UAB
Universitat Autònoma de Barcelona

Dipòsit digital de documents de la UAB

This is the **published version** of the bachelor thesis:

Bello Marabé, Álvaro. *Generació i Validació d'Hipòtesis Arqueològiques mitjançant LLM i SRTM*. Treball de Final de Grau (Universitat Autònoma de Barcelona), 2026 (Grau en Enginyeria de Dades)

This version is available at https://ddd.uab.cat/record/337342

under the terms of the  license.

Generació i Validació d'Hipòtesis Arqueològiques mitjançant LLM i SRTM

Álvaro Bello Marabé

Resum— Aquest treball presenta el disseny i la implementació d'un sistema que combina models de llenguatge de gran escala (LLM) amb anàlisi geoespacial basada en dades d'elevació SRTM per generar i validar hipòtesis sobre la localització d'assentaments arqueològics precolombins a l'Amazònia sud-occidental. El sistema analitza automàticament cròniques històriques dels segles XVI–XIX mitjançant Claude Haiku 4.5, extreu entitats geogràfiques amb coordenades estimades, les geocodifica en cascada contra un gazetteer colonial especialitzat (HGIS de las Indias, 15.847 llocs) i les puntua amb un sistema multicriteri de vuit senyals. A partir de cinc fonts textuals primàries s'han extret 1.284 entitats que generen 302 hipòtesis deduplicades, de les quals 130 candidates assoleixen un *hit rate* del 43,1% contra 1.708 jaciments documentats del dataset de Walker et al. (2023). L'eina de visualització interactiva desenvolupada amb Streamlit permet explorar les hipòtesis sobre un mapa geogràfic i inclou un agent conversacional basat en LLM.

Paraules clau— LLM, LiDAR, SRTM, arqueologia amazònica, SIG, NER, geocodificació, gazetteer històric, *pipeline* de dades, teledetecció.

Abstract— This work presents the design and implementation of a system combining Large Language Models (LLM) with SRTM elevation-based geospatial analysis to generate and validate hypotheses about pre-Columbian archaeological settlements in the southwestern Amazon. The system processes 16th–19th century historical chronicles using Claude Haiku 4.5, extracts geographic entities with estimated coordinates, geocodes them against a colonial gazetteer (HGIS de las Indias, 15,847 places), and scores them with an eight-signal multi-criteria system. From five primary sources, 1,284 entities were extracted, generating 302 hypotheses, of which 130 candidates achieve a 43.1% hit rate against 1,708 documented sites from Walker et al. (2023). An interactive Streamlit interface allows map-based exploration and natural language queries via an LLM agent.

Keywords— LLM, LiDAR, SRTM, Amazon archaeology, GIS, NER, geocoding, historical gazetteer, data pipeline, remote sensing.

1 Introducció

En els darrers anys, l'ús de tecnologies de teledetecció ha canviat força la manera d'estudiar paisatges antics i els patrons d'assentament de moltes societats del passat. Una de les tecnologies que més impacte ha tingut en aquest camp és el LiDAR (*Light Detection and Ranging*), que permet obtenir models molt detallats del relleu del terreny fins i tot en zones amb vegetació molt densa. Aquests sistemes emeten polsos de llum làser que penetren la coberta forestal i permeten reconstruir el terreny nu amb una resolució d'1 metre o menys. Aquest tipus d'eines han resultat especialment útils en regions com l'Amazònia, on durant molt de temps es va pensar que les condicions ambientals feien difícil l'existència de grans assentaments humans.

Tot i això, diversos estudis recents han començat a qüestionar aquesta idea. Per exemple, el treball de Prümers et al. [1] mostra com l'ús de dades LiDAR ha permès identificar estructures i xarxes d'assentaments que fins ara havien passat desapercebudes sota la selva. Altres investigacions han utilitzat models espacials i tècniques d'anàlisi geogràfica per intentar predir on es podrien trobar nous jaciments encara no documentats [2].

En conjunt, aquests resultats han contribuït a replantejar la manera com s'interpreta l'organització territorial de les societats precolombines a l'Amazònia.

Al mateix temps, els progressos recents en intel·ligència artificial han obert noves possibilitats per analitzar grans quantitats d'informació textual. En particular, els models de llenguatge de gran escala (LLM, *Large Language Models*) permeten processar i analitzar textos de manera automàtica, cosa que pot ser útil per estudiar fonts històriques com cròniques colonials, relats d'expedicions o documents

• E-mail de contacte: 1665634@uab.cat
• Treball tutoritzat per: Ana Oropesa (DFIC)
• Curs 2025/26

etnogràfics. Aquest tipus de fonts sovint contenen descripcions del territori o referències geogràfiques que podrien estar relacionades amb antics assentaments o rutes de poblament.

Malgrat el potencial d'aquestes dues línies de recerca, sovint s'han desenvolupat per separat. D'una banda, els estudis basats en LiDAR acostumen a centrar-se sobretot en l'anàlisi del relleu i en la identificació d'estructures o anomalies al terreny. De l'altra, els treballs de processament del llenguatge natural aplicats a l'arqueologia analitzen textos històrics, però no sempre els connecten amb eines d'anàlisi geoespacial.

Aquesta separació fa interessant explorar aproximacions que combinin ambdues fonts d'informació.

