Reporte de Avance

1INF42 Proyecto de Fin de Carrera 1

Saymon Nicho

PUCP

12 de mayo de 2025

Contenidos

1. Generalidades

Introducción

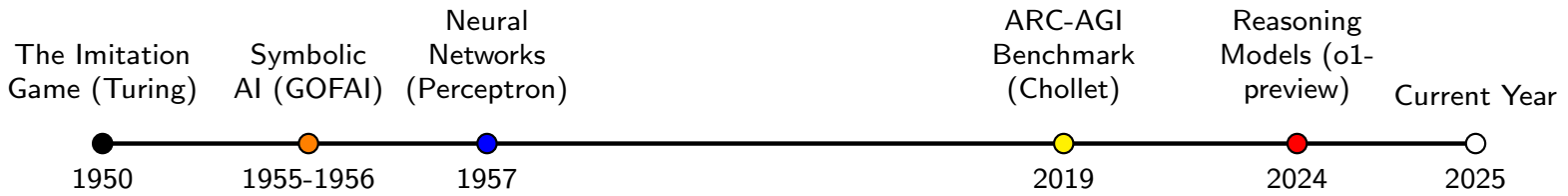

Figura 1: Historia de la IA

Propuesto por **Francois Chollet** en 2019 como un benchmark para **evaluar el razonamiento abstracto** en modelos de IA [4].

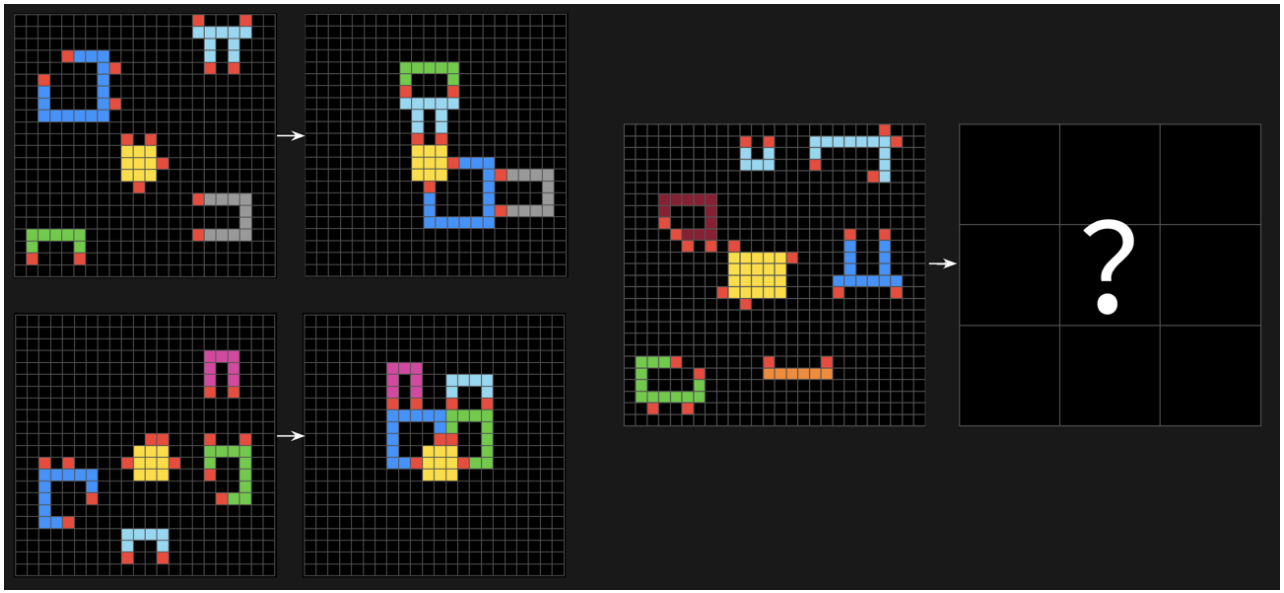

Figura 2: Benchmark ARC-AGI

Tema de Tesis

Resolución neuro-simbólica del benchmark ARC-AGI usando LLMs ligeros y representación simbólica eficiente

Asesor: César Beltrán, PhD

Proponer un **modelo ligero de IA** con **enfoque neuro-simbólico**, de modo que sirva para resolver problemas que involucren razonamiento abstracto de forma eficiente, tomando como referencia el **benchmark ARC-AGI** para su evaluación.

