dos herramientas principales:

- `tree(path)`: Esta función recibe la ruta a una carpeta raíz y devuelve una representación textual de la estructura de directorios y archivos contenida en ella, de forma análoga al comando `tree` de la línea de comandos. Esto permite al agente obtener una visión global de la organización de la bóveda.
- `read(path)`: Esta función recibe la ruta a un archivo Markdown ('.md') existente y devuelve su contenido completo como texto.

## **Escritura**

Un desafío común en los agentes que interactúan con sistemas de archivos es la tendencia del modelo a generar rutas de archivo incorrectas o con nombres similares pero no idénticos a los existentes. Para mitigar este problema, la capacidad de escritura se ha separado en dos funciones distintas:

- `mkdir(path)`: Permite al agente crear un nuevo directorio en una ubicación específica de la bóveda, asegurando que la estructura de carpetas deseada exista antes de intentar escribir un archivo.
- `write(path, content)`: Crea un nuevo archivo o sobrescribe el contenido de uno existente. El parámetro `path` de esta función se beneficia directamente de la decodificación restringida, ya que se puede guiar al modelo para que elija entre rutas sugeridas o siga un formato válido, reduciendo drásticamente los errores.

## **Obtención de contexto**

Para que el agente pueda tomar decisiones informadas, como determinar dónde crear una nueva nota o con qué notas existentes vincularla, necesita un contexto relevante de la bóveda. Dado que el contenido total de la bóveda puede exceder fácilmente la ventana de contexto del modelo, se utiliza un enfoque de "divide y vencerás".

Se propone una función que recibe un conjunto de archivos y un límite de tokens. Si el contenido combinado de los archivos excede el límite, el conjunto se divide recursivamente por la mitad hasta que los fragmentos resultantes son lo suficientemente pequeños. Una

vez que un fragmento cabe en la ventana de contexto, un modelo de lenguaje se encarga de extraer la información relevante de él. La naturaleza de esta "información relevante" se define por un nivel de **granularidad** especificado por el usuario Chen et al., «[Dense X Retrieval](#)».

Esta idea está inspirada en la técnica de Comprensión y Recuperación de Pasajes Pequeños (CPRS) de Qwen Shen et al., [QwenLong-CPRS](#). De manera similar, se proponen tres niveles de granularidad:

- **Paragraph:** Ideal para obtener resúmenes y el sentido general de un documento.
- **Sentence:** Extremadamente útil para identificar afirmaciones específicas y recuperar relaciones entre conceptos.
- **Keyword:** Perfecto para la extracción de entidades nombradas y conceptos clave.

A continuación se presenta un pseudocódigo del proceso de obtención de contexto:

```
function getContext(files, token_limit, granularity):
    if total_tokens(files) <= token_limit:
        return extract_info(files, granularity)
    else:
        mid = floor(len(files) / 2)
        part1 = files[0:mid]
        part2 = files[mid:len(files)]

        context1 = getContext(part1, token_limit, granularity)
        context2 = getContext(part2, token_limit, granularity)

        return combine(context1, context2)
```

### 2.2.3. Sobre la versatilidad del sistema

Un objetivo clave de *Importer* es su versatilidad para integrarse en diversos flujos de trabajo. El agente debe ser útil tanto para un usuario que desea construir una base de conocimiento densamente interconectada, generando múltiples relaciones y resúmenes, como para otro que simplemente quiere archivar un nuevo concepto de forma aislada, sin crear ningún vínculo.

La arquitectura basada en un agente ReAct y el sistema de obtención de contexto granular sustentan esta flexibilidad. El agente puede recibir instrucciones para ser exhaustivo en la búsqueda de conexiones o, por el contrario, para limitarse a crear una nota y guardarla en una carpeta específica. Sin embargo, esta versatilidad conlleva un incremento en la labor manual de configuración inicial, ya que el comportamiento del agente se define en gran medida a través de un archivo de instrucciones proporcionado por el usuario.

