Universidad CENFOTEC

Maestría en Software con énfasis en Inteligencia Artificial

Evaluación de la generación de casos de pruebas basados en requerimientos con modelos *Large Language Model* (LLM) para acelerar procesos en la industria aeroespacial.

Elaborado por:

Kendall Méndez Hidalgo

Fecha: Noviembre, 2025

Tabla de Contenido

Tabla de Figuras

Índice de Tablas

Resumen

El resumen será elaborado en la versión final de este documento.

Palabras Clave: modelos *LLM*, casos de prueba basados en requerimientos, industria aeroespacial, DO-178C, verificación, software crítico, ética, estandarización, eficiencia.

Capítulo 1. Introducción

En este capítulo se introduce el tema a investigar, se explica cuál es el problema, el alcance, los objetivos y el estado de la cuestión. Las siguientes secciones detallan con mayor profundidad cada uno de estos temas.

1.1 Generalidades

El aumento de la complejidad de los sistemas aviónicos en la aviación civil requiere procesos rigurosos de aseguramiento y aprobación para garantizar el cumplimiento de las regulaciones federales. Para los sistemas basados en software, la Administración Federal de Aviación (FAA) reconoce la norma DO-178C como un medio aceptable de cumplimiento, tal como se indica en la Circular de Asesoramiento AC 20-115D.

La normal DO-178C establece las pruebas basadas en requerimientos como el método más efectivo para identificar errores. Este procedimiento consiste en analizar la descripción del requerimiento y a partir de esto, identificar casos de prueba que permitan verificar completamente el comportamiento y funcionalidad documentada en los requerimientos.

Dependiendo del requerimiento, es necesario incluir más o menos casos de prueba; por ejemplo, si el requerimiento describe funcionalidades relacionadas a máquinas de estado se deberán incluir en los casos de prueba todas las transiciones posibles de estado así como las transiciones inválidas con el fin de demostrar que el sistema no permitirá dichos cambios.

La norma DO-178C especifica las pruebas mínimas que se esperan de acuerdo al tipo de requerimiento.

Identificar casos de prueba es un procedimiento predominantemente manual en el cual los ingenieros analizan el tipo de requerimiento asignado y posteriormente generan los casos de prueba que se necesitan para verificar la correcta implementación del mismo.

Los casos de prueba (*test cases*) es el primer paso en el desarrollo de pruebas (*test procedure*) y, aunque no representa un gran porcentaje de la totalidad del esfuerzo, es de gran importancia ya que constituye la base de las pruebas como tal. Si la prueba no tiene los casos necesarios para verificar correctamente la implementación del requerimiento, se expone al riesgo de no encontrar errores en etapas tempranas del proyecto y además, se aumenta el tiempo de corrección de

pruebas en etapas de revisiones formales ya que no se cumple con las pruebas mínimas requeridas según los planes del proyecto.

1.2 Antecedentes del Problema

En la industria aeroespacial, las pruebas basadas en requerimientos son el método sugerido por la guía DO-178C. Dichas pruebas comienzan con la identificación de casos de prueba, que es realizada manualmente, y representa aproximadamente un 10% del esfuerzo total del desarrollo de pruebas. La optimización de esta etapa podría traducirse en un ahorro considerable de tiempo y costos para la empresa, permitiendo reasignar recursos a otras tareas.

También, la calidad de las pruebas muchas veces está directamente relacionada a la experiencia del ingeniero, por lo que se requiere de supervisión y planes de seguimiento para las personas que no tienen tanto tiempo trabajando en este tipo de actividades.

Los errores en la identificación de casos de prueba tienen consecuencias directas: incrementan el tiempo y los costos de revisión, pueden conducir a retrabajos costosos, retrasos en la obtención de certificaciones regulatorias y, en el peor de los casos, podrían derivar en fallos funcionales no detectados durante las pruebas, comprometiendo la seguridad y el cumplimiento normativo.

1.3 Definición y Descripción del Problema

El problema central de este trabajo radica en investigar y evaluar la efectividad de los *LLM* en generar casos de prueba que permitan agilizar los procesos de verificación sin comprometer la calidad. Dado que este es un nuevo enfoque con pocos estudios en el tema, una de las dificultades principales es la limitada disponibilidad de datos que respalden buenas prácticas. Por ello, es necesario definir también una metodología para implementar y evaluar este tipo de tecnología.

Los modelos *LLM* ofrecen una solución potencial, al interpretar requerimientos en lenguaje natural y generar casos de prueba automáticamente. Sin embargo, su implementación enfrenta desafíos específicos:

- **Ambigüedad en requerimientos técnicos:** Los *LLM* pueden malinterpretar términos técnicos, generando casos de prueba que no cumplen con los

estándares de trazabilidad.

- **Falta de estandarización:** No existen procedimientos consolidados para integrar *LLMs* en procesos de prueba aeroespaciales, lo que complica su adopción.

- **Preocupaciones éticas:** El uso de *LLMs* en sistemas críticos plantea riesgos, como la generación de casos de prueba incorrectos que podrían comprometer la seguridad aérea.

El enfoque alternativo se selecciona para evaluar métodos de *LLM* cuantitativa (precisión, sensibilidad) y cualitativamente (aceptación del usuario), aprovechando las fortalezas de los métodos mixtos para abordar necesidades técnicas y contextuales, y específicamente, para mitigar la ambigüedad en requerimientos técnicos.

La investigación aplicada es adecuada para proponer una guía que permita incorporar este tipo de tecnologías y diseñar un prototipo funcional para la generación de casos de prueba, aunque su limitación radica en la necesidad de validación exhaustiva para garantizar el cumplimiento con DO-178C. Este enfoque asegura una solución integral adaptada a los requisitos operativos y regulatorios de la industria de la aviación.

1.4 Justificación

Este proyecto es relevante porque aborda una necesidad crítica en la industria aeroespacial: reducir el tiempo y los costos asociados con la generación de casos de prueba, manteniendo altos estándares de calidad. La generación de pruebas mediante *LLMs* puede disminuir significativamente el tiempo de desarrollo de pruebas y mejorar la cobertura de requerimientos. Además, la estandarización de procesos mediante una herramienta innovadora permite cumplir con normativas como DO-178C, garantizando seguridad y eficiencia. Este Trabajo Final de Graduación (TFG) aporta innovación al integrar tecnologías de punta en un sector donde la fiabilidad es crucial, beneficiando a empresas, ingenieros y usuarios finales.

