Anteproyecto

UCASAL

Alumno: Alejo Torres

# Título del trabajo

El título del trabajo a desarrollar será: “Arquitectura Transformer: Análisis del estado del arte de modelos LLM y su integración en el desarrollo de aplicaciones impulsadas por IA”.

# Profesor guía propuesto

La Mg. Lic. María Lorena Talamé de la Universidad Católica de Salta ha sido seleccionada como directora propuesta para liderar el seguimiento y acompañamiento del proyecto, Esto debido a la relación con el docente y su amplio conocimiento en materia de investigación e inteligencia artificial.

# Breve descripción del trabajo

El presente trabajo tiene por objetivo explorar los fundamentos de la arquitectura Transformer, explicando cuales fueron los hitos previos en el mundo del Machine Learning que dieron lugar a su surgimiento y realizar un análisis del estado del arte actual de los LLMs más populares, tanto de naturaleza Open Source como freemium. Se busca conocer con certeza el alcance actual de una tecnología que hoy en día se encuentra en boca de todos. A su vez, explorar el ecosistema de framewokrs, herramientas y abstracciones creadas por la comunidad que permiten aprovechar los modelos de lenguaje en el desarrollo de aplicaciones impulsadas por Inteligencia Artificial, como lo son motores de búsqueda, asistentes “ChatBot”, apps para resumir información o extracción de ideas principales etc.

El resultado final de este trabajo es producir un documento que condense la información actual y explique en profundidad esta arquitectura que da sustento a modelos como GPT3, Llama2,BERT,Falcon40B,entro otros. Acompañado del desarrollo de pequeñas pruebas de concepto y aplicaciones web como las mencionadas anteriormente, que permitan demostrar el alcance actual de esta tecnología y como incorporarla en el desarrollo de aplicaciones.

## Objetivo General

El objetivo principal será producir un documento que condense la información disponible y explique en profundidad esta arquitectura, acompañado de pruebas de concepto y aplicaciones web, que permitan demostrar el alcance actual de esta tecnología y como incorporarlas en el flujo de desarrollo de software.

Formando un potente recurso que apunta a lograr un entendimiento completo del tema. Y que, en un futuro, pueda servir como material bibliográfico para cualquiera que esté interesado en estudiar esta rama de los modelos generativos.

## Objetivos Específicos

Los objetivos propuestos en el marco del presente trabajo serán:

1. **Investigación sobre la arquitectura Transformer:** El trabajo requiere investigar principalmente sobre los papers originales, complementándolos con artículos científicos, masterclasses, seminarios y charlas informativas, con el objetivo de recopilar toda la información necesaria para obtener un dominio completo sobre el tema.
2. **Desarrollo de Notebooks:** Utilizare Jupyter Notebooks para explorar los diferentes componentes de la arquitectura, realizar pruebas de concepto y demostraciones prácticas escritas en Python, junto a explicaciones detalladas.
3. **Investigar sobre los LLMs más populares:** El objetivo es realizar una investigación de los modelos de lenguaje más populares usados por la comunidad actualmente, tanto Open Source como freemium, con el objetivo de realizar un análisis comparativos, teniendo en cuenta sus dimensiones y rendimiento bajo una serie de pruebas orientadas a la practicidad y utilidad del modelo.
4. **Explorar el ecosistema de las aplicaciones impulsadas por IA:** Contando con un entendimiento bien formado sobre los LLM me dispondré a explorar el ecosistema de herramientas y librerías como LangChain, CTransformer, BitsAndBytes, StreamLit, ChainLit, entre otras. Para comprender cual es el paradigma a seguir a la hora de pensar y desarrollar servicios, aplicaciones y abstracciones sobre modelos del lenguaje de estas características.
5. **Desarrollo de aplicaciones impulsadas por IA:** El objetivo será, utilizando los conocimientos sobre el ecosistema de desarrollo, construir aplicaciones traten de resolver problemáticas reales o que puedas extenderse con tal fin. Esto incluirá asistentes “ChatBots”, motores de búsqueda, aplicaciones orientadas al tratamiento de información: resúmenes, extracción de ideas principales, creación de herramientas de estudio, etc.