## Capítulo 3

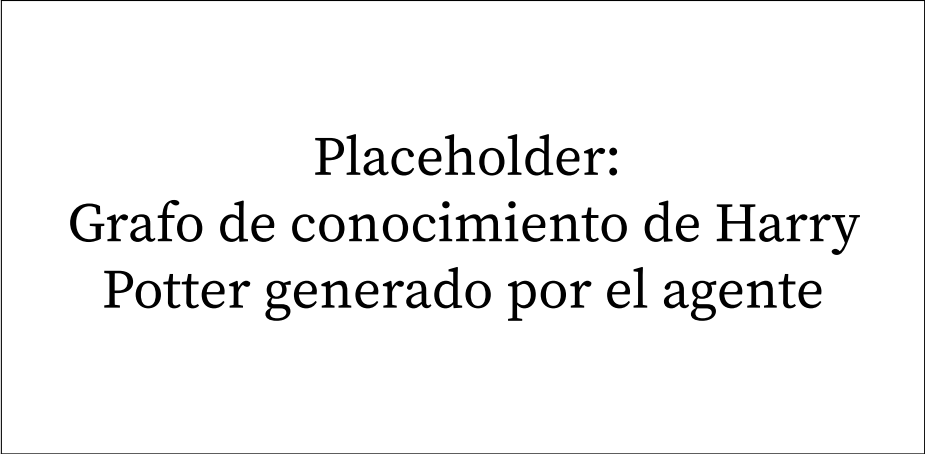
# Experimentos

En este capítulo se presentan una serie de experimentos diseñados para evaluar el rendimiento y las capacidades del agente *Importer*. Es importante aclarar que no se realizó un análisis holístico y cuantitativo del sistema. En su lugar, se optó por explorar distintos casos de uso que demuestran la versatilidad de la herramienta y ponen de manifiesto las interesantes posibilidades que surgen al integrar un agente de lenguaje autónomo en una bóveda de conocimiento de Obsidian.

### 3.1. Base estructurada sobre Harry Potter

En el primer experimento, se utilizó la versatilidad del agente para una tarea clásica: la creación de una base de conocimientos estructurada sobre el universo de Harry Potter. El objetivo era proporcionar al agente los textos de los libros y pedirle que generara notas para personajes, lugares y conceptos clave, vinculándolos entre sí. Este caso de uso potencia a Obsidian como una herramienta para la exploración visual del conocimiento a través de su vista de grafo.

Es relevante mencionar que, por defecto, Obsidian no muestra etiquetas en las aristas del grafo, lo que dificulta la interpretación de la naturaleza de las relaciones. Para solventar esto, se utilizó un plugin de la comunidad que añade dicha funcionalidad. El grafo resultante, como se muestra en la Figura 3.1, fue creado exitosamente. Una prueba de la correcta identificación de la relevancia de los conceptos se observa en el tamaño de los nodos: los nodos correspondientes a "Harry Potter" y "Hogwarts" son visiblemente más grandes, indicando un mayor número de conexiones y, por tanto, una mayor centralidad en la narrativa.



## Placeholder: Grafo de conocimiento de Harry Potter generado por el agente

Figura 3.1. Grafo resultante del experimento de Harry Potter, mostrando la centralidad de nodos clave.

### 3.2. Añadiendo artículos a una *Wiki* tecnológica

Este experimento se diseñó para simular el mantenimiento y crecimiento de una base de conocimiento existente. Se partió de una bóveda de Obsidian que funcionaba como una *wiki* personal sobre temas de tecnología (programación, inteligencia artificial, etc.). Progresivamente, se le proporcionaron al agente nuevos artículos, acompañados de un breve resumen en audio describiendo su contenido.


El objetivo era doble. Por un lado, se buscaba demostrar que el agente es capaz de analizar el nuevo contenido, comprenderlo en el contexto de la bóveda existente y determinar correctamente con qué notas preexistentes debía vincularlo (Figura 3.2). Por otro lado, y de igual importancia, se quería verificar que el agente no siempre establece conexiones. Si un nuevo artículo trata sobre un tema completamente ajeno a lo que ya existe en la bóveda, el resultado deseable es que se cree una nota aislada, sin vínculos forzados o incorrectos (Figura 3.3). Este comportamiento es crucial para mantener la integridad y la calidad del grafo de conocimiento.

### 3.3. Projects, Areas, Resources, Archived

Este experimento refleja un caso de uso más cotidiano y de productividad personal, siguiendo la metodología P.A.R.A. (Projects, Areas, Resources, Archived) popularizada por Tiago Forte en su libro *Building a Second Brain* Forte, *Building a Second Brain*. El objetivo era evaluar si el sistema puede ser utilizado para mantener una bóveda organizada bajo esta estructura, siguiendo instrucciones de voz.