1.5 Viabilidad

A nivel personal, se cuenta con 9 años de experiencia trabajando en la industria de aviación, principalmente, suministrando servicios de verificación de software para ayudar a los clientes a obtener la certificación de sus productos.

Se ha trabajado con múltiples empresas y proyectos, obteniendo experiencia y exposición a una amplia variedad de requerimientos. La experiencia en servicios de verificación y entendimiento de la guía DO-178C respaldan la iniciativa de esta investigación. El criterio de experto permite evaluar los resultados obtenidos y definir una guía para la implementación de los LLMs en la generación de casos de prueba.

1.5.1 Punto de Vista Técnico

El equipo de investigación cuenta con conocimientos en inteligencia artificial, desarrollo de software y metodologías ágiles, adquiridos durante la formación en Ingeniería Informática. Además, se tiene acceso a modelos *LLM* de código abierto (como *LLaMA y Mistral*) y comerciales (como *GPT-4o y GPT-5*), así como a *frameworks* modernos como *LangChain* y *LangGraph*. La capacitación adicional en técnicas de *prompt engineering* se realiza mediante recursos en línea y documentación técnica, asegurando la competencia técnica para ejecutar el proyecto.

1.5.2 Punto de Vista Operativo

El proyecto se lleva a cabo sin interrumpir las operaciones de la empresa aeroespacial. Las pruebas piloto se realizan en un entorno controlado, utilizando datos simulados de requerimientos. La colaboración con ingenieros de pruebas se coordina mediante reuniones virtuales, minimizando la necesidad de recursos físicos.

Los requerimientos a utilizar son ficticios por lo que no hay riesgo de utilizar propiedad privada de los clientes. Dichos requerimientos fueron generados tomando en cuenta la variación en formato y tipos de requerimientos observados en otros proyectos.

1.5.3 Punto de Vista Económico

El costo teórico del proyecto se detalla en la Tabla 1. La mayor parte del costo corresponde a las horas de consultoría del investigador, asumidas por el propio estudiante. Los costos asociados a la colaboración de ingenieros de la empresa son mínimos, ya que se limitan a horas de revisión y validación. Adicional a las horas de investigador y consultoría, se debe pagar un costo por utilizar los *API* de *OpenAI* dado que no son gratuitos. El investigador asume dichos costos.

Tabla 1. *Costos de la investigación*

Rubro	***Costo***	***Responsable***
Horas de consultor (investigador)	*$24,000*	*Investigador*
Capacitación metodológica	*$0 (recursos en línea)*	*Investigador*
Literatura (artículos científicos)	*$200*	*Investigador*
Transportes	*$0 (trabajo remoto)*	*Investigador*
Licencias de software	*$300*	*Investigador*
Equipo computacional	*$2000 (equipo personal)*	*Investigador*

Las horas de consultor se calculan asumiendo 400 horas a $60/hora. Los costos de literatura se cubren mediante acceso a bases de datos académicas. Se estima que el trabajo tardará aproximadamente 5 meses en completarse, con el investigador dedicando 20 horas semanales.

1.6 Objetivos

Los objetivos se redactan utilizando la taxonomía de Bloom revisada, que clasifica los procesos cognitivos en niveles como recordar, comprender, aplicar, analizar, evaluar y crear. Esta taxonomía se selecciona por su claridad para estructurar objetivos desde lo básico hasta lo complejo, asegurando que el proyecto progrese de manera lógica hacia la solución del problema.

1.6.1 Objetivo General

Evaluar la generación de casos de pruebas basados en requerimientos con modelos *Large Language Model* (*LLM*) para acelerar procesos de verificación en la industria aeroespacial.

1.6.2 Objetivos Específicos

- Identificar las limitaciones y desafíos sobre el uso de modelos *LLM* para la generación de casos de prueba de software crítico en la industria aeroespacial.
- Describir el proceso de generación de casos de prueba recomendado por el estándar DO-178C con el fin de establecer una base metodológica para el uso de modelos LLM en la generación de casos de pruebas.
- Clasificar los requerimientos funcionales en categorías según el tipo de entradas y variables, con el fin de establecer una estrategia sistemática de generación de casos de prueba que facilite la evaluación posterior mediante modelos LLM.
- Analizar y comparar, a partir de la literatura existente, diferentes paradigmas de aplicación de los *LLM* para la generación de casos de prueba, evaluando su aplicabilidad a los requerimientos de la industria aeroespacial.
- Diseñar y desarrollar un prototipo funcional y una guía práctica para la integración de modelos *LLM* en la generación de casos de prueba, considerando las normativas DO-178C, aspectos éticos y de seguridad.
- Elaborar una guía de generación de casos de prueba según el tipo de requerimiento analizado que sirva de base para el entrenamiento del personal y desarrollo de herramientas impulsadas por *LLM* en la industria aeroespacial.

1.7 Alcances y Limitaciones

El presente trabajo utiliza datos simulados para garantizar la confidencialidad y reproducibilidad del estudio. Se dispone de ayuda por parte de la empresa Avionyx para la validación de los resultados. La guía y el prototipo están orientados a su contexto operativo y normativo, pero los datos de prueba y requerimientos utilizados no corresponden a proyectos o clientes reales.

1.7.1 Alcances

- Un documento escrito que detalla el marco conceptual para integrar modelos *LLM* en la generación de casos de prueba para sistemas aeroespaciales, incluyendo una revisión sistemática de literatura.

- Una guía práctica para ingenieros de pruebas en la empresa Avionyx, describiendo cómo implementar *LLMs* en procesos de prueba cumpliendo con la normativa DO-178C.

- Un prototipo funcional de una herramienta basada en *LangChain*, Langgraph y *React* que genera casos de prueba automáticamente a partir de requerimientos en lenguaje natural.

- Un informe técnico con los resultados de la evaluación comparativa de métodos de uso de *LLMs*, incluyendo métricas de precisión y exactitud.

1.7.2 Limitaciones

El proyecto no abarca las siguientes áreas:

- Desarrollo de un sistema completo listo para producción, ya que el enfoque es un prototipo de prueba de concepto.