# Alcances del trabajo

El alcance del trabajo incluirá los siguientes aspectos:

1. **Exploración de la Arquitectura Transformer y su contexto histórico:** El trabajo se enfocará en investigar los fundamentos de la arquitectura Transformer, incluyendo hitos anteriores en el campo del Machine Learning que llevaron a su desarrollo. Esto requerirá revisar los papers originales, así como artículos científicos, seminarios y charlas informativas para obtener un conocimiento completo y profundo de esta arquitectura.
2. **Análisis del estado del arte de LLMs:** Se examinará el panorama actual de los Modelos de Lenguaje de Aprendizaje Profundo (LLMs) tanto Open Source como freemium. Esto incluye una investigación exhaustiva sobre modelos más populares como GPT-3, Llama2, BERT, Falcon40B y otros. Se realizarán análisis comparativos, considerando aspectos como dimensiones y rendimiento mediante pruebas prácticas.
3. **Exploración del ecosistema de desarrollo de IA:** Se Explorará el ecosistema de herramientas, frameworks y abstracciones creadas por la comunidad para aprovechar los modelos de lenguaje en aplicaciones de Inteligencia Artificial. Se estudiarán librerías como LangChain, CTransformer, BitsAndBytes, StreamLit y ChainLit, con el objetivo de documentar y comprender los paradigmas y enfoques para desarrollar aplicaciones impulsadas por modelos de lenguaje.
4. **Desarrollo de aplicaciones prácticas de IA:** El trabajo incluirá la creación de pequeñas pruebas de concepto y aplicaciones web que demuestren el alcance actual de la tecnología de LLMs y cómo se pueden incorporar en aplicaciones del mundo real. Esto comprenderá la construcción de asistentes "ChatBot", motores de búsqueda y aplicaciones orientadas al procesamiento de información, como resúmenes y extracción de ideas principales.

# Institución, empresa u organización

Todos Los avances, documentos, aplicaciones y pruebas de concepto obtenidos serán puestos a disposición del Departamento de Ingeniería en Informática de la Universidad Católica de Salta. Con el objetivo de servir como una guía de estudio del tema.

# Metodología y/o procedimiento a utilizar

Para abordar este proyecto de manera efectiva, se puede adoptar una metodología en la que se combinen enfoques iterativos y fases específicas para la etapa de investigación. La etapa construcción de software, utilizara una adaptación del enfoque en cascada. Esto permitirá una adaptación constante a los nuevos hallazgos de la investigación en un ámbito tan dinámico como los es el de las IA generativas text-to-text y un proceso controlado en desarrollo de las aplicaciones.

**Fase de Investigación:**

1. **Definición de Objetivos de Investigación:**

Establecer claramente los objetivos específicos de la investigación, como la comprensión profunda de la arquitectura Transformer, la identificación de hitos históricos relevantes y el análisis del estado actual de los LLMs.

2. **Recopilación de Información:**

Investigar los papers originales sobre la arquitectura Transformer y sus precursores en el campo del Machine Learning. Complementando con artículos científicos, tutoriales, seminarios y charlas informativas para condensar la información y obtener una comprensión completa del tema.

3. **Análisis Comparativo de LLMs:**

Investigar y comparar los modelos de lenguaje más populares (GPT-3, Llama2, BERT, Falcon40B, entre otros). Evaluar sus características, ventajas y limitaciones. Realizar pruebas prácticas y mediciones de rendimiento bajo diversos escenarios.

4. **Exploración del Ecosistema de Desarrollo:**

Investigar herramientas y librerías utilizadas para desarrollar aplicaciones impulsadas por IA, como LangChain, CTransformer, BitsAndBytes, StreamLit, etc. Comprender sus características y cómo se integran con los modelos de lenguaje.

**Fase de Construcción de Software:**

Para cada una de las aplicaciones y pruebas de conceptos que complementaran la investigación se opta por basarse en el enfoque de "desarrollo en cascada". Este enfoque implica la realización de actividades secuenciales y claramente agrupadas en fases o ciclos. La decisión de utilizar este enfoque se debe a la naturaleza de las aplicaciones a construir y al hecho de que el equipo de desarrollo contara con una sola persona, lo cual no requiere priorizar las comunicaciones internas dentro del equipo. Con este enfoque metodológico, se espera lograr un desarrollo estructurado y efectivo de las aplicaciones impulsadas por IA.

1. **Requisitos de la Aplicación:** En esta fase inicial, se definirán los requisitos específicos de cada aplicación. Su objetivo principal, funcionalidades clave y cuál es el caso de uso que pretende demostrar.
2. **Diseño:** Se creará un diseño detallado de la arquitectura de cada aplicación. Esto incluirá la estructura general, la interacción de los componentes y cómo se integrarán los modelos de lenguaje. Además, se diseñará la interfaz de usuario con consideraciones de usabilidad y experiencia del usuario.
3. **Implementación:** En esta fase, se llevará a cabo la codificación de cada aplicación según los diseños previamente establecidos. Poniendo en práctica las herramientas del ecosistema sobre las cuales se investigó previamente.
4. **Pruebas:** Se realizarán pruebas para verificar el rendimiento de las aplicaciones enfocándose en la interacción con los LLMs y las directivas utilizadas para definir cada una de las tareas dentro del flujo.