Problemática

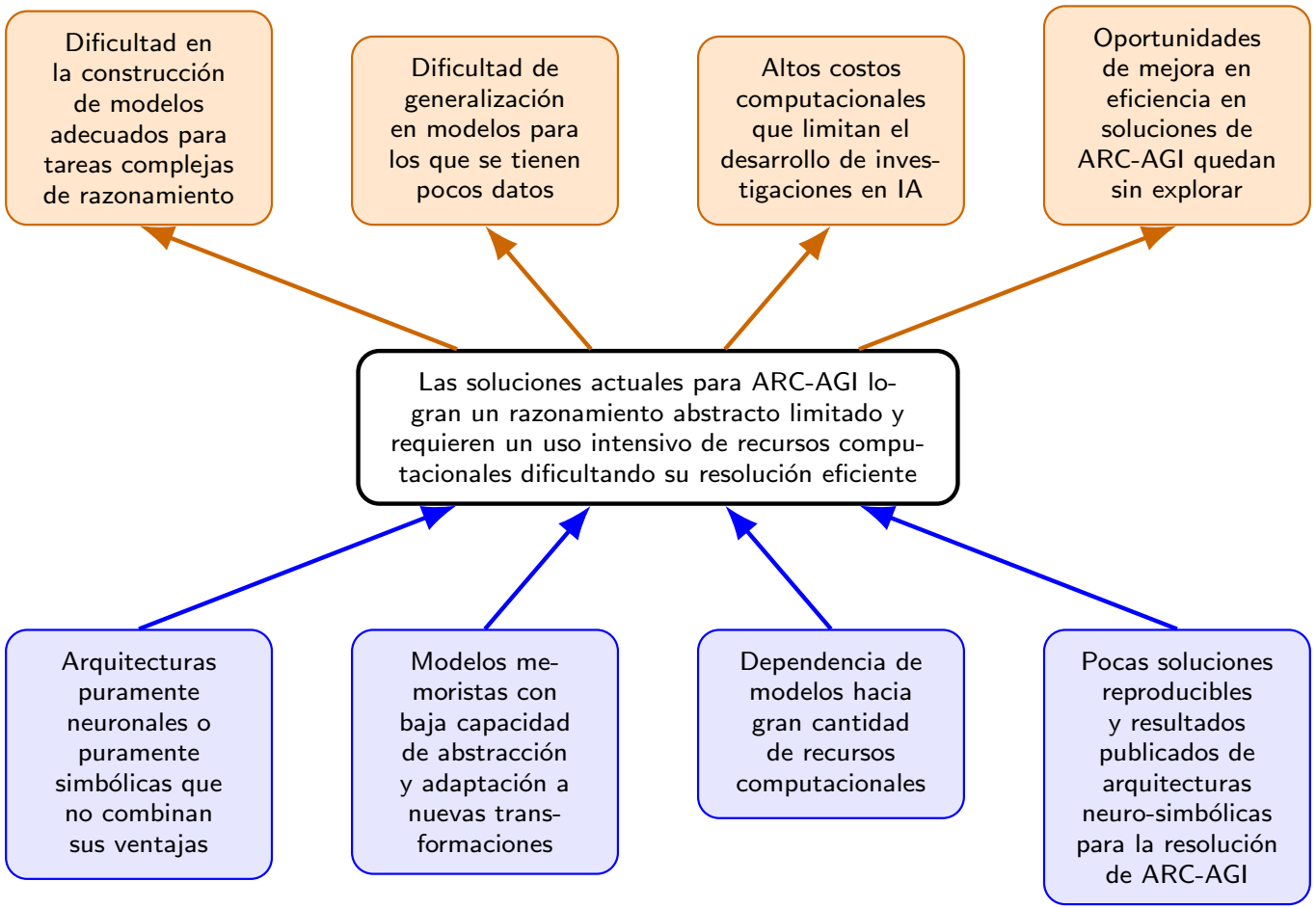

Figura 3: Árbol de Problemas

Problema Seleccionado

La problemática principal es la **dificultad actual que presentan los modelos de IA** para realizar **razonamiento abstracto** y **generalizar** efectivamente a partir de pocos ejemplos.

En benchmarks como ARC-AGI se evidencia cómo las arquitecturas actuales, principalmente LLMs, tienen **limitaciones significativas en términos de eficiencia**, además de sufrir dependencia de grandes volúmenes de datos para su entrenamiento y operación.