- Validación de los casos de prueba generados en hardware real o entornos operativos, debido a la naturaleza simulada de los datos.

- Integración de la herramienta en los sistemas existentes, ya que esto requiere recursos adicionales fuera del alcance del TFG.

- Evaluación de *LLMs* en dominios no aeroespaciales, enfocándose exclusivamente en sistemas críticos de aeronaves.

1.8 Marco de Referencia Organizacional y Socioeconómico

En este apartado se incluye la historia, tipo de negocio, misión, visión, valores de la empresa que respaldan el estudio y algunas políticas de la empresa que se deben tomar en cuenta para este trabajo de la empresa Avionyx donde se estará implementando la prueba de concepto.

1.8.1 Historia

Fundada en 1989, Avionyx inició sus operaciones en Florida. Tras varios años de crecimiento constante, la empresa se expandió con la apertura de nuevas oficinas en Costa Rica. Impulsada por el desarrollo acelerado y el sólido apoyo de la comunidad local, todas las operaciones se trasladaron en 2008 a la Zona Franca de América. Desde sus inicios, Avionyx se ha especializado en el desarrollo y la verificación de software y firmware de aviónica, cumpliendo rigurosos estándares de la industria como DO-178C, DO-278A y DO-254. La empresa ha apoyado a numerosos fabricantes de aviónica, ofreciendo servicios que abarcan todo el ciclo de vida del desarrollo de software, incluyendo el análisis de requisitos, el diseño, la verificación y el soporte para la certificación. En 2022, Avionyx fue adquirida por Joby Aviation, empresa californiana dedicada al desarrollo de aeronaves eléctricas de despegue y aterrizaje vertical (*eVTOL*).

1.8.2 Tipo de Negocio y Mercado Meta

Avionyx es una empresa especializada en servicios de desarrollo y verificación de software en la industria aeroespacial. Los clientes son empresas responsables del desarrollo de equipos de aviación que necesitan ayuda para certificar sus productos de software. Avionyx brinda todo tipo de soporte necesario para desarrollar software con base en la guía DO-178C.

1.8.3 Misión, Visión y Valores

La misión de Avionyx se define como:
“Avionyx es una empresa de servicios de ingeniería que combina estrictos procesos de calidad, el personal más capacitado y soluciones innovadoras para desarrollar y dar soporte a software, hardware y sistemas de seguridad y misión crítica para la industria aeroespacial. Avionyx se esfuerza por maximizar el valor para sus clientes

y empleados mediante el compromiso, el respeto, la integridad, el profesionalismo y el trabajo en equipo para lograr la satisfacción general y relaciones duraderas.”

La visión de la compañía es:
“Aprovechar nuestra experiencia en servicios de ingeniería para ayudar a nuestros clientes a maximizar la calidad y el valor al ofrecer los productos de aviación más seguros, confiables y sostenibles del mundo.”

Los valores de la empresa son la calidad de servicio, compromiso, respeto, innovación, profesionalismo y trabajo en equipo.

1.8.4 Políticas Institucionales

Los nombres de los clientes y proyectos no se pueden mencionar. Además de las restricciones mencionadas, Avionyx mantiene un conjunto de políticas corporativas alineadas con sus prácticas de ingeniería y cumplimiento normativo:

- Protección de la confidencialidad: no está permitido utilizar información de clientes (requerimientos, especificaciones, documentación) salvo en entornos autorizados y con *NDA* vigentes.
- Cumplimiento normativo y certificatorio: todas las actividades de verificación y desarrollo de software deben adherirse a estándares DO-178C/DO-254, manteniendo trazabilidad completa desde requisitos hasta evidencia de verificación.
- Asistencia a auditorías: los equipos deben estar preparados para presentar evidencia de cumplimiento en auditorías internas y externas (*FAA*, *EASA*).
- Gestión de herramientas calificadas: se exige el uso de herramientas certificadas o calificadas según DO-330, y cualquier herramienta nueva que reemplace al humano en actividades de verificación debe seguir un proceso formal de evaluación y validación.

1.9 Revisión de literatura

1.9.1 Revisión sistemática

La revisión sistemática de literatura se realiza siguiendo la plantilla propuesta por Biolchini, Gomes, Cruz y Horta (2005), que estructura el proceso en preguntas de investigación, fuentes, criterios de selección y análisis.

- **Pregunta de investigación:** ¿Cuáles son los desafíos, mejores prácticas y resultados reportados en la literatura sobre el uso de modelos *LLM* para la generación de casos de prueba en software crítico, particularmente en la industria aeroespacial?

- **Fuentes:** Bases de datos académicas como I*EEE Xplore, ACM Digital Library, Springer y Google Scholar*, complementadas con repositorios de preprints como *arXiv* para estudios recientes.

- **Criterios de inclusión:** Artículos publicados entre 2023 y 2025, en inglés, que aborden *LLMs*, generación de casos de prueba y/o software crítico, priorizando estudios con resultados empíricos o aplicaciones en industrias reguladas (e.g., aeroespacial, médica).

- **Criterios de exclusión:** Artículos sin datos empíricos, centrados en dominios no críticos o que no mencionen *LLMs* explícitamente.

- **Proceso:** Se empleó una búsqueda avanzada en las fuentes académicas mencionadas anteriormente donde se filtró resultados por medio de las siguientes palabras clave: "*verification", "LLM", "Test case generation", "aerospace", "Artificial Intelligence", "requirement based test".* Estos se analizarán en detalle para extraer desafíos, métodos y resultados.

1.9.2 Estado de la cuestión

Dada la capacidad de los LLM para procesar texto, seguir instrucciones y extraer información del lenguaje natural, estos se han investigado como una herramienta para generar casos de prueba a partir de requerimientos textuales en diferentes industrias.

Trabajos previos han demostrado que los LLM son competentes en la generación de texto, la comprensión del lenguaje, el razonamiento aritmético y lógico, así como en la comprensión contextual (Chang et al., 2024).

Korraprolu, Pinninti, y Reddy (2025) examinaron las capacidades de varios modelos para generar casos de prueba utilizando únicamente requerimientos textuales. En esta investigación, se empleó un solo prompt con diferentes modelos (BARD, ChatGPT3.5, Claude, Gemini, ChatGPT4.0 y Llama3) para evaluar los casos de prueba generados por cada uno.