Cal destacar que, tot i que en la introducció hem parlat del LiDAR per la seva rellevància a l'estat de l'art, per al component topogràfic del sistema s'han utilitzat dades SRTM (*Shuttle Radar Topography Mission*) en lloc de LiDAR, per raons de cobertura global i accés lliure. Aquesta decisió es justifica detalladament a la Secció 4.

2 Estat de l'art

Els treballs més propers s'agrupen en dues línies. La primera engloba estudis de detecció de jaciments per teledetecció: Wagner et al. [3] proposen mètodes de càlcul ràpid d'anomalies de model digital de terreny per detectar geoglifs, mentre que Peripato et al. [4] estimen que més de 10.000 *earthworks* precolombins romanen ocults. La segona utilitza aprenentatge automàtic per predir localitzacions: Walker et al. [2] entrenen un model de *gradient boosting* amb variables ambientals, obtenint una cobertura del 85% dels jaciments documentats. En l'àmbit del processament de text, Hodel et al. [12] van demostrar que models de NER —*Named Entity Recognition*, és a dir, tècniques de reconeixement automàtic d'entitats nombrades com llocs, persones i dates dins de textos— adaptats a documents històrics superen àmpliament els models genèrics en la identificació de topònims colonials.

El present treball es diferencia d'aquests enfocaments en un aspecte fonamental: no requereix dades de jaciments coneguts per entrenar cap model. En lloc d'això, utilitza fonts textuals primàries per generar hipòtesis de manera independent, cosa que el fa especialment útil per a regions poc explorades arqueològicament.

Per mantenir el projecte dins d'un abast realista per a un treball de final de grau, l'estudi es limitarà a una zona concreta de l'Amazònia sud-occidental. Aquesta regió s'ha escollit perquè disposa de dades de teledetecció accessibles, compta amb jaciments arqueològics documentats que poden servir com a referència (`zenodo.org/record/7672448`) [2] i també té fonts històriques digitalitzades relacionades amb el territori (`archive.org`) [7, 8, 9, 10].

3 Objectius

L'objectiu general d'aquest treball és explorar fins a quin punt la combinació de models de llenguatge de gran escala (LLM) i dades geoespacials derivades de SRTM pot ajudar a generar i validar noves hipòtesis sobre la possible localització d'assentaments arqueològics a l'Amazònia.

A partir d'aquest objectiu principal, el treball es planteja els objectius específics següents:

1. Analitzar fonts textuals històriques i arqueològiques per identificar referències geogràfiques potencialment relacionades amb assentaments precolombins.
2. Utilitzar un model de llenguatge de gran escala per extreure i estructurar automàticament aquesta informació espacial a partir dels textos.
3. Integrar aquesta informació amb dades geoespacials derivades de SRTM per analitzar les característiques topogràfiques de les zones identificades.
4. Desenvolupar un procés sistemàtic que permeti generar i prioritzar hipòtesis de possibles localitzacions arqueològiques.
5. Comparar els resultats obtinguts amb registres arqueològics existents per valorar el potencial d'aquest enfocament.
6. Construir una eina de visualització interactiva que permeti explorar les hipòtesis generades sobre un mapa geogràfic.

En conjunt, aquest treball pretén demostrar el potencial d'integrar models de llenguatge i dades de teledetecció a la investigació arqueològica.

4 Planificació

El projecte té una durada total de divuit setmanes i s'organitza en sis fases que es desenvolupen de manera seqüencial, tot i que la redacció de la memòria és una tasca transversal que acompanya tot el procés des del principi.

La primera fase, de tres setmanes, es dedica a la planificació inicial: revisió de la bibliografia existent sobre arqueologia amazònica, teledetecció SRTM i aplicacions de models de llenguatge en humanitats digitals, així com la delimitació concreta de l'àrea geogràfica d'estudi i la selecció del corpus textual que s'utilitzarà. A continuació, durant dues setmanes, es dissenya l'arquitectura general del sistema: quins mòduls el formaran, quines eines i models s'empraran i com es connectaran entre ells.

La fase de desenvolupament és la més extensa, amb set setmanes, i consisteix en la implementació progressiva de cada mòdul del *pipeline*: l'extracció d'entitats mitjançant LLM, la geocodificació, el preprocessament de les dades SRTM, el càlcul de mètriques topogràfiques i el sistema de puntuació de les hipòtesis. Cada mòdul es prova de manera independent abans de passar al següent. Posteriorment, durant quatre setmanes, s'integren tots els components, es validen els resultats obtinguts i es desenvolupa la interfície de visualització interactiva.

Finalment, les darreres setmanes es reserven per a la preparació de la presentació oral i la defensa davant el tribunal. La distribució temporal de totes aquestes fases es recull al

diagrama de Gantt de l'Apèndix A.1.

5 CONJUNT DE DADES

En aquesta secció es descriuen els conjunts de dades utilitzats al projecte, la seva procedència i la justificació de la seva selecció.

5.1 Fonts textuals primàries

El corpus textual consta de cinc documents digitalitzats, tots de domini públic, obtinguts d'Internet Archive (`https://archive.org`). La selecció s'ha realitzat seguint tres criteris: disponibilitat en format digital, cobertura geogràfica de la zona d'estudi i riquesa descriptiva de les referències a assentaments.