Objetivos

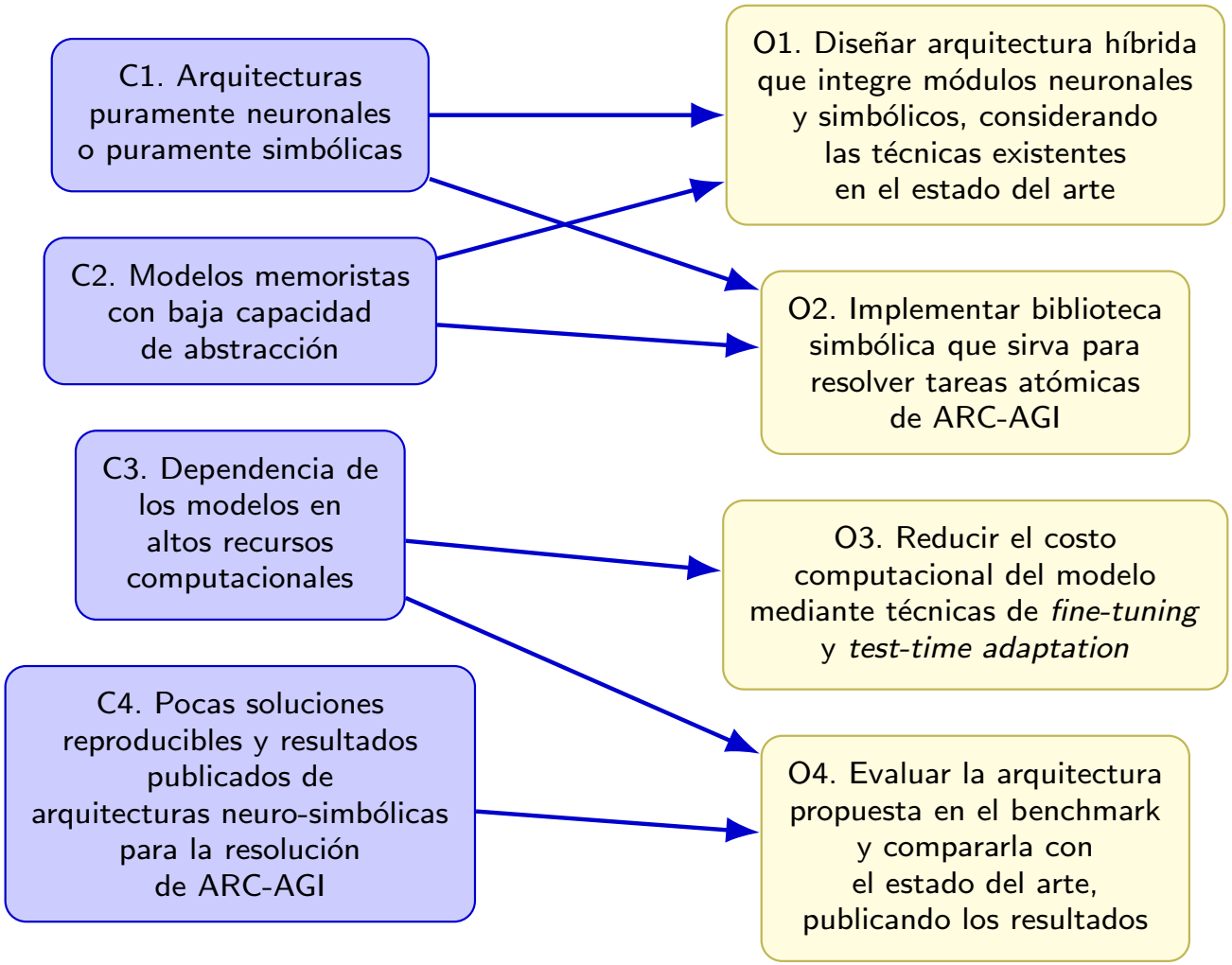

Figura 4: Relación entre Problemas y Objetivos

Resultados Esperados para el O1

O1. Diseñar una arquitectura híbrida que integre redes neuronales ligeras y módulos de razonamiento simbólico para abordar el benchmark ARC-AGI considerando las técnicas existentes en el estado del arte.		
Resultado Esperado	**Medio de Verificación**	**IOV**
R1. Selección de técnicas y algoritmos usados en el estado del arte con enfoque neuro-simbólico.	- Informe de selección de técnicas con análisis comparativo.	- Informe en el que se presenten al menos **4 técnicas** usadas en arquitecturas neuro-simbólicas con justificación. **Este debe ser aprobado por un experto en IA**.
R2. Selección de datasets incluyendo el original de ARC-AGI.	- Repositorio en Google Drive.	- Documento donde se presenten al menos **3 datasets adicionales** a ARC-AGI con justificación y análisis.
R3. Pipeline base donde se integren los módulos simbólicos y los LLM seleccionados.	- Código fuente en Python del pipeline base implementado, junto a un informe técnico.	- Repositorio en línea con código fuente accesible. - Informe técnico que detalle la arquitectura propuesta y su justificación. **Este debe ser aprobado por un experto en IA**.

Cuadro 1: Resultados Esperados, Medios de Verificación e IOV para el Objetivo 1

Resultados Esperados para el O2

O2. Implementar una biblioteca de operaciones y transformaciones genéricas y reutilizables para la resolución de los distintos tipos de problemas propuestos en ARC-AGI, de forma que estas actúen como módulo de razonamiento simbólico en la arquitectura.		
Resultado Esperado	**Medio de Verificación**	**IOV**
R1. Biblioteca en Python con transformaciones elementales como rotación, reflexión, coloreo, y otras operaciones que permitan la resolución de problemas de ARC-AGI.	- Código fuente en Python de la biblioteca.	- Repositorio en línea con código fuente accesible donde se implementen al menos **10 operaciones atómicas**.
R2. Validación de la biblioteca mediante su evaluación en subtareas preseleccionadas de ARC-AGI.	- Notebook para evaluación automatizada sobre un subconjunto de subtareas de los datasets elegidos.	- **Precisión superior al 80 %** por parte de las funciones de la biblioteca para la resolución de tareas atómicas.
R3. Documentación que detalle el uso de la biblioteca, las funciones que incluye y ejemplos de uso.	- Notebooks que demuestren el uso de la biblioteca con ejemplos.	- Notebooks publicados en el repositorio del proyecto que demuestren el uso de **todas las funciones implementadas**.

Cuadro 2: Resultados Esperados, Medios de Verificación e IOV para el Objetivo 2

Resultados Esperados para el O3

O3. Reducir el costo computacional de la arquitectura mediante técnicas de fine-tuning y test-time adaptation de manera que la solución sea eficiente en términos de recursos y tiempo de ejecución.		
Resultado Esperado	**Medio de Verificación**	**IOV**
R1. Aplicación de *fine-tuning* para ajuste de capas y parámetros, optimizando así la abstracción.	- Registro de hiperparámetros y pesos finales del modelo ajustado.	- Informe que incluya **al menos 2 métricas comparativas** entre el modelo base y el modelo ajustado. **Debe ser aprobado por un experto en IA**.
R2. Evaluación cuantitativa de recursos usados (consumo de GPU/CPU, memoria y tiempos de inferencia).	- Reporte con análisis comparativo entre los recursos usados por el modelo base y el ajustado. - Informe que detalle la relación costo/beneficio de la arquitectura propuesta en las fases de entrenamiento y evaluación.	- Documento donde se presenten **al menos 3 métricas comparativas** entre el modelo base y el ajustado. - Informe que detalle el proceso de evaluación y los resultados obtenidos. **Debe ser aprobado por un experto en IA**.