El método de evaluación consistió en generar casos de prueba utilizando LLMs y luego compararlos con los obtenidos mediante Simulink, una herramienta comercial. La calidad de los casos de prueba se evaluó principalmente a través de la cobertura de código, analizando las porciones del código ejecutadas por los casos de prueba generados.

Además, los requerimientos se clasificaron en diferentes categorías con el fin de investigar la relación entre el tipo de requerimiento y la calidad de los casos de prueba correspondientes. Dado que la cobertura se utilizó como la métrica objetiva principal para evaluar la calidad de las pruebas, estas clasificaciones de requerimientos se derivaron con base en la complejidad ciclomática del código asociado a cada requerimiento.

Los resultados de este experimento mostraron que era posible generar casos de prueba utilizando LLMs; sin embargo, el porcentaje de cobertura disminuía conforme aumentaba la complejidad de los requerimientos, lo que indica que los LLMs presentaban dificultades para comprender estructuras lógicas complejas dentro de los requerimientos.

Los hallazgos también revelaron que los LLMs tenían problemas para capturar relaciones entre requerimientos, lo cual limitaba su capacidad para generar casos de prueba con cobertura completa. Además, los requerimientos que involucran múltiples entradas, iteraciones o restricciones de tiempo fueron excluidos del alcance de este estudio.

Los LLMs también han sido utilizados por la industria automotriz para generar casos de prueba a partir de requerimientos textuales. El estudio realizado por Malmberg (2025) propone utilizar GPT-4.0 para generar casos de prueba a partir de los requerimientos incluidos en un documento.

Los documentos se clasifican en tres categorías dependiendo del nivel de complejidad de los requerimientos que contienen. Se consideraron de baja

complejidad aquellos documentos asociados con menos de cinco casos de prueba esperados. Aquellos con seis a dieciséis casos de prueba esperados se clasificaron como de complejidad media, mientras que los documentos con más de dieciséis casos de prueba esperados se categorizaron como de alta complejidad.

En este estudio se utilizó un solo prompt, y los casos de prueba generados se analizaron en términos de cobertura de requerimientos, completitud, corrección y precisión (identificar señales relevantes sin detalles innecesarios).

Los resultados mostraron que el LLM fue capaz de generar pruebas utilizables en el 63 % de los casos evaluados para requerimientos más simples. Consistente con los hallazgos reportados por Korraprolu et al. (2025), la calidad de las pruebas disminuyó a medida que aumentaba la complejidad de los requerimientos. El estudio también confirmó que los LLMs pueden acelerar significativamente el proceso de generación de pruebas, reduciendo el tiempo requerido en un factor de 45 a 180. Sin embargo, debido a limitaciones como alucinaciones, complejidad de los requerimientos y los estrictos estándares de la industria, la supervisión humana sigue siendo necesaria para garantizar que los casos de prueba generados cumplan con los requisitos de la industria.

Aunque la mayoría de los requerimientos textuales describen la lógica del sistema y el comportamiento esperado en lenguaje natural, ciertos requerimientos dependen de diagramas para representar la lógica de máquinas de estado y procesos secuenciales. Para abordar esto, se propuso un método que utiliza LLMs para generar casos de prueba a partir de diagramas de máquinas de estado en la industria aeroespacial (Su et al., 2024).

En este método se siguieron cinco pasos:

1. Extraer criterios de seguridad de los estándares de aviación (por ejemplo, DO-178C, ARP-4754A).
2. Extender las máquinas de estado con los criterios de seguridad utilizando LLMs.
3. Generar una nueva máquina de estado con los criterios de seguridad incorporados.
4. Realizar un recorrido de profundidad primero por los bordes del diagrama extendido para crear rutas de prueba.
5. Usar LLMs para optimizar los hiper parámetros y los procesos de mutación en algoritmos genéticos.

El método propuesto automatiza de manera efectiva la extensión de los diagramas de estado de seguridad y la generación de casos de prueba, reduciendo la

dependencia de la experiencia del *tester* y mejorando la calidad de las pruebas. Además, redujo en un 90 % el número máximo de iteraciones al utilizar el algoritmo genético.

Considerando que los LLMs también han sido entrenados con datos de código y lenguajes de programación, distintos estudios han utilizado esta característica para evaluar el desempeño de los LLMs en la generación de casos de prueba para pruebas unitarias.

Li y Yuan (2024) llevaron a cabo un experimento para evaluar los casos de prueba generados a partir de la descripción de una función, en el cual se emplearon varios LLMs (StarChat, CodeLLama, GPT-3.5, GPT-4). En este estudio, se utilizaron dos conjuntos de datos de preguntas y respuestas de código para evaluar la corrección de las respuestas generadas por los LLMs. Los resultados indicaron que, aunque los modelos producían respuestas variadas para ejemplos simples, los modelos más avanzados (por ejemplo, GPT-3.5 y GPT-4) generaban casos de prueba adicionales para escenarios complejos.

Se propuso un método alternativo basado en el *prompting* ReAct a través del *framework* TestChain, que emplea un LLM para la identificación de entradas de prueba y otro para la predicción de salidas. Al dividir las tareas entre modelos, este enfoque introdujo especialización en los distintos módulos del LLM. Los resultados mostraron que aplicar TestChain con GPT-4 mejoró el desempeño en un 13,84 % en comparación con el uso de GPT-4 de manera aislada sobre el mismo conjunto de datos (Li & Yuan, 2024).

1.10 Métricas de Evaluación

Las métricas que se van a emplear para llevar a cabo la evaluación son precisión, sensibilidad y puntuación F1.

$$Precisión = \frac{TP}{TP + FP} \quad (1)$$

$$Sensibilidad = \frac{TP}{TP + FN} \quad (2)$$

$$Puntuación\ F1 = 2 \times \left(\frac{Precisión \times Sensibilidad}{Precisión + Sensibilidad}\right) \quad (3)$$

Las ecuaciones (1), (2) y (3) muestran la implementación de cada métrica en donde *TP* representa el número de casos de prueba correctos, *FP* el número de casos de prueba sugeridos por el *LLM* que son incorrectos y *FN* el número de casos de prueba faltantes.