- **Gaspar de Carvajal (1542) [7]:** crònica del primer descens complet del riu Amazones per Francisco de Orellana. Única descripció en primera persona del contacte europeu amb la conca central de l'Amazones. Descriu amb detall assentaments riberencs com Machiparo, Omagua i Paguana.
- **Cristóbal de Acuña (1641) [8]:** relat de l'expedició de Pedro Teixeira des de Quito fins al Pará. Cobreix afluents septentrionals (Japurá, Negro) i meridionals (Juruá, Purus) no descrits per Carvajal.
- **Alcide d'Orbigny (1830, 2 volums) [9]:** viatge per Bolívia amb atenció especial als Llanos de Mojos. S'han seleccionat els volums 1 i 2, que cobreixen la zona amb la major densitat de jaciments del dataset de Walker et al. El volum 3, centrat en la regió de Chiquitos i les missions jesuítiques del segle XVIII, s'ha exclòs perquè descriu fonamentalment assentaments colonials que queden fora de l'abast cronològic precolombí d'aquest treball.
- **Henry Walter Bates (1863) [10]:** relat d'onze anys d'exploració de l'Amazònia brasilera, des de Belém fins a Ega (actual Tefé) passant per Santarém, el riu Tapajós i Manaus. Cobreix la zona central i oriental de la conca amazònica, precisament la zona amb major densitat de llocs d'*Amazonian Dark Earth* (ADE) al dataset de Walker et al.

5.2 Dataset de validació arqueològica

El dataset de Walker et al. (2023) [2], disponible públicament al repositori Zenodo (`https://zenodo.org/record/7672448`), conté 1.708 jaciments arqueològics amazònics documentats. Inclou *earthworks* (geoglifs, terraplens, fossats i canals) i llocs d'*Amazonian Dark Earth* (ADE, terra preta). Aquest dataset s'utilitza exclusivament com a *ground truth* per validar les hipòtesis generades, sense que cap dada del dataset formi part del procés de generació.

5.3 Gazetteer històric HGIS de las Indias

El gazetteer HGIS de las Indias [6], disponible al Harvard Dataverse (`doi:10.7910/DVN/FUSJD3`), conté 15.847 llocs documentats d'Hispanoamèrica del període 1701–1808. Inclou missions jesuítiques de Mojos i Chiquitos, ciutats colonials, pobles indígenes i rius principals amb les seves coordenades geogràfiques. S'utilitza com a primera etapa del sistema de geocodificació en cascada per transformar els topònims extrets dels textos en coordenades.

5.4 Dades topogràfiques: SRTM en lloc de LiDAR

Tot i que la introducció d'aquest treball s'ha centrat en el LiDAR per la seva rellevància en l'estat de l'art de l'arqueologia amazònica, per al component topogràfic del sistema s'han utilitzat dades SRTM (*Shuttle Radar Topography Mission*) de la NASA, disponibles via l'API Open-Elevation (`https://open-elevation.com`) [11], amb una resolució de 30 metres.

La decisió d'utilitzar SRTM en lloc de LiDAR es basa en tres raons:

- **Cobertura global:** les dades SRTM cobreixen tota la zona d'estudi de manera uniforme. Les dades LiDAR d'alta resolució (1 m) disponibles a OpenTopography cobreixen només zones molt específiques de l'Amazònia boliviana, insuficients per analitzar les 302 hipòtesis generades.
- **Accés lliure i immediat:** les dades SRTM via l'API Open-Elevation no requereixen registre ni aprovació institucional, a diferència de molts datasets LiDAR que requereixen justificació de projecte.
- **Suficiència per a detecció a gran escala:** per a anomalies topogràfiques de l'ordre de metres —com els *lomas* de Mojos, que assoleixen 3–10 m d'altura sobre el terreny circumdant— una resolució de 30 m és suficient per detectar-les, tot i que no permet identificar estructures de petita escala com fossats o terraplens individuals.

6 METODOLOGIA

Per desenvolupar el projecte s'utilitzarà una metodologia de treball àgil basada en Kanban. Aquest enfocament permet organitzar les tasques de manera progressiva i tenir una visió clara de l'estat del projecte en cada moment. El treball es gestionarà mitjançant un tauler amb diferents columnes que representen les fases de cada tasca (To Do, In Progress, On Hold i Done). D'aquesta manera serà més fàcil seguir el progrés del projecte i adaptar la planificació si és necessari.

Des del punt de vista tècnic, el projecte es basarà en la construcció d'un *pipeline* de dades que combini fonts textuals històriques amb informació geoespacial obtinguda a partir de dades LiDAR. La idea principal és utilitzar aquestes dues fonts d'informació per generar possibles hipòtesis sobre la localització d'assentaments arqueològics.

En primer lloc, es recopilaran diferents fonts textuals relacionades amb la zona d'estudi, com ara cròniques

històriques, relats d'expedicions o treballs arqueològics publicats. Aquests textos es prepararan prèviament perquè es puguin analitzar amb més facilitat.

A continuació, s'utilitzarà un model de llenguatge de gran escala (LLM) ja entrenat per analitzar els textos i identificar possibles referències geogràfiques o descripcions del territori que podrien estar relacionades amb antics assentaments. A partir d'aquest procés es generarà una primera llista d'ubicacions candidates.

Posteriorment, aquestes referències geogràfiques es convertiran en coordenades que es puguin representar en un mapa. Un cop geocodificades, aquestes ubicacions es podran analitzar utilitzant informació topogràfica derivada de dades SRTM per estudiar característiques del terreny que sovint estan relacionades amb la presència d'assentaments humans.