Cuadro 3: Resultados Esperados, Medios de Verificación e IOV para el Objetivo 3

Resultados Esperados para el O4

O4. Evaluar la arquitectura propuesta en el benchmark ARC-AGI y compararla con el estado del arte, publicando los resultados en un repositorio público para su uso por la comunidad científica.		
Resultado Esperado	**Medio de Verificación**	**IOV**
R1. Evaluación de la arquitectura en los datasets público y semi-privado del benchmark.	Código en Python para la evaluación del modelo final.	- Repositorio en línea con código accesible donde se **evidencie la evaluación del modelo**. - Reporte de resultados en el repositorio con las métricas obtenidas y su justificación. **Este debe ser aprobado por un experto en IA**.
R2. Comparación de resultados con los de arquitecturas del estado del arte.	- Reporte con análisis comparativo entre los resultados obtenidos y los del estado del arte.	- Documento con análisis estadístico entre la solución propuesta y el estado del arte, incluyendo tablas y gráficos comparativos. **Este debe ser aprobado por un experto en IA**.
R3. Discusión final sobre limitaciones encontradas y oportunidades de mejora.	- Informe de conclusiones donde se discuta la efectividades de las técnicas utilizadas.	- Informe en el repositorio con **al menos 3 limitaciones** identificadas y **2 oportunidades de mejora**.
R4. Publicación del proyecto en un repositorio abierto.	- Repositorio público en GitHub con licencia MIT, código fuente, reportes e informes.	- Repositorio público en GitHub subido tras la finalización del proyecto. - README que resuma el repositorio.