Los casos de prueba sugeridos por los *LLM* serán analizados utilizando dichas métricas para determinar cuál paradigma de implementación y *LLM* obtuvo mejores resultados, y para cuantificar los casos faltantes e incorrectos.

1.11 Diseño del Estudio de Caso

Para llevar a cabo esta evaluación, es necesario definir un *benchmark* de requerimientos con sus casos de prueba esperados correspondientes. Las descripciones de los requerimientos pueden incluir operadores booleanos, múltiples condiciones, máquinas de estados, algoritmos, funcionalidades de temporización, etc.

Se proponen tres categorías para la clasificación de requerimientos:

- Requerimientos con condiciones: Incluyen entradas y condiciones bajo las cuáles se deben implementar ciertas acciones.
- Requerimientos sin condiciones: el sistema debe implementar las acciones documentadas independientemente de condiciones o estado de las entradas.
- Requerimientos con algoritmos u operaciones matemáticas: involucran operaciones matemáticas, ecuaciones, algoritmos, etc.

También se requiere una lista de verificación para establecer los criterios que se seguirán al evaluar las respuestas de los LLM basándose en la guía DO-178C. La Tabla 2 muestra la lista que se utilizará para la verificación de casos de prueba sugeridos por los *LLM.*

Tabla 2. *Criterios de evaluación de los casos de prueba*

Criterio ID	Criterio de evaluación
1	Los casos de prueba de verificación incluyen representaciones de clases de equivalencia para las entradas.
2	Los casos de prueba verifican correctamente los límites.
3	Los casos de prueba verifican correctamente los valores mínimos y máximos del rango.
4	Los casos de prueba toman en consideración aspectos relacionados con variables que involucran tiempo.
5	Los casos de prueba cubren operaciones Booleanas.
6	Los casos de prueba incluyen el resultado esperado.
7	Cuando es posible los casos de prueba incluyen representaciones de clases de equivalencia para entradas fuera de los límites.

8	Los casos de prueba predicen correctamente los resultados esperados basándose en la lógica descrita en el requerimiento y en los valores de entrada especificados en cada caso de prueba.

Fuente: Confección propia

Se utilizan tres métricas cuantitativas—precisión, puntuación F1 y exhaustividad—para analizar las respuestas de los LLM. Para los casos de prueba faltantes o incorrectos, se documentan los tipos de errores identificados con el fin de proporcionar una evaluación cualitativa del rendimiento del LLM.

Para llevar a cabo esta evaluación, se consideran dos experimentos diferentes. El primero consiste en evaluar las distintas combinaciones de LLMs y paradigmas de aplicación utilizando los requerimientos de para comparar los resultados entre los diferentes enfoques y modelos. El segundo es un experimento similar, pero esta vez añadiendo otra variable: las respuestas a las mismas preguntas por parte de ingenieros de verificación con diferentes niveles de experiencia en la industria aeroespacial.

El *benchmark* de requisitos se dividirá en dos grupos: uno para la comparación entre LLMs y otro para la comparación entre LLMs y humanos. El número de requisitos utilizado para la comparación LLM–humano es significativamente menor que el utilizado para la comparación solo entre LLMs, ya que la evaluación implica administrar una prueba a los participantes. Aunque el tipo y nivel de complejidad de los requisitos serán equivalentes en ambos experimentos, los requisitos específicos empleados serán diferentes.

Capítulo 2. Marco Conceptual

Dado el enfoque alternativo de este trabajo, se opta por un marco conceptual, que consiste en la definición y clarificación de los principales conceptos que estructuran la investigación sobre la integración de modelos de lenguaje grande (*LLMs*) en la generación de casos de prueba para software crítico en la industria aeroespacial.

2.2 Metodología para la elaboración del marco conceptual

Siguiendo las recomendaciones metodológicas de la Universidad Cenfotec, este marco conceptual se elabora a partir de:

- **Revisión sistemática de literatura:** Se identificaron los conceptos clave relacionados con *LLMs*, pruebas basadas en requerimientos, verificación, DO-178C, y paradigmas de aplicación de los *LLMs*.

- **Nube y jerarquización de conceptos:** Se construyó una nube de conceptos y un mapa conceptual jerárquico para organizar los términos desde los más generales hasta los más específicos.

- **Definición top-down:** Cada concepto se presenta y desarrolla de lo general a lo particular, asegurando coherencia y fluidez en la exposición.

- **Criterios de fuentes:** Las definiciones y discusiones de cada concepto provienen de literatura académica y normativas reconocidas, conforme a criterios de rigor científico y autoridad.

2.3 Conceptos fundamentales

La DO-178C es la norma internacional que regula el desarrollo y la certificación de software para sistemas aeronáuticos. Exige procesos rigurosos de verificación y validación, asegurando la trazabilidad entre requerimientos, código y pruebas. Cumplir con DO-178C es obligatorio para obtener certificaciones como la de la *FAA* o *EASA*. (Visure, 2025).

La prueba basada en requerimientos es el enfoque principal en industrias reguladas, ya que garantiza que cada requerimiento del sistema sea verificado a través de uno o más casos de prueba específicos. Según la guía DO-178C, este método es esencial para lograr trazabilidad y cobertura, aspectos críticos en sistemas aeroespaciales (RTCA, 2011).

Las pruebas basadas en requerimientos son el método preferido por DO-178C ya que se ha demostrado ser el más efectivo para la detección de errores (RTCA, 2011). Este proceso comienza con un fase previa que es la generación de casos de prueba. Este proceso consiste en identificar los escenarios que van a ejecutarse en las pruebas basado en un análisis de los requerimientos. Los casos de prueba son un resumen de dichos escenarios y usualmente describen información de los valores de entrada de los requerimientos y las salidas esperadas en el sistema basado en la lógica y funcionalidad del requerimiento analizado. La guía DO-178C establece los criterios a tomar en cuenta para seleccionar la cantidad mínima de casos de prueba.

Esta fase inicial es una tarea repetitiva, laboriosa y depende en gran medida de la experiencia y práctica del ingeniero de verificación. Por esta razón, se motiva a analizar la efectividad de herramientas que asistan al ingeniero en esta tarea.