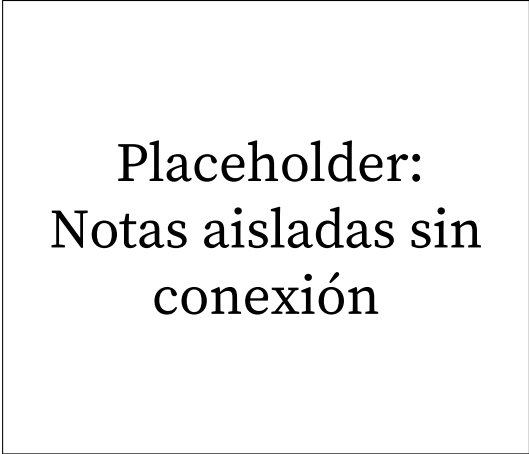
Se probó la capacidad del agente para realizar tareas como:

- Incluir nuevos archivos (recursos) en la bóveda, clasificándolos correctamente en la



Placeholder:  
Cluster de notas  
relacionadas

Figura 3.2. Nuevas notas integradas y conectadas con el conocimiento existente en la wiki.



Placeholder:  
Notas aisladas sin  
conexión

Figura 3.3. Nuevas notas añadidas que permanecen como "islas" de conocimiento, al no tener relación directa con el contenido previo.

carpeta 'Resources'.

- Vincular estos nuevos recursos con notas existentes en las carpetas 'Projects' o 'Areas'.
- Sintetizar información a partir de varias notas existentes para generar un nuevo resumen o una nota de trabajo.

Este escenario pone a prueba no solo la interacción con el sistema de archivos, sino también la capacidad del agente para comprender la intención del usuario y ejecutar flujos de trabajo más complejos y estructurados.

### **3.4. Note augmented *LLMs* are computationally universal**

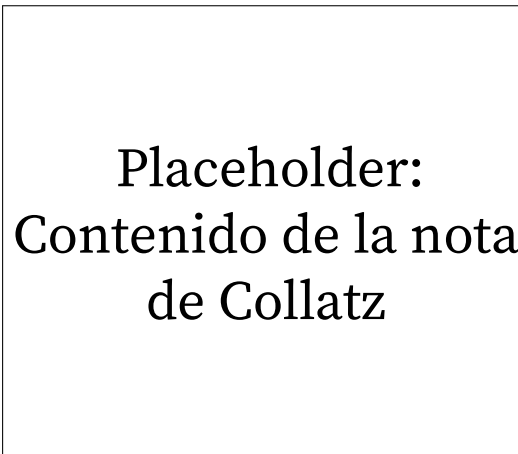
Este último experimento, de naturaleza más teórica, se inspira en la idea de que los modelos de lenguaje, cuando se aumentan con una memoria externa, pueden volverse

computacionalmente universales (Schuurmans, *Memory Augmented Large Language Models Are Computationally Universal*).

Para explorar esta idea, se le pidió al agente que evaluara la Conjetura de Collatz para un número dado, utilizando una nota de Obsidian como su "memoria de trabajo." registro. En cada paso de la secuencia, el agente leía el número actual de la nota, calculaba el siguiente número, y sobrescribía la nota con el nuevo valor.

Se evaluó el caso para  $n = 10$ , que se resolvió correctamente en 6 iteraciones (Figuras 3.4 y 3.5). También se evaluaron casos más largos como  $n = 11$  (14 iteraciones) y  $n = 27$  (111 iteraciones). Este último caso excedió los límites de cuota de la API y requirió reiniciar el proceso, lo cual demostró una capacidad emergente del sistema: la posibilidad de "dejar una tarea y seguirla luego", ya que el estado se preserva en la nota de Obsidian.

La intuición subyacente es que la capacidad de releer el estado desde un archivo fiable reduce drásticamente la probabilidad de error en un proceso iterativo largo, de manera análoga a cómo una CPU depende de leer y escribir correctamente en sus registros. En este caso se usó un modelo potente en razonamiento algebraico (Gemini 2.5 Flash). Si bien un modelo más modesto podría fallar en esta tarea matemática, la misma arquitectura podría permitirle automatizar con alta fiabilidad otra tarea iterativa que se alinee con sus fortalezas (por ejemplo, un análisis de sentimiento refinado a lo largo de múltiples revisiones), siempre que sea proficiente en las operaciones básicas de leer y escribir en su memoria externa.