Cuadro 4: Resultados Esperados, Medios de Verificación e IOV para el Objetivo 4

Métodos, Herramientas y Técnicas

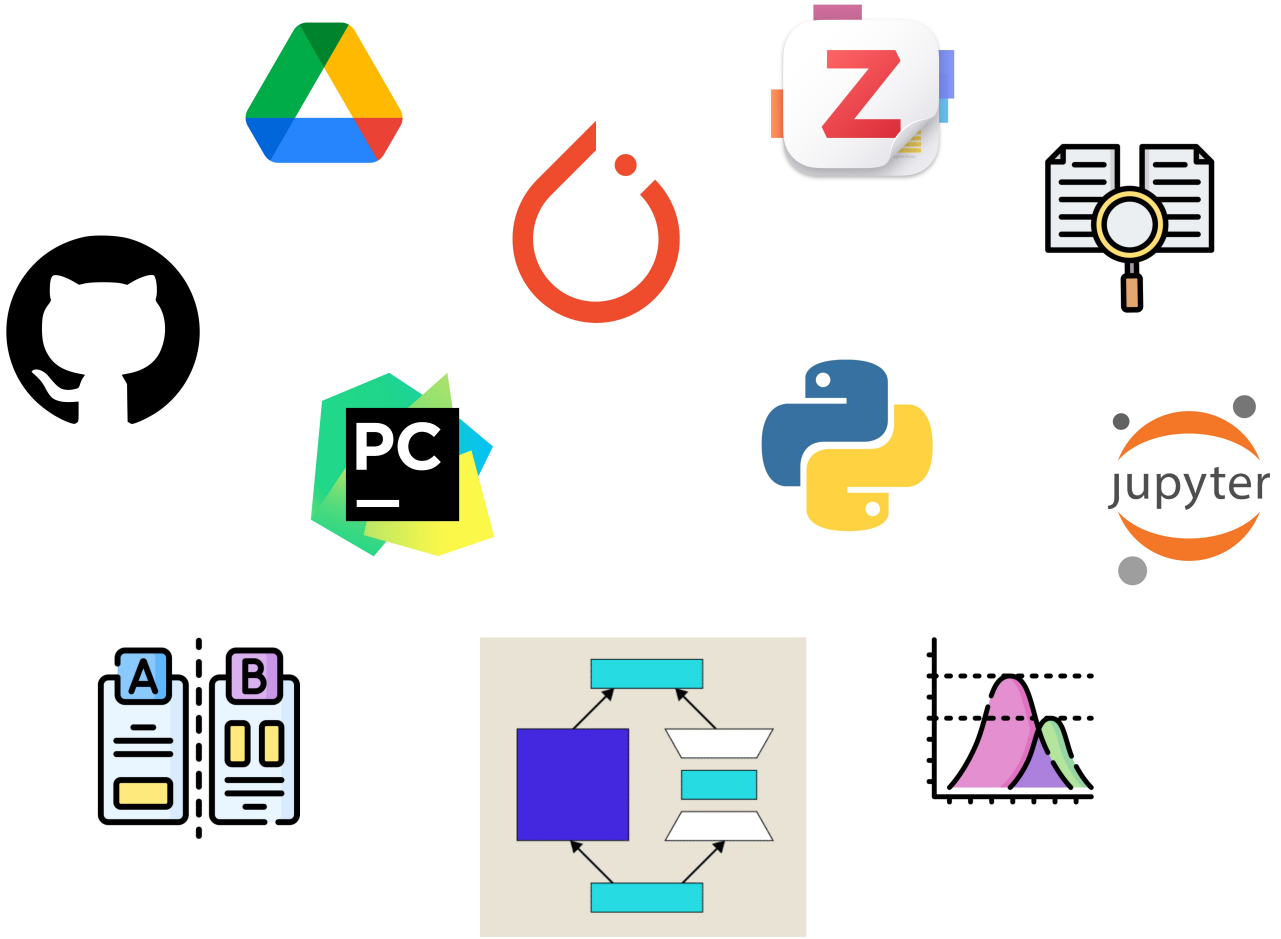

Figura 5: Métodos, Herramientas y Técnicas

2. Marco Referencial

Marco Teórico

- *Artificial Fluid Intelligence*
- Razonamiento Abstracto
- *Symbolic AI*
- Redes Neuronales
- *Neuro-Symbolic Integration Theory*

Marco Conceptual

- *ARC-AGI*
- Generalización
- Eficiencia
- *Lightweight Large Language Models*
- *Few-shot Learning*

3. Estado del Arte

Estado del Arte

Se realizó una revisión sistemática de la literatura (SLR).

El objetivo principal fue proporcionar una evaluación exhaustiva y crítica de la literatura existente relacionada con **arquitecturas neuro-simbólicas que busquen resolver problemas que requieran razonamiento abstracto** y que sean similares a los del benchmark ARC-AGI.

Preguntas de Investigación

- **P1**: ¿Cuáles son las arquitecturas neuro-simbólicas existentes que han sido propuestas para resolver problemas de razonamiento abstracto y que funcionen con LLMs?
- **P2**: ¿Qué benchmarks y conjuntos de datos se han utilizado para evaluar estas arquitecturas y cómo se comparan con el original propuesto en ARC-AGI?
- **P3**: ¿Cuál ha sido la eficiencia computacional de cada enfoque en términos de tiempo de entrenamiento, consumo de recursos y complejidad de los modelos?
- **P4**: ¿Cuáles son las métricas de rendimiento y resultados obtenidos por cada enfoque en términos de *accuracy*, generalización y robustez?
- **P5**: ¿Qué brechas, limitaciones y direcciones futuras de investigación se han identificado en la literatura existente?