# Plan de trabajo

A continuación se proporciona un plan de trabajo tentativo para el proyecto

Fase de Investigación:

1. Definición de Objetivos de Investigación (Duración: 1 semana):

- Establecer los objetivos específicos de la investigación, detallando lo que se espera lograr en cada uno.

2. Recopilación de Información (Duración: 3 semanas):

- Investigar papers originales de la arquitectura Transformer y su contexto histórico en el campo del Machine Learning.

- Complementar la investigación con artículos científicos, tutoriales, seminarios y charlas informativas para obtener una comprensión profunda.

3. Análisis Comparativo de LLMs (Duración: 2 semanas):

- Definir, Investigar y comparar los modelos de lenguaje más populares

- Evaluar sus características, ventajas y limitaciones.

- Realizar pruebas prácticas y mediciones de rendimiento bajo diferentes escenarios.

4. Exploración del Ecosistema de Desarrollo (Duración: 2 semanas):

- Investigar herramientas y librerías utilizadas para desarrollar aplicaciones impulsadas por IA, como LangChain, CTransformer, BitsAndBytes, StreamLit, etc.

- Comprender cómo se integran con los modelos de lenguaje y cuál es el paradigma a seguir en el desarrollo de aplicaciones impulsadas por IA.

**Fase de Construcción de Software:**

1. Requisitos de la Aplicación (Duración: 1 semana):

- Definir claramente los requisitos específicos de cada aplicación, incluyendo su objetivo principal y funcionalidades clave.

2. Diseño (Duración: 2 semanas):

- Crear un diseño detallado de la arquitectura de cada aplicación, incluyendo la interacción de los componentes y la integración de los modelos de lenguaje.

- Diseñar la interfaz de usuario para garantizar la usabilidad y la experiencia del usuario.

3. Implementación (Duración: 4 semanas):

- Codificar cada aplicación según los diseños previamente establecidos.

- Utilizar las herramientas, librerías y APIs del ecosistema de desarrollo de IA investigadas en la fase anterior.

4. Pruebas (Duración: 3 semanas):

- Realizar pruebas exhaustivas para verificar el rendimiento y la funcionalidad de las aplicaciones.

- Enfocarse en la interacción con los LLMs y en la precisión de las tareas definidas para cada aplicación.

**Evaluación y Presentación Final:**

1. Evaluación del Proyecto (Duración: 1 semana):

- Comparar los resultados obtenidos con los objetivos establecidos en la fase de investigación y desarrollo.

- Identificar posibles ajustes y mejoras necesarias en las aplicaciones.

2. Elaboración del Documento Final (Duración: 2 semanas):

- Crear un documento completo que condense la información sobre la arquitectura Transformer, el análisis de LLMs y la descripción detallada de las aplicaciones desarrolladas.

- Incluir resultados, conclusiones y recomendaciones.

Duración Total del Proyecto: Aproximadamente 21 semanas

Este plan de trabajo permitirá abordar de manera efectiva la investigación, el desarrollo de aplicaciones y la presentación final, abarcando el objetivo del proyecto en su totalidad. Cabe aclarar que las estimaciones de tiempo son tentativas y pueden sufrir modificaciones.

# Bibliografía

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877-1901

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., ... & Scialom, T. (2023). Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ba, J. L., Kiros, J. R., & Hinton, G. E. (2016). Layer normalization. *arXiv preprint arXiv:1607.06450*.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778).

Pytorch Documentation: <https://pytorch.org/docs/stable/index.html>

LangCahin Documentation: <https://docs.langchain.com/docs/>

StreamLit Documentation: <https://docs.streamlit.io/>

NextJs Documentation: https://nextjs.org/docs

Andrej Karpathy blogpost: <https://karpathy.ai/zero-to-hero.html>

<https://github.com/karpathy/nanoGPT>

<https://github.com/karpathy/makemore>

https://github.com/huggingface/transformers

Material de estudio utilizado en la Universidad Politécnica de Madrid en las asignaturas de Aprendizaje automático II y Sistemas inteligentes, en la carrera de Data Science e inteligencia Artificial dictadas por el Dr. Francisco García Serradilla

Camacho, J. D. V. (2020). Desarrollo en cascada (waterfall) VS desarrollo agile-SCRUM.