A rectangular box with a thin black border, containing the text 'Placeholder: Contenido de la nota de Collatz' in a serif font, centered within the box.

Placeholder:  
Contenido de la nota  
de Collatz

Figura 3.4. Estado de la nota de trabajo durante la evaluación de Collatz.



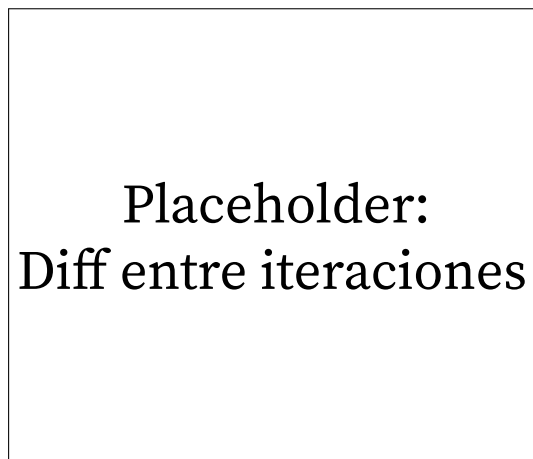


Figura 3.5. Diferencia en el contenido de la nota entre dos pasos de la iteración.

# Bibliografía

- Ahrens, Sönke. *How to Take Smart Notes: One Simple Technique to Boost Writing, Learning and Thinking – for Students, Academics and Nonfiction Book Writers*. Feb. de 2017 (ver páginas [1](#), [2](#)).
- Forte, Tiago. *Building a Second Brain*. Profile, jun. de 2022 (ver páginas [1](#), [2](#), [19](#)).
- Grundspenkis, Janis. «Agent Based Approach for Organization and Personal Knowledge Modelling: Knowledge Management Perspective». En: *Journal of Intelligent Manufacturing* 18.4 (jul. de 2007), págs. 451-457. ISSN: 0956-5515, 1572-8145. DOI: [10.1007/s10845-007-0052-6](#) (ver página [1](#)).
- Jansen, Renée S., Daniel Lakens y Wijnand A. IJsselsteijn. «An Integrative Review of the Cognitive Costs and Benefits of Note-Taking». En: *Educational Research Review* 22 (nov. de 2017), págs. 223-233. ISSN: 1747938X. DOI: [10.1016/j.edurev.2017.10.001](#) (ver página [2](#)).
- Pauk, Walter y Ross J. Q. Owens. *How to Study in College*. 10th Ed. Boston, MA: Cengage Learning, 2010. ISBN: 978-1-4390-8446-5 978-0-495-90302-4 (ver página [2](#)).
- Bush, Vannevar. *As We May Think*. Ene. de 1945 (ver página [3](#)).
- Davies, Stephen, Javier Velez-Morales y Roger King. «Building the Memex Sixty Years Later: Trends and Directions in Personal Knowledge Bases». En: (2005) (ver página [3](#)).
- Davies, Stephen. «Still Building the Memex». En: *Communications of the ACM* 54.2 (feb. de 2011), págs. 80-88. ISSN: 0001-0782, 1557-7317. DOI: [10.1145/1897816.1897840](#) (ver página [3](#)).
- Pyne, Yvette y Stuart Stewart. «Meta-Work: How We Research Is as Important as What We Research». En: *The British Journal of General Practice* 72.716 (feb. de 2022), págs. 130-131. ISSN: 0960-1643. DOI: [10.3399/bjgp22X718757](#) (ver página [3](#)).
- OpenAI et al. *GPT-4 Technical Report*. Mar. de 2024. DOI: [10.48550/arXiv.2303.08774](#). arXiv: [2303.08774 \[cs\]](#) (ver página [4](#)).
- Team, Gemini et al. *Gemini: A Family of Highly Capable Multimodal Models*. Jun. de 2024. DOI: [10.48550/arXiv.2312.11805](#). arXiv: [2312.11805 \[cs\]](#) (ver página [4](#)).
- DeepSeek-AI et al. *DeepSeek-V3 Technical Report*. Dic. de 2024. DOI: [10.48550/arXiv.2412.19437](#). arXiv: [2412.19437 \[cs\]](#) (ver página [4](#)).