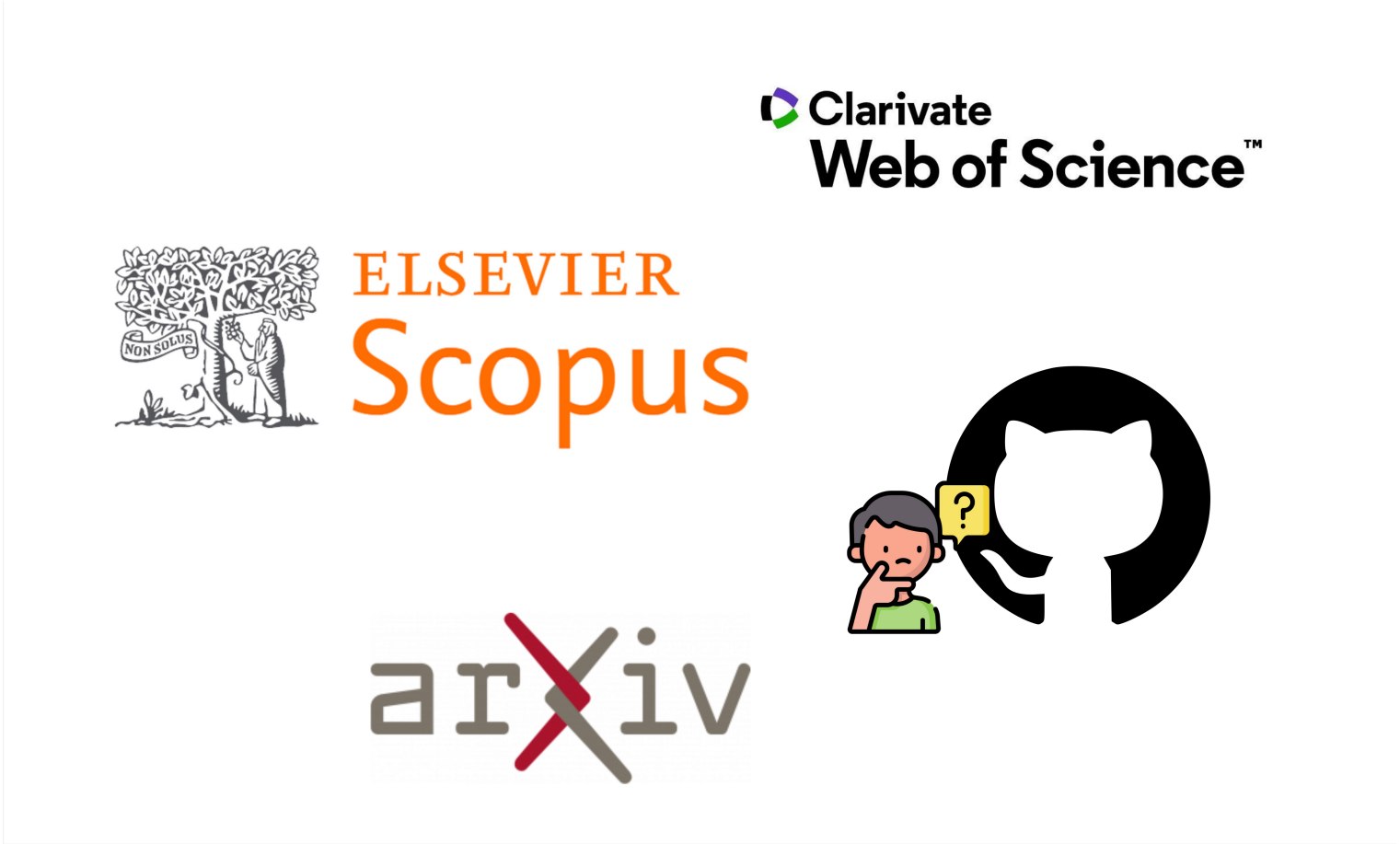

Figura 6: Motores de búsqueda

Cadenas de Búsqueda

Motor de búsqueda	Cadena de búsqueda	Número de Resultados
Scopus	(‘‘artificial intelligence’’ OR ai OR ‘‘machine learning’’ OR ‘‘deep learning’’) AND (‘‘neuro-symbolic AI’’ OR ‘‘neuro-symbolic reasoning’’ OR ‘‘neuro-symbolic learning’’ OR ‘‘compositional generalization’’ OR ‘‘compositional reasoning’’) AND (‘‘ARC-AGI’’ OR arc OR ‘‘abstract reasoning’’ OR ‘‘abstract reasoning benchmark’’) AND PUBYEAR > 2017 AND PUBYEAR < 2026 AND (LIMIT-TO (DOCTYPE , ‘‘ar’’) OR LIMIT-TO (DOCTYPE , ‘‘cp’’) OR LIMIT-TO (DOCTYPE , ‘‘bk’’)) AND (LIMIT-TO (LANGUAGE , ‘‘English’’))	108
Web of Science	((TS=((artificial intelligence OR AI OR ‘‘machine learning’’ OR ‘‘deep learning’’) AND (‘‘neuro-symbolic AI’’ OR ‘‘neuro-symbolic reasoning’’ OR ‘‘neuro-symbolic learning’’ OR ‘‘compositional generalization’’ OR ‘‘compositional reasoning’’))) AND LA=(English OR Spanish)) AND DOP=(2018/2024)	51
arXiv	(order: -announced_date_first; size: 100; date_range: from 2018-01-01 to 2025-12-31; classification: Computer Science (cs); include_cross_list: True; terms: AND all=((‘‘neuro-symbolic’’ OR compositional OR ‘‘deep learning’’ OR ‘‘artificial intelligence’’ OR AI) AND (‘‘ARC-AGI’’ OR ‘‘abstraction and reasoning corpus’’)))	39

Cuadro 5: Cadenas y resultados en diferentes motores de búsqueda.

Criterios de Inclusión y Exclusión

Los criterios de inclusión son:

- **Relevancia temática**: Los estudios deben abordar el tema de IA neuro-simbólica y el razonamiento abstracto.
- **Lenguaje**: Los estudios deben estar escritos en inglés o español.
- **Tipo de publicación**: Se aceptan artículos, trabajos de conferencias, reportes técnicos y preprints de alta calidad.
- **Fecha de publicación**: Estudios publicados entre 2018 y 2025.
- **Calidad metodológica**: Los estudios deben presentar una metodología clara y resultados que respalden sus conclusiones.

Los criterios de exclusión son:

- **Contenido irrelevante**: Los estudios que no aborden el tema de IA neuro-simbólica o que no estén relacionados con el benchmark ARC-AGI, ni problemas de razonamiento abstracto.
- **Poca documentación**: Los estudios que no presenten datos suficientes o que no se puedan reproducir.
- **Accesibilidad limitada**: Los estudios que no sean accesibles en su totalidad o que solo presenten resúmenes o *abstracts*.

Documentos seleccionados

Motor de búsqueda	**Cantidad de documentos**	**Porcentaje**
Scopus	8	33.33 %
Web of Science	2	8.33 %
arXiv	11	45.83 %
Otras fuentes	3	12.50 %
Total	**24**	**100 %**

Cuadro 6: Cantidad de documentos seleccionados por motor de búsqueda.

Formulario de extracción de datos

Campo	Descripción	Pregunta
ID	Identificador primario del estudio	General
Título	Título del estudio	General
Autores	Autores que formaron parte del estudio	General
Año	Año de publicación	General
Fuente	Nombre de la revista o conferencia	General
Enlace de consulta	URL al artículo completo	General
Abstract	Resumen del estudio	General
Citaciones	Número de citas recibidas	General
Arquitectura	Arquitectura neuro-simbólica propuesta	P1
Enfoque	Enfoque de razonamiento utilizado	P1
Dataset y benchmark	Dataset y benchmark utilizado para la evaluación	P2
Eficiencia computacional	Tiempo de entrenamiento, consumo de recursos y complejidad del modelo	P3
Resultados y métricas	Resultados obtenidos y métricas de rendimiento	P4
Limitaciones y brechas	Limitaciones y brechas identificadas en el estudio	P5
Direcciones futuras	Direcciones futuras de investigación propuestas	P5

Cuadro 7: Formulario de extracción de datos para la revisión sistemática.