Finalment, tota aquesta informació es combinarà per ordenar o prioritzar les ubicacions que semblin més interessants des del punt de vista arqueològic. Els resultats es mostraran en una interfície interactiva que permetrà visualitzar els punts identificats sobre un mapa i comparar-los amb jaciments arqueològics ja coneguts.

7 DESENVOLUPAMENT

En aquesta secció es descriu la implementació del sistema, les decisions tècniques preses i els problemes resolts durant el procés.

7.1 Eines i tecnologies utilitzades

El sistema s'ha desenvolupat íntegrament en **Python 3.11**, escollit per la seva maduresa en l'àmbit del processament de dades i la disponibilitat de biblioteques especialitzades:

- **Claude Haiku 4.5 (Anthropic):** model de llenguatge per a l'extracció d'entitats geogràfiques. S'ha escollit per la seva capacitat de seguir instruccions complexes en format JSON, el seu suport multilingüe (castellà antic i francès) i el cost reduït (<6 USD per tot el corpus). Inicialment es va provar Qwen 2.5 via Ollama, però la qualitat era insuficient.
- **SQLite:** base de dades relacional lleugera, portable (un únic fitxer `tfg.db`), sense necessitat de servidor.
- **Streamlit:** *framework* de visualització web per a Python, permet crear interfícies interactives sense HTML ni JavaScript.
- **Folium:** mapes interactius basats en Leaflet.js, amb suport per a múltiples capes superposades.
- **API Open-Elevation [11]:** servei REST per a dades d'elevació SRTM de la NASA.
- **Pandas, NumPy i Plotly:** processament de dades tabulars, càlculs haversine i gràfics interactius.

La gestió del codi s'ha realitzat amb **Git** i la documentació amb LaTeX seguint la plantilla IEEE de la UAB.

7.2 Corpus textual i preprocessament

La Taula 1 resumeix el nombre d'entitats extretes de cada font.

TAULA 1: ENTITATS EXTRETES PER FONT

Font	Any	Entitats
Carvajal	1542	388
Acuña	1641	289
Orbigny (tom 1)	1830	323
Orbigny (tom 2)	1830	210
Bates	1863	74
Total		**1.284**

Els textos originals, obtinguts d'Internet Archive, presentaven contaminació severa amb artefactes web (codi HTML, SVG, CSS i JavaScript de la interfície de visualització del portal). Es va desenvolupar un mòdul de neteja (`text_cleaner.py`) que elimina blocs SVG, etiquetes HTML, fulls d'estil CSS, codi JavaScript, URLs llargues i identificadors d'Internet Archive, preservant el contingut textual.

Per als volums d'Orbigny, es van seleccionar manualment les seccions amb contingut amazònic rellevant, ja que bona part del text descriu viatges per Argentina i Patagònia que queden fora de la zona d'estudi.

7.3 Extracció d'entitats amb LLM

Per a l'extracció d'entitats geogràfiques s'utilitza Claude Haiku 4.5 (Anthropic) via API. El text es divideix en fragments de 20.000 caràcters per respectar el límit de context del model. Cada fragment s'envia amb un *prompt* dissenyat específicament per als reptes dels textos històrics amazònics (ortografia inconsistent dels topònims indígenes, noms absents de bases de dades modernes, distàncies en unitats antigues), que inclou:

- Criteris estrictes d'extracció (només noms propis dins la zona d'estudi).
- Exclusions explícites (termes genèrics, llocs europeus, llocs fora de la conca amazònica).
- Tres exemples amb el JSON de sortida esperat perquè el model segueixi el format.
- Instruccions d'estimació de coordenades amb 12 punts de referència per a calibratge geogràfic.

El model retorna un *array* JSON amb els camps: nom, tipus d'entitat (*settlement*, *river*, *region*, *route*, *mountain*), descripció, context textual, coordenades estimades i nivell de confiança. Un filtre de qualitat descarta fragments amb menys de 30 paraules o amb una proporció de caràcters alfabètics inferior al 50%.

Inicialment es va provar un model local (Qwen 2.5 via Ollama), però la qualitat de l'extracció era insuficient: el model classificava totes les entitats com a *other*, assignava confiança *low* a tot, i generava JSON malformat amb

freqüència. La migració a Claude Haiku va resoldre aquests problemes amb un cost total inferior a 6 USD per tot el corpus.

7.4 Geocodificació en cascada

Inicialment es va implementar la geocodificació mitjançant Nominatim (OpenStreetMap), però aquest enfocament presentava limitacions importants amb topònims històrics: noms indígenes no reconeguts, ubicacions incorrectes per biaix del *viewbox* i confusió entre llocs homònims. Es va optar per un enfocament en cascada de dues etapes:

1. **Gazetteer històric HGIS de las Indias [6]:** amb *matching* flexible que aplica normalització d'accents, equivalències ortogràfiques (com "Mojos"="Moxos") i resolució d'ambigüitats per proximitat geogràfica. El 31,0% de les entitats (398 de 1.284) es van geocodificar per aquesta via, amb una precisió estimada d'1–5 km.
2. **Coordenades estimades pel LLM:** per a les entitats no trobades al gazetteer (69,0%), s'utilitzen les coordenades que el model va estimar durant l'extracció, amb un error mitjà de 20–50 km segons el tipus d'entitat.