- Bai, Jinze et al. *Qwen Technical Report*. Sep. de 2023. DOI: [10.48550/arXiv.2309.16609](https://doi.org/10.48550/arXiv.2309.16609). arXiv: [2309.16609](https://arxiv.org/abs/2309.16609) [cs] (ver página 4).
- Grattafiori, Aaron et al. *The Llama 3 Herd of Models*. Nov. de 2024. DOI: [10.48550/arXiv.2407.21783](https://doi.org/10.48550/arXiv.2407.21783). arXiv: [2407.21783](https://arxiv.org/abs/2407.21783) [cs] (ver página 4).
- Brown, Tom B. et al. «Language Models Are Few-Shot Learners». En: *Proceedings of the 34th International Conference on Neural Information Processing Systems*. NIPS '20. Red Hook, NY, USA: Curran Associates Inc., dic. de 2020, págs. 1877-1901. ISBN: 978-1-7138-2954-6 (ver página 4).
- Min, Sewon et al. «Rethinking the Role of Demonstrations: What Makes In-Context Learning Work?» En: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Ed. por Yoav Goldberg, Zornitsa Kozareva y Yue Zhang. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, dic. de 2022, págs. 11048-11064. DOI: [10.18653/v1/2022.emnlp-main.759](https://doi.org/10.18653/v1/2022.emnlp-main.759) (ver página 4).
- Kong, Aobo et al. *Better Zero-Shot Reasoning with Role-Play Prompting*. Mar. de 2024. DOI: [10.48550/arXiv.2308.07702](https://doi.org/10.48550/arXiv.2308.07702). arXiv: [2308.07702](https://arxiv.org/abs/2308.07702) [cs] (ver página 4).
- Lu, Albert et al. *Bounding the Capabilities of Large Language Models in Open Text Generation with Prompt Constraints*. Feb. de 2023. DOI: [10.48550/arXiv.2302.09185](https://doi.org/10.48550/arXiv.2302.09185). arXiv: [2302.09185](https://arxiv.org/abs/2302.09185) [cs] (ver página 4).
- Nye, Maxwell et al. *Show Your Work: Scratchpads for Intermediate Computation with Language Models*. Nov. de 2021. DOI: [10.48550/arXiv.2112.00114](https://doi.org/10.48550/arXiv.2112.00114). arXiv: [2112.00114](https://arxiv.org/abs/2112.00114) [cs] (ver página 5).
- Wei, Jason et al. *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models*. Ene. de 2023. DOI: [10.48550/arXiv.2201.11903](https://doi.org/10.48550/arXiv.2201.11903). arXiv: [2201.11903](https://arxiv.org/abs/2201.11903) [cs] (ver página 5).
- Kojima, Takeshi et al. *Large Language Models Are Zero-Shot Reasoners*. Ene. de 2023. DOI: [10.48550/arXiv.2205.11916](https://doi.org/10.48550/arXiv.2205.11916). arXiv: [2205.11916](https://arxiv.org/abs/2205.11916) [cs] (ver página 5).
- Wang, Lei et al. *Plan-and-Solve Prompting: Improving Zero-Shot Chain-of-Thought Reasoning by Large Language Models*. Mayo de 2023. DOI: [10.48550/arXiv.2305.04091](https://doi.org/10.48550/arXiv.2305.04091). arXiv: [2305.04091](https://arxiv.org/abs/2305.04091) [cs] (ver páginas 5, 15).
- Fu, Yao et al. *Complexity-Based Prompting for Multi-Step Reasoning*. Ene. de 2023. DOI: [10.48550/arXiv.2210.00720](https://doi.org/10.48550/arXiv.2210.00720). arXiv: [2210.00720](https://arxiv.org/abs/2210.00720) [cs] (ver página 5).
- Zhou, Denny et al. *Least-to-Most Prompting Enables Complex Reasoning in Large Language Models*. Abr. de 2023. DOI: [10.48550/arXiv.2205.10625](https://doi.org/10.48550/arXiv.2205.10625). arXiv: [2205.10625](https://arxiv.org/abs/2205.10625) [cs] (ver página 5).
- Wang, Xuezhi et al. *Self-Consistency Improves Chain of Thought Reasoning in Language Models*. Mar. de 2023. DOI: [10.48550/arXiv.2203.11171](https://doi.org/10.48550/arXiv.2203.11171). arXiv: [2203.11171](https://arxiv.org/abs/2203.11171) [cs] (ver página 5).