Los *Large Language Models (LLMs)* son modelos de inteligencia artificial diseñados para la arquitectura de *transformers* que les permite predecir la siguiente palabra con mayor probabilidad de aparición en una secuencia de datos. Estos modelos son entrenados con grandes volúmenes de texto, capaces de comprender y generar lenguaje natural de manera autónoma. Se han utilizado para tareas de traducción de lenguajes, resumen y análisis de texto, y conversaciones(Geroimenko, 2025).

Debido a que la gran mayoría de requerimientos son representados de manera textual, los *LLM* tienen potencial de asistir a los ingenieros de verificación en este tipo de tareas.

Existen diferentes métodos para emplear los *LLM* en el desarrollo de tareas requeridas por el usuario. Estos métodos se presentan en este trabajo como

paradigmas de aplicación. En esta investigación, se analizarán tres métodos: *prompts, RAG y* agentes.

Los prompts son las entradas e instrucciones utilizadas para dirigir a los *LLM* a realizar las tareas solicitadas por el usuario. Existen buenas prácticas y técnicas de prompt engineering que pueden emplearse para mejorar la calidad de las respuestas generadas por los *LLM (*Geroimenko, 2025*)*.

RAG es un marco de trabajo para LLM que combina un método de recuperación con un generador de secuencia a secuencia que utiliza los *LLM* para predecir la siguiente palabra de la secuencia (Oche et al., 2025). Este método permite que el *LLM* amplíe su base de conocimientos hacia datos para los cuales no fue entrenado. Este enfoque reduce las alucinaciones y mejora la precisión y la relevancia de las respuestas de los *LLM* dentro de dominios específicos del conocimiento.

Los agentes LLM tienen la capacidad de percibir su entorno y tomar decisiones. Los agentes poseen características que les permiten manejar solicitudes más complejas y tareas de resolución de problemas durante el proceso de aplicación. Un agente tiene características individuales que provienen del conocimiento que se le proporcionó al modelo, de sus roles y de sus objetivos. Recibe información de herramientas o fuentes externas para comprender el entorno que lo rodea. Además, posee la capacidad de planificar acciones basadas en su percepción del entorno y su estado interno, lo que le permite ejecutar tareas que contribuyen a alcanzar sus objetivos (As Li et al., 2024).

Una limitación y problema de los *LLM* es que, en algunos casos, generan información incorrecta. Este fenómeno se llama alucinación y ocurre porque los *LLM* predicen palabras basándose en patrones que aprendieron durante su entrenamiento. Si el usuario realiza una pregunta que el modelo nunca ha visto antes, hará una conjetura estadística basada en los patrones observados durante su entrenamiento. Este proceso puede llevar a la producción de información inexacta que parece coherente y lógicamente consistente, a pesar de ser incorrecta (Coursaris et al., 2025).

2.5 Conclusión del Marco Conceptual

Este marco conceptual proporciona las bases necesarias para entender los desafíos, oportunidades y consideraciones éticas en la aplicación de *LLMs* en el proceso de generación de pruebas en la industria aeroespacial. Sirve como referencia para el análisis y discusión de los resultados obtenidos en el presente trabajo, así como para la formulación de recomendaciones y futuras líneas de investigación.

Capítulo 3. Marco Metodológico

Este capítulo describe la metodología adoptada para evaluar *LLMs* en pruebas aeroespaciales. Se articula un enfoque alternativo con un tipo de investigación aplicada. Se plantea realizar una evaluación de los modelos *LLMs* a partir de casos de estudio.

3.1 Tipo de Investigación

La investigación es aplicada y orientada a la solución de un problema práctico en un contexto regulado (DO-178C). Se asume la teoría de la complejidad y un pragmatismo metodológico que permite combinar mediciones objetivas y valoraciones expertas para obtener evidencia accionable.

3.2 Alcance Investigativo

El alcance es **descriptivo y exploratorio**. Además de lo señalado en los objetivos, se abordará explícitamente:

Descripción (estado actual y caracterización):

- **Proceso “as-is”** de identificación manual de casos de prueba en Avionyx (roles, artefactos, herramientas, tiempos por fase).

- **Tipología de requerimientos**: estructurados (p. ej., **EARS**) vs. no estructurados; atributos como ambigüedad, longitud y criticidad.

- **Métricas base**: tiempos actuales, precisión, exactitud, criterios de aceptación.

Exploración (oportunidades y limitaciones específicas):

- **Oportunidades**: reducción de tiempo en la identificación inicial de casos; mejora de cobertura por requerimiento; plantillas y prompts reutilizables; registro auditable de decisiones (prompts, versiones).

- **Limitaciones**: ambigüedad semántica, alucinaciones, reproducibilidad de resultados, cumplimiento DO-178C/DO-330 en el uso de herramientas, sesgos del modelo, seguridad y confidencialidad de datos, costos y latencia.

Resultados esperados del alcance: mapa de condiciones en que los *LLMs* aportan valor, umbrales mínimos de desempeño para su adopción parcial y lineamientos de integración con los procesos existentes.

3.3 Enfoque

El enfoque es alternativo ya que se enfoca en la evaluación de diferentes modelos y paradigmas de aplicación de LLMs por medio de casos de estudio. En este análisis, los resultados que se desean medir no pueden ser cubiertos únicamente con métricas cuantitativas ya que hay otros factores a tomar en cuenta tales como la calidad de las pruebas,evaluación de capacidades emergentes, coherencia,etc,

3.4 Diseño

El diseño de caso aplicado, flexible y adaptativo, con iteraciones tipo sprint. En cada iteración se comparan resultados, se ajustan prompts y flujos y se documentan decisiones.

3.5 Población y Muestreo

- **Población de referencia**: personal técnico de Avionyx vinculado a verificación de software en entornos DO-178C (ingenieros de verificación, de software y QA/Procesos).

- **Muestreo**: **no probabilístico por conveniencia**.

Cuantificación estimada *(ajustable según disponibilidad)*:

- **n ≈ 6–12 participantes**, distribuidos de la siguiente manera:
    - **6–10** ingenieros/as de verificación.
- **Criterios de inclusión**: experiencia en DO-178C y participación en pruebas basadas en requerimientos.
- **Criterios de variación**: diversidad en años de experiencia, proyecto y rol.
- **Entrevistas cualitativas**: hasta alcanzar **saturación temática** (estimado 4–8 entrevistas).