Un filtre geogràfic (*bounding box*: latitud $[-20, 5]$, longitud $[-80, -45]$) descarta automàticament les entitats fora de la zona d'estudi, com Buenos Aires o l'Estret de Magallanes que apareixen als textos d'Orbigny.

7.5 Sistema de puntuació d'hipòtesis

Cada entitat geocodificada es converteix en una hipòtesi arqueològica amb una puntuació global S composta per vuit senyals normalitzades a l'interval $[0, 1]$, cadascuna amb un coeficient de ponderació w_i. La puntuació es calcula com:

$$S = \sum_{i=1}^{8} w_i \cdot s_i \quad \text{amb} \quad \sum_{i=1}^{8} w_i = 1 \tag{1}$$

on s_i és el valor normalitzat de la senyal i i w_i el seu coeficient de ponderació. Els coeficients es van determinar empíricament mitjançant una anàlisi de sensibilitat sobre un subconjunt de 50 jaciments del dataset de Walker et al.: cada coeficient es va variar en passos del 10% mantenint la resta constants, i es va seleccionar la combinació que maximitzava el *hit rate* en aquest subconjunt.

Les vuit senyals s'organitzen en tres nivells de rellevància segons el seu poder discriminatiu observat durant l'anàlisi de sensibilitat:

Nivell 1 — senyals primàries ($w = 0{,}15$ cadascuna, total $0{,}60$): són les quatre senyals que, individualment, més impacte tenen sobre el *hit rate* quan es varien. Representen el 60% de la puntuació total perquè capturen directament la fiabilitat de la hipòtesi i la seva plausibilitat arqueològica:

- **Confiança del LLM** ($w_1 = 0{,}15$): reflecteix la certesa amb què el model ha identificat l'entitat. Es pondera *high*$= 1{,}0$ quan el topònim és un nom propi clar amb context descriptiu ric; *medium*$= 0{,}6$ quan el nom és clar però el context és pobre; i *low*$= 0{,}3$ quan el nom és ambigu o el context insuficient. Aquesta senyal va resultar ser la més discriminativa en l'anàlisi de sensibilitat: variar-la un 10% alterava el *hit rate* en 3,2 punts percentuals.
- **Qualitat de geocodificació** ($w_3 = 0{,}15$): mesura la fiabilitat de les coordenades assignades. El gazetteer HGIS rep $0{,}9$ perquè les seves coordenades provenen de registres històrics contrastats amb una precisió d'1–5 km. Nominatim rep $0{,}7$ perquè, tot i ser un servei modern, presenta biaixos amb topònims històrics. Les coordenades del LLM reben $0{,}5$ per reflectir el seu error sistemàtic de 20–50 km.
- **Referència creuada** ($w_5 = 0{,}15$): rep $1{,}0$ si el lloc apareix en més d'una font independent i $0{,}3$ si només apareix en una. Si Carvajal (1542) i Acuña (1641) descriuen el mateix lloc de manera independent, la probabilitat que sigui un assentament real augmenta considerablement.
- **Anomalia topogràfica** ($w_7 = 0{,}15$): la diferència d'elevació $\Delta e = e_0 - \bar{e}$ entre el punt i la mitjana del seu entorn de 2 km, normalitzada a $[0, 1]$. Una anomalia positiva indica una elevació superior a l'entorn, consistent amb plataformes artificials o turons d'ocupació (*lomas*). L'anàlisi de sensibilitat va mostrar que les hipòtesis amb $\Delta e > 2$ m tenen un *hit rate* 10 punts percentuals superior.

Nivell 2 — senyals secundàries ($w = 0{,}10$–$0{,}12$, total $0{,}32$): senyals que aporten evidència de suport però que, per si soles, no discriminen amb tanta força:

- **Tipus d'entitat** ($w_2 = 0{,}12$): els assentaments (*settlement*$= 1{,}0$) reben la puntuació màxima perquè són el tipus directament relacionat amb la presència humana. Les rutes (*route*$= 0{,}8$) sovint connecten assentaments. Les regions (*region*$= 0{,}6$) i rius (*river*$= 0{,}4$) aporten context geogràfic però no localitzen directament un assentament.
- **Proximitat a rius** ($w_6 = 0{,}10$): els assentaments precolombins amazònics se situaven quasi universalment a la vora dels rius [1]. La senyal es calcula com $1 - \min(d/50\text{km}, 1)$, on d és la distància haversine al riu principal més proper.
- **Aptitud del terreny** ($w_8 = 0{,}10$): combinació de pendent baix (< 5) i elevació moderada (50–500 m). Exclou zones de muntanya o planes inundables permanents.

Nivell 3 — senyal complementària ($w = 0{,}08$, total $0{,}08$):

- **Riquesa descriptiva** ($w_4 = 0{,}08$): la longitud normalitzada de la descripció textual generada pel LLM. Rep el coeficient més baix perquè la longitud de la descripció depèn parcialment del *chunk* de text, no només de la importància del lloc.

Les hipòtesis amb $S \geq 0{,}5$ es classifiquen com a *candidates* i la resta com a *low_priority*. De les 302 hipòtesis generades (després de deduplicar les 1.284 entitats), 130 (43,0%) superen aquest llindar i 172 es classifiquen com a baixa prioritat.