Respuesta a la pregunta P1

P1. ¿Cuáles son las arquitecturas neuro-simbólicas existentes que han sido propuestas para resolver problemas de razonamiento abstracto y que funcionen con LLMs?

- *Program Induction*: Arquitecturas neuro-simbólicas que combinan modelos de lenguaje visual con módulos simbólicos para inducir programas desde ejemplos.
- Un enfoque reciente propone un *ensemble* entre un modulo de transducción y uno de inferencia para resolver ARC-AGI [10].
- Hay propuestas que se enfocan en *test time fine-tuning* y *data augmentation* durante el tiempo de inferencia [8, 6].
- Muchas arquitecturas usan técnicas de reducción dimensional para mejorar la eficiencia y el rendimiento [8].

Respuesta a la pregunta P2

P2. ¿Qué benchmarks y conjuntos de datos se han utilizado para evaluar estas arquitecturas y cómo se comparan con el original propuesto en ARC-AGI?

- La mayoría de estudios usan como benchmark principal el ARC-AGI original y variantes como **ConceptARC y LARC** [2, 1, 7, 6].
- También se utilizan datasets basados en *Raven's Progressive Matrices* (como RAVEN e I-RAVEN) [13, 9].
- Otros conjuntos, como GSM8K, SVAMP, AQuA y DARG, que evalúan habilidades en matemáticas, lenguaje natural o razonamiento en grafos [12, 11].

Respuesta a la pregunta P3

P3. ¿Cuál ha sido la eficiencia computacional de cada enfoque en términos de tiempo de entrenamiento, consumo de recursos y complejidad de los modelos?

- The ARCHitects utilizan **2 GPUs Nvidia T4 (16GB)** con LLMs como **Mistral-NeMo-Minitron-8B** y **Llama3.2-3B**, adaptadores LoRA (**64–256**) y cuantización a **4 bits** [8].
- El modelo de Li et al. (2024) mejora el *performance* en un **20–30 %** durante inferencia (20,000 muestras) al aumentar el presupuesto [10].
- La solución de Cole y Osman (2025) implementa **LongT5 en una GPU P100 (16GB)** y logra la mejor eficiencia y métricas en el dataset privado de ARC-AGI [6].
- El modelo de Zhang et al. (2022) usa **4 GPUs A100 (40GB)** y **LoRA** para reducir el uso de memoria y mejorar la eficiencia [13].

Respuesta a la pregunta P4

P4. ¿Cuáles son las métricas de rendimiento y resultados obtenidos por cada enfoque en términos de accuracy, generalización y robustez?

- Zhang et al. (2022) logra **78.45 %** en sistematicidad, **79.95 %** en productividad y **80.5 %** en localismo usando RPMs [13].
- Hersche et al. (2023) alcanza **87.7 %** y **88.1 %** de precisión en RAVEN e I-RAVEN, los mejores resultados en estos benchmarks [9].
- En ARC-AGI, Cole y Osman (2025) logran el mejor resultado con **58.5 %** de precisión usando *test time fine-tuning* [6].
- *The ARCHitects* alcanza **56.5 %** [8].
- Otros modelos oscilan entre **30–40 %** [3].

Respuesta a la pregunta P5

P5. ¿Qué brechas, limitaciones y direcciones futuras de investigación se han identificado en la literatura existente?

- Las arquitecturas híbridas muestran buena precisión, pero tienen problemas de escalabilidad y alta dependencia de ejemplos anotados.
- El enfoque de *test-time fine-tuning* implica altos costos computacionales.
- Se sugiere investigar métodos de inducción simbólica automatizada y pipelines con múltiples enfoques.
- Se propone ampliar el benchmark ARC-AGI para evaluar más aspectos de la inteligencia y fomentar generalización sistemática [5].

Conclusiones

- Las arquitecturas híbridas neuro-simbólicas muestran alto potencial en tareas de razonamiento abstracto y generalización sistemática.
- El rendimiento mejora al combinar modelos ligeros, técnicas de inducción algebraica y estrategias como *test-time fine-tuning*.
- Es recomendable usar ARC-AGI, junto con sus variantes dados ue incluyen problemas más diversos y complementarios.
- Persisten limitaciones en escalabilidad, adaptabilidad y eficiencia computacional.