- Yao, Shunyu et al. *Tree of Thoughts: Deliberate Problem Solving with Large Language Models*. Dic. de 2023. DOI: [10.48550/arXiv.2305.10601](#). arXiv: [2305.10601 \[cs\]](#) (ver página 5).
- Besta, Maciej et al. «Graph of Thoughts: Solving Elaborate Problems with Large Language Models». En: *Proceedings of the AAAI Conference on Artificial Intelligence* 38.16 (mar. de 2024), págs. 17682-17690. ISSN: 2374-3468, 2159-5399. DOI: [10.1609/aaai.v38i16.29720](#). arXiv: [2308.09687 \[cs\]](#) (ver página 5).
- Press, Ofir et al. *Measuring and Narrowing the Compositionality Gap in Language Models*. Oct. de 2023. DOI: [10.48550/arXiv.2210.03350](#). arXiv: [2210.03350 \[cs\]](#) (ver página 5).
- Madaan, Aman et al. *Self-Refine: Iterative Refinement with Self-Feedback*. Mayo de 2023. DOI: [10.48550/arXiv.2303.17651](#). arXiv: [2303.17651 \[cs\]](#) (ver página 5).
- Zheng, Huaixiu Steven et al. *Take a Step Back: Evoking Reasoning via Abstraction in Large Language Models*. Mar. de 2024. DOI: [10.48550/arXiv.2310.06117](#). arXiv: [2310.06117 \[cs\]](#) (ver página 5).
- Gao, Peizhong et al. *Meta Reasoning for Large Language Models*. Jun. de 2024. DOI: [10.48550/arXiv.2406.11698](#). arXiv: [2406.11698 \[cs\]](#) (ver página 5).
- Aytes, Simon A., Jinheon Baek y Sung Ju Hwang. *Sketch-of-Thought: Efficient LLM Reasoning with Adaptive Cognitive-Inspired Sketching*. Mar. de 2025. DOI: [10.48550/arXiv.2503.05179](#). arXiv: [2503.05179 \[cs\]](#) (ver página 5).
- Zhang, Yifan et al. *Cumulative Reasoning with Large Language Models*. Mar. de 2025. DOI: [10.48550/arXiv.2308.04371](#). arXiv: [2308.04371 \[cs\]](#) (ver página 5).
- Liu, Nelson F et al. «Lost in the Middle: How Language Models Use Long Contexts». En: (2024) (ver página 5).
- Lewis, Patrick et al. *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*. Abr. de 2021. DOI: [10.48550/arXiv.2005.11401](#). arXiv: [2005.11401 \[cs\]](#) (ver página 6).
- Team, Gemini et al. *Gemini 1.5: Unlocking Multimodal Understanding across Millions of Tokens of Context*. Dic. de 2024. DOI: [10.48550/arXiv.2403.05530](#). arXiv: [2403.05530 \[cs\]](#) (ver página 6).
- Yang, An et al. *Qwen2.5-1M Technical Report*. Ene. de 2025. DOI: [10.48550/arXiv.2501.15383](#). arXiv: [2501.15383 \[cs\]](#) (ver página 6).
- Agarwal, Rishabh et al. *Many-Shot In-Context Learning*. Oct. de 2024. DOI: [10.48550/arXiv.2404.11018](#). arXiv: [2404.11018 \[cs\]](#) (ver página 6).
- Chen, Tong et al. «Dense X Retrieval: What Retrieval Granularity Should We Use?» En: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Ed. por Yaser Al-Onaizan, Mohit Bansal y Yun-Nung Chen. Miami, Florida, USA:

- Association for Computational Linguistics, nov. de 2024, págs. 15159-15177. DOI: [10.18653/v1/2024.emnlp-main.845](#) (ver páginas 6, 17).
- Shen, Weizhou et al. *QwenLong-CPRS: Towards  $\infty$ -LLMs with Dynamic Context Optimization*. Mayo de 2025. DOI: [10.48550/arXiv.2505.18092](#). arXiv: [2505.18092 \[cs\]](#) (ver páginas 6, 17).
- Zhuang, Yufan et al. *Self-Taught Agentic Long Context Understanding*. Feb. de 2025. DOI: [10.48550/arXiv.2502.15920](#). arXiv: [2502.15920 \[cs\]](#) (ver página 6).
- Sun, Zhiqing et al. *Recitation-Augmented Language Models*. Feb. de 2023. DOI: [10.48550/arXiv.2210.01296](#). arXiv: [2210.01296 \[cs\]](#) (ver página 6).
- Karpas, Ehud et al. *MRKL Systems: A Modular, Neuro-Symbolic Architecture That Combines Large Language Models, External Knowledge Sources and Discrete Reasoning*. Mayo de 2022. DOI: [10.48550/arXiv.2205.00445](#). arXiv: [2205.00445 \[cs\]](#) (ver página 6).
- Gao, Luyu et al. *PAL: Program-aided Language Models*. Ene. de 2023. DOI: [10.48550/arXiv.2211.10435](#). arXiv: [2211.10435 \[cs\]](#) (ver página 7).
- Gou, Zhibin et al. *ToRA: A Tool-Integrated Reasoning Agent for Mathematical Problem Solving*. Feb. de 2024. DOI: [10.48550/arXiv.2309.17452](#). arXiv: [2309.17452 \[cs\]](#) (ver página 7).
- Yao, Shunyu et al. *ReAct: Synergizing Reasoning and Acting in Language Models*. Mar. de 2023. DOI: [10.48550/arXiv.2210.03629](#). arXiv: [2210.03629 \[cs\]](#) (ver páginas 7, 15).
- Shinn, Noah et al. *Reflexion: Language Agents with Verbal Reinforcement Learning*. Mar. de 2023. DOI: [10.48550/arXiv.2303.11366](#). arXiv: [2303.11366](#) (ver página 7).
- Sumers, Theodore R. et al. *Cognitive Architectures for Language Agents*. Mar. de 2024. DOI: [10.48550/arXiv.2309.02427](#). arXiv: [2309.02427 \[cs\]](#) (ver página 7).
- Van Durme, Benjamin y Lenhart Schubert. «Open Knowledge Extraction through Compositional Language Processing». En: *Semantics in Text Processing. STEP 2008 Conference Proceedings*. Ed. por Johan Bos y Rodolfo Delmonte. College Publications, 2008, págs. 239-254 (ver página 7).
- Song, Linfeng et al. «OpenFact: Factuality Enhanced Open Knowledge Extraction». En: *Transactions of the Association for Computational Linguistics* 11 (jun. de 2023), págs. 686-702. ISSN: 2307-387X. DOI: [10.1162/tac1\\_a\\_00569](#) (ver página 7).
- Qian, Kun et al. *Open Domain Knowledge Extraction for Knowledge Graphs*. Oct. de 2023. DOI: [10.48550/arXiv.2312.09424](#). arXiv: [2312.09424 \[cs\]](#) (ver página 8).
- Zhang, Bowen y Harold Soh. *Extract, Define, Canonicalize: An LLM-based Framework for Knowledge Graph Construction*. Oct. de 2024. DOI: [10.48550/arXiv.2404.03868](#). arXiv: [2404.03868 \[cs\]](#) (ver página 8).
- Kommineni, Vamsi Krishna, Birgitta König-Ries y Sheeba Samuel. *From Human Experts to Machines: An LLM Supported Approach to Ontology and Knowledge Graph Construction*.