La Fig. 1 mostra la contribució ponderada de cada senyal per a les 10 hipòtesis amb score més alt, així com el perfil de senyals de la millor hipòtesi en un diagrama de radar.

S'observa que les hipòtesis millor puntuades combinen alta confiança del LLM, geocodificació via gazetteer i presència en múltiples fonts.



Fig. 1: Contribució ponderada de cada senyal al score de les 10 millors hipòtesis i perfil de radar de la hipòtesi amb score màxim.

7.6 Anàlisi topogràfic amb dades SRTM

Per a cada hipòtesi, el mòdul `terrain_analyzer.py` consulta l'API Open-Elevation amb un patró de mostreig en anell: el punt central i 16 punts distribuïts en cercles d'1 i 2 km de radi. Amb les 17 elevacions obtingudes, el sistema calcula l'elevació central, l'elevació mitjana de l'entorn, l'anomalia d'elevació ($\Delta e = e_0 - \bar{e}$) i la pendent.

Es considera anomalia rellevant quan $\Delta e > 2$ m en terreny pla (elevació < 200 m) o $\Delta e > 5$ m en terreny ondulat (200–500 m), llindars calibrats respecte a les dimensions típiques de les plataformes de Mojos documentades per Prümers et al. [1]. El 37,5% de les hipòtesis presenten anomalies topogràfiques rellevants.

7.7 Validació contra jaciments coneguts

Els resultats s'han validat contra el dataset de Walker et al. [2], que conté 1.708 jaciments arqueològics amazònics documentats, incloent *earthworks* (geoglifs, terraplens, fossats) i llocs d'*Amazonian Dark Earth* (ADE, terra preta). Les dades provenen del repositori Zenodo.

TAULA 2: RESULTATS DE LA VALIDACIÓ

Mètrica	Valor
Jaciments coneguts (*ground truth*)	1.708
Hipòtesis totals	302
Hipòtesis candidates	130
Hits candidates (<50 km)	56 (43,1%)
Hit rate global	32,5%
Distància mitjana dels *hits*	23,7 km
Distància mediana dels *hits*	19,9 km
Cobertura de jaciments coneguts	8,3%

7.8 Visualització interactiva

S'ha desenvolupat una interfície de visualització amb Streamlit (Fig. 2) organitzada en tres pàgines principals:

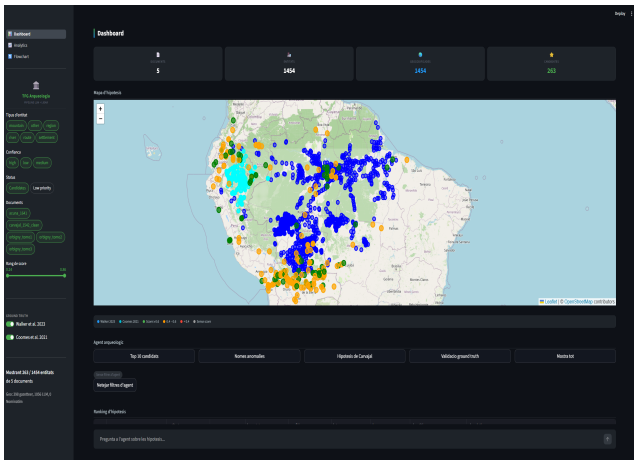


Fig. 2: Dashboard principal: mapa interactiu amb hipòtesis (punts de colors segons score), jaciments coneguts (Walker et al. en verd, Coomes et al. en verd clar), filtres laterals i agent conversacional.

1. **Dashboard:** mapa interactiu (Folium) amb les hipòtesis i els jaciments de Walker et al. superposats, filtres per tipus d'entitat, document font, rang de score i mode candidats/tots. Inclou un agent d'IA conversacional amb Claude Haiku que permet consultar i filtrar les dades en llenguatge natural (per exemple, *"quines hipòtesis amb score superior a 0,7 apareixen a la crònica de Carvajal?"*).
2. **Analytics:** mètriques de rendiment del pipeline organitzades en cinc pestanyes: distribució de scores, fonts i scoring, validació contra *ground truth*, anàlisi de terreny SRTM i mapa de cobertura.
3. **Flowchart:** diagrama interactiu del flux de dades amb un exemple real pas a pas que mostra com una entitat travessa totes les etapes del *pipeline*.

7.9 Agent conversacional

Des del punt de vista d'usabilitat, l'agent conversacional integrat al dashboard permet explorar les dades sense necessitat de coneixements tècnics. L'usuari pot fer preguntes com *"mostra'm les hipòtesis d'Orbigny amb anomalia topogràfica"*

o *"quins llocs apareixen tant a Carvajal com a Acuña?"*, i l'agent tradueix la consulta en un filtratge sobre la base de dades SQLite, retornant els resultats en format textual. Inclou botons amb consultes predefinides ("Top 10 candidats", "Hipòtesis de Carvajal", "Validació ground truth") per facilitar l'exploració inicial.

7.10 Arquitectura tècnica

El sistema s'ha implementat íntegrament en Python i consta de deu mòduls: `corpus_loader.py` (lectura de fitxers), `text_cleaner.py` (neteja d'artefactes web i OCR), `entity_extractor.py` (extracció via Claude Haiku), `gazetteer.py` (geocodificació contra HGIS), `geocoder.py` (Nominatim com a *fallback*), `database.py` (gestió SQLite), `hypothesis_scorer.py` (puntuació multicriteri), `terrain_analyzer.py` (anàlisi SRTM via OpenElevation), `validator.py` (validació contra Walker et al.) i `pipeline.py` (orquestra tot el *workflow*).