3.6 Instrumentos de Recolección de Datos

(a) Matrices de evaluación (con rúbricas y pesos)

Se aplicarán a cada requerimiento y al conjunto:

1. **Precisión (0–1)**
   Porcentaje de casos generados que verifican correctamente el requerimiento.
    - **1.0 - 0.80**:alta; **0.79 – 0.7**: media; **<0.7**: baja.
2. **Sensibilidad (0–1)**
   Porcentaje de casos generados que verifican correctamente el requerimiento con respecto a los casos faltantes no identificados por los *LLM*.
    - **1.0 - 0.80**:alta; **0.79 – 0.7**: media; **<0.7**: baja.
3. **Puntuación F1 (0–1)**
   Balance entre precisión y sensibilidad. Combina ambas métricas en un solo número para obtener una representación general del desempeño.
    - **1.0 - 0.80**:alta; **0.79 – 0.7**: media; **<0.7**: baja.

(b) Entrevistas abiertas/semiestructuradas *(30–45 min)*

Ejes y ejemplos de preguntas:

- **Utilidad**: “¿En qué escenarios considera que los LLM son útiles para generar casos de prueba y en cuáles no? ¿Por qué?”

- **Carga cognitiva y flujo de trabajo**: “¿El LLM reduce su carga o introduce pasos extra?”
- **Riesgos/ética**: “¿Qué riesgos ve para seguridad/cumplimiento?”
- **Adopción**: “¿Qué umbrales y controles pondría para uso en producción?”

(c) Encuestas

Ítems Likert (1–5) sobre: utilidad percibida, facilidad de uso, claridad de casos, intención de uso.

(d) Registros de tiempo y eficiencia

- **Definición operacional**: tiempo desde recepción del requerimiento hasta tener lista la **lista de casos** (no ejecución).
- **Comparación**: manual vs. LLM por requerimiento.

3.7 Técnicas de Análisis de Información

- **Cuantitativo**: estadísticas descriptivas (precisión, sensibilidad), comparación pareada manual vs. LLM (ganancia de tiempo y diferencia en precisión/sensibilidad).
- **Cualitativo**: Clasificación de errores, relación entre métricas y tipos de requerimientos.

Capítulo 4. Análisis del Diagnóstico

El análisis de diagnóstico constituye una investigación previa sobre el tema de generación de casos de prueba en la industria aeroespacial con el fin de considerar aspectos importantes previo al desarrollo de este trabajo.

Dentro de los temas a considerar se encuentran los criterios de aceptación de los casos de prueba, normativas dentro de la industria aeroespacial, regulaciones del uso de inteligencia artificial e impacto en la calidad y seguridad de las pruebas.

La cualificación de herramientas es necesaria cuando las actividades de desarrollo o verificación de software se reducen, eliminan o automatizan mediante herramientas cuyo resultado no es verificado según lo descrito en la guía DO-178C (RTCA, 2011a).

Los errores introducidos por este tipo de herramientas pueden afectar negativamente la funcionalidad del software y la seguridad del sistema. El proceso

de cualificación de herramientas proporciona guías para el desarrollo y la verificación de herramientas con el fin de asegurar que su integridad y funcionalidad prevista no se vean comprometidas (RTCA,2011b). El proceso y los objetivos que deben seguirse están documentados en la DO-330.

Durante este proceso, se debe demostrar que las salidas de las herramientas cumplen con los objetivos descritos en la DO-330. Las salidas deben ser consistentes cuando se proporcionan las mismas entradas en el mismo entorno. Para las salidas que muestran variaciones, se debe demostrar que no tienen un impacto negativo en el proceso que utiliza dichas salidas.

Dado que los LLM generan contenido basado en probabilidades, las herramientas impulsadas por LLM no deben considerarse para reemplazar, eliminar o automatizar ningún proceso de verificación de software, ya que no es posible obtener salidas determinísticas.

La FAA ha establecido una hoja de ruta que define los principios rectores para garantizar la seguridad de la inteligencia artificial en la aviación (Federal Aviation Administration, 2024). La hoja de ruta enfatiza la importancia de centrarse en la garantía de la seguridad y en las mejoras de seguridad, particularmente en áreas donde las aplicaciones de IA tienen el potencial de mejorar la seguridad operativa. También destaca la importancia de evitar la personificación de los sistemas de IA, ya que estas tecnologías no deben considerarse ni tratarse como seres humanos. Las decisiones críticas y el control total de sistemas o procesos no deben delegarse a herramientas de IA. En su lugar, los diseñadores e ingenieros deben diferenciar claramente las responsabilidades asignadas a los sistemas de IA de aquellas asignadas a los operadores humanos para evitar cualquier compromiso con la seguridad.

La FAA recomienda adoptar un enfoque incremental para integrar la IA en los procesos de aviación. Esta estrategia progresiva permite a los especialistas y a la FAA acumular conocimiento y experiencia a partir de implementaciones de menor escala antes de desplegar sistemas basados en IA en dominios más críticos para la seguridad o de mayor impacto.

La guía DO-178C establece los criterios de aceptación que deben utilizarse como marco de referencia para la evaluación de resultados obtenidos por medio de *LLM.* El objetivo es analizar los resultados con mayor objetividad para determinar el valor aportado por herramientas que utilizan inteligencia artificial.

El análisis de diagnóstico revela la situación actual de las normativas en la industria aeroespacial con respecto a la verificación de software e implementación de herramientas impulsadas por IA en torno a esta actividad.

Se concluye que debido al impacto que tiene las actividades de verificación de software, la naturaleza probabilística de los *LLM* y las sugerencias por parte de la FAA, no se debe diseñar herramientas que busquen automatizar actividades de verificación.

El ser humano tiene la decisión final y es responsable por cada prueba desarrollada. Por lo tanto, este trabajo se enfoca en cómo herramientas impulsadas por IA pueden asistir al ser humano y agilizar los procesos de verificación sin comprometer la calidad de las pruebas.

También, se seleccionó un grupo de ingenieros al cual se le realizaron pruebas preliminares para analizar el tipo de ayuda que este tipo de herramientas debe brindar.

Los ingenieros se dividieron en tres categorías dependiendo de su nivel de experiencia en el área de verificación: menos de 1 año, de 1 año a 3 años y más de 3 años. El objetivo es identificar relaciones entre la experiencia y la calidad de casos de prueba identificados.