- Mar. de 2024. DOI: [10.48550/arXiv.2403.08345](#). arXiv: [2403.08345 \[cs\]](#) (ver página 8).
- Li, Zixuan et al. *KnowCoder: Coding Structured Knowledge into LLMs for Universal Information Extraction*. Mar. de 2024. DOI: [10.48550/arXiv.2403.07969](#). arXiv: [2403.07969 \[cs\]](#) (ver página 8).
- Luo, Yujie et al. *OneKE: A Dockerized Schema-Guided LLM Agent-based Knowledge Extraction System*. Feb. de 2025. DOI: [10.48550/arXiv.2412.20005](#). arXiv: [2412.20005 \[cs\]](#) (ver página 8).
- Li, Diya et al. *Automated Clinical Data Extraction with Knowledge Conditioned LLMs*. Nov. de 2024. DOI: [10.48550/arXiv.2406.18027](#). arXiv: [2406.18027 \[cs\]](#) (ver página 8).
- Feng, Xiaohan, Xixin Wu y Helen Meng. *Ontology-Grounded Automatic Knowledge Graph Construction by LLM under Wikidata Schema*. Dic. de 2024. DOI: [10.48550/arXiv.2412.20942](#). arXiv: [2412.20942 \[cs\]](#) (ver página 8).
- McCusker, Jamie. *LOKE: Linked Open Knowledge Extraction for Automated Knowledge Graph Construction*. Nov. de 2023. DOI: [10.48550/arXiv.2311.09366](#). arXiv: [2311.09366 \[cs\]](#) (ver página 8).
- Abolhasani, Mohammad Sadeq y Rong Pan. *Leveraging LLM for Automated Ontology Extraction and Knowledge Graph Generation*. Dic. de 2024. DOI: [10.48550/arXiv.2412.00608](#). arXiv: [2412.00608 \[cs\]](#) (ver página 8).
- Shu, Dong et al. *Knowledge Graph Large Language Model (KG-LLM) for Link Prediction*. Ago. de 2024. DOI: [10.48550/arXiv.2403.07311](#). arXiv: [2403.07311 \[cs\]](#) (ver página 8).
- He, Zhongmou et al. *LinkGPT: Teaching Large Language Models To Predict Missing Links*. Jun. de 2024. DOI: [10.48550/arXiv.2406.04640](#). arXiv: [2406.04640 \[cs\]](#) (ver página 9).
- Bi, Baolong et al. *LPNL: Scalable Link Prediction with Large Language Models*. Feb. de 2024. DOI: [10.48550/arXiv.2401.13227](#). arXiv: [2401.13227 \[cs\]](#) (ver página 9).
- Carta, Salvatore et al. *Iterative Zero-Shot LLM Prompting for Knowledge Graph Construction*. Jul. de 2023. DOI: [10.48550/arXiv.2307.01128](#). arXiv: [2307.01128 \[cs\]](#) (ver página 9).
- Park, Gilchan et al. *Enhancing Future Link Prediction in Quantum Computing Semantic Networks through LLM-Initiated Node Features*. Oct. de 2024. DOI: [10.48550/arXiv.2410.04251](#). arXiv: [2410.04251 \[cs\]](#) (ver página 9).
- Zhu, Yuqi et al. «LLMs for Knowledge Graph Construction and Reasoning: Recent Capabilities and Future Opportunities». En: *World Wide Web* 27.5 (sep. de 2024), pág. 58. ISSN: 1386-145X, 1573-1413. DOI: [10.1007/s11280-024-01297-w](#) (ver página 9).

- Machado, Marcelo et al. «{LLM Store}: Leveraging Large Language Models as Sources of {Wikidata}-Structured Knowledge». En: (2024) (ver página 9).
- Fraga, Felipe et al. «On the Automatic Generation of Knowledge Connections:» en: *Proceedings of the 25th International Conference on Enterprise Information Systems*. Prague, Czech Republic: SCITEPRESS - Science and Technology Publications, 2023, págs. 43-54. ISBN: 978-989-758-648-4. DOI: [10.5220/0011781100003467](https://doi.org/10.5220/0011781100003467) (ver página 11).
- Geng, Saibo et al. *Grammar-Constrained Decoding for Structured NLP Tasks without Finetuning*. Ene. de 2024. DOI: [10.48550/arXiv.2305.13971](https://doi.org/10.48550/arXiv.2305.13971). arXiv: [2305.13971](https://arxiv.org/abs/2305.13971) [cs] (ver página 15).
- Schuurmans, Dale. *Memory Augmented Large Language Models Are Computationally Universal*. Ene. de 2023. DOI: [10.48550/arXiv.2301.04589](https://doi.org/10.48550/arXiv.2301.04589). arXiv: [2301.04589](https://arxiv.org/abs/2301.04589) [cs] (ver página 21).