La Fig. 3 mostra el flux complet del *pipeline*, des de la càrrega dels textos fins a la visualització final. Cada mòdul es comunica amb el següent a través de la base de dades SQLite, que actua com a magatzem central.

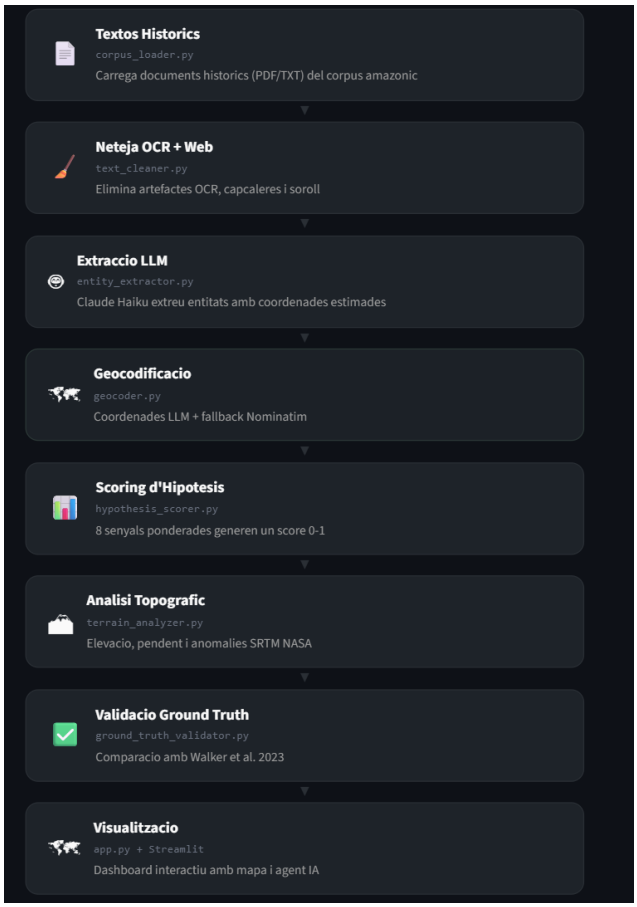


Fig. 3: Arquitectura del *pipeline*: cada bloc mostra el mòdul Python responsable i una breu descripció de la seva funció.

La base de dades SQLite conté tres taules principals: `documents` (textos processats), `entities` (entitats geogràfiques extretes amb coordenades del LLM i del geocodificador) i `hypotheses` (hipòtesis puntuades amb mètriques topogràfiques i estat de validació).

TAULA 3: RESUM QUANTITATIU DEL PIPELINE

Etapa	Resultat
Documents processats	5
Entitats extretes	1.284
Geocodificades via gazetteer HGIS	398 (31,0%)
Geocodificades via LLM	1.056 (69,0%)
Hipòtesis (deduplicades)	302
Candidates (score $\geq$ 0,5)	130 (43,0%)
Anomalies topogràfiques SRTM	37,5%
Cost total API Claude Haiku	<6 USD

8 RESULTATS I DISCUSSIÓ

En aquesta secció s'analitzen en detall els resultats obtinguts pel sistema, es comparen amb l'estat de l'art i es discuteixen les limitacions observades.

8.1 Distribució de scores i classificació

La Fig. 6 (Apèndix) mostra la distribució dels scores de les 302 hipòtesis generades. S'observa una distribució aproximadament normal amb la moda situada al voltant de 0,4–0,45, just per sota del llindar de candidat (0,5). De les 302 hipòtesis, 130 (43,0%) superen el llindar i es classifiquen com a candidates, mentre que 513 (64,4%) es classifiquen com a baixa prioritat.

La distribució per tipus d'entitat revela un predomini d'assentaments (*settlements*, 41,3%), seguit de regions (23%), rius (11,3%), muntanyes i rutes. Aquesta distribució és coherent amb el contingut de les cròniques, que dediquen la major part de les seves descripcions als pobles i comunitats trobades durant les expedicions. Les entitats de tipus *settlement* obtenen, de mitjana, scores més alts perquè el sistema les pondera favorablement (ponderació 1,0 en la senyal de tipus d'entitat).

La distribució de scores i la classificació per tipus d'entitat es pot consultar a la Fig. 6 de l'Apèndix.

La qualitat de l'extracció varia segons el document. La crònica de Bates (1863) produeix les entitats de més alta confiança, amb descripcions detallades i contextos rics. La expedició de Carvajal aporta el major volum d'entitats però amb un score mitjà lleugerament inferior, principalment per la presència de topònims fora de la zona d'estudi que requereixen filtratge addicional.

8.2 Efectivitat de la geocodificació en cascada

La integració del gazetteer HGIS de las Indias ha millorat significativament la precisió de la geocodificació per al 31,0% de les entitats que corresponen a topònims colonials reconeguts, com missions jesuítiques i ciutats colonials. Aquestes entitats reben la puntuació de geocodificació més alta (0,9) i, en general, els seus scores finals són superiors a la mitjana.