4. Cronograma

Cronograma

Semana	Sesión de Clase	Entregables	Actividades	Envío del avance preliminar al asesor	Envío al asesor/publicación de la siguiente semana (antes del mediodía)	Revisor
1	**Exposición de tema, cronograma y estado de avance**	**Cronograma de trabajo del curso**		**Miércoles**	**Viernes (antes del mediodía) / Lunes de la siguiente semana (antes del mediodía)**	Asesor
2	Exposición de tema, cronograma y estado de avance.	E1: avance del 40%	- Revisión de literatura - Elaboración del formulario de extracción - Definición de preguntas de investigación	Miércoles		-
3	Exposición de tema, cronograma y estado de avance.	E1: avance del 90%	- Elaboración del *problem statement* - Análisis de las relaciones causa-efecto en el estudio a realizar - Definición de resultados esperados	Miércoles		-
4	**Exposición 1**	**E1: Problemática, estado del arte, objetivos (general y específicos), resultados esperados y cronograma de trabajo en el curso**	**Revisión final del documento y ajuste de correcciones según retroalimentación**	**Miércoles**	**Viernes (antes del mediodía) / Lunes de la siguiente semana (antes del mediodía)**	Profesor del curso
5	Exposición 2	E2 (avance 20%): Desarollo del marco conceptual y teórico	- Identificación de conceptos clave - Búsqueda de fuentes para la elaboración de definiciones	Miércoles		-
6	Exposición 3	E2 (avance 60%): Definición de herramientas, métodos y procedimientos	- Avance de la metodología a utilizar - Definición y justificación de las herramientas a utilizar	Miércoles		-
7	Exposición 4	E2 (avance 90%): Finalización de revisión del marco teórico y metodológico	- Conclusión de la metodología a utilizar - Corrección de observaciones	Miércoles		-
8	**Exposición 5**	**E2: Levantamiento de observaciones del E1, marco conceptual/teórico/legal, herramientas, métodos y procedimientos**	**Revisión final del documento y ajuste de correcciones según retroalimentación**	**Miércoles**	**Viernes (antes del mediodía) / Lunes de la siguiente semana (antes del mediodía)**	Profesor del curso
9	EXÁMENES PARCIALES					-
10	Exposición 6	E3 (avance 20%): Justificación y viabilidad técnica y económica	- Evaluación de factibilidad del proyecto - Análisis de costos y de recursos	Miércoles		-
11	Exposición 7	E3 (avance 60%): Definición del alcance y EDT	- Desarrollo del EDT - Validación del alcance del proyecto - Identificación detallada de las tareas	Miércoles		-
12	Exposición 8	E3 (avance 90%): Elaboración del cronograma detallado y análisis de riesgos	- Planificación detallada de las tareas - Asignación de tiempos y recursos - Identificación de dependencias - Elaboración del cronograma	Miércoles		-
13	**Exposición 9**	**E3: Proyecto de fin de carrera completo incluyendo: todas las correcciones y el Anexo de Plan de Proyectos**	**Revisión final del documento y ajuste de correcciones según retroalimentación**	**Miércoles**	**Viernes (antes del mediodía) / Lunes de la siguiente semana (antes del mediodía)**	Jurado
14	Exposición 10					-
15 y 16	Exposiciones finales					Jurado
17	EXÁMENES FINALES					-

Figura 7: Cronograma

Gracias

¿Preguntas?

Referencias I

[1]. Mattia Atzeni, Mrinmaya Sachan, and Andreas Loukas.
Infusing lattice symmetry priors in attention mechanisms for sample-efficient abstract geometric reasoning.
In *International Conference on Machine Learning*, pages 1200–1217. PMLR, 2023.

[2]. Giacomo Camposampiero, Loïc Houmard, Benjamin Estermann, Joël Mathys, and Roger Wattenhofer.
Abstract visual reasoning enabled by language.
In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2643–2647, 2023.

[3]. Francois Chollet, Mike Knoop, Gregory Kamradt, and Bryan Landers.
Arc prize 2024: Technical report, 2025.

[4]. François Chollet.
On the measure of intelligence.
arXiv preprint arXiv:1911.01547, 2019.

[5]. François Chollet.
Announcing arc agi 2 and arc prize 2025, 2025.

Referencias II

[6]. Jack Cole and Mohamed Osman.
Don't throw the baby out with the bathwater: How and why deep learning for arc.
`https://github.com/MohamedOsman1998/deep-learning-for-arc/blob/main/deep_learning_for_arc.pdf`, 2025.
Manuscript hosted on GitHub.

[7]. Sébastien Ferré.
Tackling the abstraction and reasoning corpus (arc) with object-centric models and the mdl principle.
In *International Symposium on Intelligent Data Analysis*, pages 3–15. Springer, 2024.

[8]. Daniel Franzen, Jan Disselhoff, and David Hartmann.
The llm architect: Solving arc-agi is a matter of perspective.
2024.

Referencias III

[9]. Michael Hersche, Mustafa Zeqiri, Luca Benini, Abu Sebastian, and Abbas Rahimi.
A neuro-vector-symbolic architecture for solving raven's progressive matrices.
Nature Machine Intelligence, 5(4):363–375, 2023.

[10]. Wen-Ding Li, Keya Hu, Carter Larsen, Yuqing Wu, Simon Alford, Caleb Woo, Spencer M Dunn, Hao Tang, Michelangelo Naim, Dat Nguyen, et al.
Combining induction and transduction for abstract reasoning.
arXiv preprint arXiv:2411.02272, 2024.

[11]. Linlu Qiu, Liwei Jiang, Ximing Lu, Melanie Sclar, Valentina Pyatkin, Chandra Bhagavatula, Bailin Wang, Yoon Kim, Yejin Choi, Nouha Dziri, et al.
Phenomenal yet puzzling: Testing inductive reasoning capabilities of language models with hypothesis refinement.
arXiv preprint arXiv:2310.08559, 2023.

Referencias IV

[12]. Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou.
Self-consistency improves chain of thought reasoning in language models.
arXiv preprint arXiv:2203.11171, 2022.

[13]. Chi Zhang, Sirui Xie, Baoxiong Jia, Ying Nian Wu, Song-Chun Zhu, and Yixin Zhu.
Learning algebraic representation for systematic generalization in abstract reasoning.
In *European Conference on Computer Vision*, pages 692–709. Springer, 2022.