Se utilizaron las mismas métricas de precisión y sensibilidad descritas en este documento, así como el tiempo que se necesitó para identificar los casos de prueba.

Los resultados preliminares muestran que hay una relación entre el grado de experiencia de los ingenieros y la calidad de pruebas identificadas. Los ingenieros con más de tres años de experiencia obtuvieron una puntuación F1 promedio de 0.88. Los ingenieros entre uno y tres años de experiencia obtuvieron una puntuación F1 promedio de 0.8, mientras que los ingenieros con menos de un año de experiencia obtuvieron una calificación de 0.55.

Capítulo 5. Propuesta de Solución

Con base en la investigación realizada en este documento, se presenta a continuación, la propuesta para cumplir con los objetivos de este trabajo.

Se realizará una evaluación de tres paradigmas de aplicación de los LLMs para asistir a los ingenieros en la generación de casos de prueba. Los paradigmas de aplicación son: RAG, *prompts* y agentes.

Se evaluarán las respuestas obtenidas utilizando el checklist de la Tabla 2 y de acuerdo a este criterio de evaluación se emplearán las métricas de precisión, sensibilidad y puntaje F1 para analizar el desempeño de cada uno de los paradigmas y modelos empleados.

Los resultados obtenidos se compararán con las evaluaciones previas que se realizaron a los ingenieros con el fin de determinar el impacto que estas herramientas pueden tener en la productividad de los ingenieros. Adicionalmente, se realizarán entrevistas para analizar el impacto que tienen estas herramientas en el trabajo de los ingenieros.

Finalmente, se propone realizar un diseño de una aplicación que utilice *LLMs* para asistir a los ingenieros en la generación de casos de prueba. El diseño de esta aplicación debe estar orientado a asistir al ingeniero y no reemplazarlo o automatizar estas tareas para cumplir con los estándares requeridos. El humano debe estar en flujo de trabajo de la aplicación para revisar y modificar los casos de prueba sugeridos por la aplicación.

Capítulo 6. Conclusiones de trabajo PIA-01

El trabajo realizado presenta una investigación que propone cómo emplear tecnologías emergentes como LLMs para asistir a los ingenieros de la industria aeroespacial en generación de casos de prueba.

La metodología utilizada propone un método que combina métricas y resultados cuantitativos con un análisis cualitativo de la calidad y validez de los casos de prueba. Esta metodología busca analizar el desempeño de diferentes métodos para asistir al ingeniero con los casos de prueba y comparar estos resultados con el rendimiento actual de un grupo de ingenieros que serán parte del grupo de estudio.

Se realizó una investigación previa de los paradigmas de aplicación de los *LLMs* que se plantean utilizar durante la evaluación. El objetivo es analizar ventajas y desventajas de cada aplicación que serán criterios a considerar para el desarrollo del prototipo funcional.

Se definió el alcance y objetivo del prototipo que puede ayudar a los ingenieros de verificación con el fin de no comprometer la calidad de las pruebas y, así mismo, monitorear y corregir cualquier error introducido por la aplicación.

Glosario

IA - Inteligencia Artificial

FAA - Federal Aviation Administration

LLM - Large Language Model

RAG - Retrieval Augmented Generation

Referencias

1. Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., Ye, W., Zhang, Y., Chang, Y., Yu, P. S., Yang, Q., & Xie, X. (2024). A survey on evaluation of large language models. ACM Transactions on Intelligent Systems and Technology, 15(3), Article 39. https://doi.org/10.1145/3641289.
2. Korraprolu, B. R., Pinninti, P., & Reddy, Y. R. (2025). Test case generation for requirements in natural language – An LLM comparison study. In Proceedings of the 18th Innovations in Software Engineering Conference (ISEC '25) (Article 9, pp. 1–5). Association for Computing Machinery. https://doi.org/10.1145/3717383.3717389
3. Malmberg, H. (2025). Automated Test Generation From Software Requirements Using an LLM : Generating Black-Box Test Cases From Real-World Automotive Software Requirements With GPT-4o (Dissertation). Obtenido de https://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-368031
4. Su, Q., Li, X., Ren, Y., Ouyang, X., Hu, C., & Yin, Y. (2024). State diagram extension and test case generation based on large language models for improving test engineers' efficiency in safety testing. In 2024 IEEE 35th International Symposium on Software Reliability Engineering Workshops (ISSREW) (pp. 287–294). IEEE. https://doi.org/10.1109/ISSREW63542.2024.00092
5. Li, K., & Yuan, Y. (2024). Large language models as test case generators: Performance evaluation and enhancement. arXiv. https://doi.org/10.48550/arXiv.2404.13340.
6. RTCA, Incorporated. (2011). *DO-178C: Software considerations in airborne systems and equipment certification.* RTCA, Inc.
7. Visure Solutions. (s.f.). *¿Qué es DO-178C?* Recuperado el 28 de octubre de 2025, de https://visuresolutions.com/aerospace-and-defense/do-178c/
8. Geroimenko, V. (2025). The Essential Guide to Prompt Engineering: Key Principles, Techniques, Challenges, and Security Risks. Springer Nature Switzerland AG. https://doi.org/10.1007/978-3-031-86206-9

9. Oche, A. J., Folashade, A. G., Ghosal, T., & Biswas, A. (2025). A systematic review of key retrieval-augmented generation (RAG) systems: Progress, gaps, and future directions (arXiv preprint arXiv:2507.18910). https://doi.org/10.48550/arXiv.2507.18910.
10. Li, X., Wang, S., Zeng, S., Wu, Y., & Yang, Y. (2024). A survey on LLM-based multi-agent systems: Workflow, infrastructure, and challenges. Vicinagearth, 1, Artículo 9. https://doi.org/10.1007/s44336-024-00009-2.
11. Coursaris, C. K., Beringer, J., Léger, P.-M., & Öz, B. (Eds.). (2025). The design of human-centered artificial intelligence for the workplace (Vol. 7933, Progress in IS). Springer Nature Switzerland AG. https://doi.org/10.1007/978-3-031-83512-4
12. RTCA. (2011). *DO-330: Software tool qualification considerations*. RTCA, Incorporated.
13. Federal Aviation Administration. (2024, July 23). *Roadmap for artificial intelligence safety assurance – Version I*. https://www.faa.gov/media/82891