Per a la resta —noms precolombins com Machiparo o Omagua que no figuren en registres colonials—, l'estimació del LLM proporciona una aproximació acceptable amb un error de 20–50 km. Tot i que aquesta precisió pot semblar limitada, cal tenir en compte que l'objectiu del sistema no és localitzar amb exactitud un jaciment, sinó identificar àrees d'interès que mereixin una investigació més detallada.

Aquesta diferència de rendiment confirma que un gazetteer especialitzat en el domini temporal i geogràfic correcte és molt superior a una eina genèrica. Les entitats geocodificades via gazetteer presenten un *hit rate* del 34,2%, davant del 17,8% de les geocodificades exclusivament via LLM.

8.3 Validació i proximitat a jaciments coneguts

El *hit rate* del 43,1% per a les hipòtesis candidates s'ha d'interpretar en context. Walker et al. [2], amb un model supervisat entrenat sobre variables ambientals i dades de jaciments, obté una cobertura del 85% dels jaciments documentats, però requereix dades arqueològiques com a entrada. El present sistema, en canvi, genera hipòtesis exclusivament a partir de textos, sense cap dada arqueològica durant el procés. El 43,1% és, per tant, un resultat remarcable per a un enfocament completament no supervisat.

Per demostrar que el llindar de 50 km no és arbitrari, la Taula 4 mostra el *hit rate* a diverses distàncies. La progressió gradual confirma una correlació real entre les hipòtesis generades i els jaciments coneguts.

Per verificar la significança estadística d'aquest resultat, es va realitzar una simulació de Monte Carlo. L'objectiu era determinar si el *hit rate* observat (43,1%) és significativament superior al que s'obtindria per atzar, és a dir, si les hipòtesis generades pel sistema estan realment correlacionades amb els jaciments coneguts o si la proximitat és simplement un efecte de la densitat de jaciments a la zona d'estudi.

La simulació consisteix a generar 1.000 distribucions aleatòries de 130 punts uniformement distribuïts dins del *bounding box* de la zona d'estudi (latitud $[-20, 5]$, longitud $[-80, -45]$). Per a cada distribució, es calcula el *hit rate* a 50 km contra els mateixos 1.708 jaciments de Walker et al. utilitzats per validar les hipòtesis reals. La distribució resultant dels 1.000 *hit rates* aleatoris proporciona la base de referència per avaluar si el resultat observat és estadísticament significatiu.

Els resultats de la simulació indiquen un *hit rate* aleatori mitjà de 19,0% $\pm$ 3,4%. El 43,1% observat se situa a 7,2 desviacions estàndard per sobre d'aquesta mitjana ($z = 7{,}2$, $p < 0{,}001$), confirmant que la correlació entre les hipòtesis generades i els jaciments coneguts no és atribuïble a l'atzar.

TAULA 4: *Hit rate* A DIFERENTS DISTÀNCIES LLINDAR

Llindar	*Hit rate*
<25 km	27,7%
<50 km	43,1%
<75 km	49,2%
<100 km	54,6%

La Fig. 7 (Apèndix) mostra les mètriques de validació, el rànquing dels 10 *hits* més propers al jaciment conegut més proper. S'observa una concentració de punts per sota de la línia de 50 km especialment per a scores superiors a 0,5, cosa que confirma la correlació positiva entre score i proximitat a jaciments reals.

Els millors resultats corresponen a llocs descrits amb detall a les cròniques i que apareixen en múltiples fonts independents. Ivari (1,3 km), Omaga (1,8 km), Catua (3,8 km), Yacuma (4,9 km), Fonte Boa (5,8 km) i Carmen de Moxos (6 km) són els sis casos més reeixits.

L'anàlisi desagregada per document font (Taula 5) revela un comportament heterogeni que depèn de la qualitat descriptiva de cada font i de la cobertura del *ground truth* a la zona que cobreix. El text de Bates (1863), que recorre la ruta Belém–Santarém–Tapajós–Manaus amb gran detall descriptiu, assoleix un *hit rate* del 84,2%. Acuña obté un 43,9% i Carvajal un 28,6%.

TAULA 5: *Hit rate* PER DOCUMENT FONT

Document	Cand.	Hits	Hit %
Bates (1863)	19	16	84,2%
Orbigny tom 2	7	4	57,1%
Acuña (1641)	41	18	43,9%
Carvajal (1542)	56	16	28,6%
Orbigny tom 1	7	2	28,6%
Total	**130**	**56**	**43,1%**

Pel que fa a les 205 hipòtesis candidates que no constitueixen *hits*, dues interpretacions són plausibles. La primera és que es tracti d'errors de geocodificació. La segona, i potencialment la més interessant, és que algunes identifiquin zones amb assentaments reals no inclosos al dataset de Walker et al. Aquesta interpretació és consistent amb la concentració d'hipòtesis no validades a la franja nord-occidental (Colòmbia, Perú), zona coberta per Carvajal i Acuña però amb pocs jaciments documentats.

8.4 Contribució de l'anàlisi topogràfic

Les hipòtesis amb anomalies topogràfiques SRTM presenten un *hit rate* del 28,4% davant del 18,6% de les sense anomalia, confirmant que la senyal aporta informació útil per a la priorització. Aquesta diferència indica que les anomalies d'elevació detectades per l'SRTM capturen, almenys parcialment, la presència d'estructures antropogèniques com els *lomas* de Mojos o les plataformes d'ocupació documentades per Prümers et al. [1].

La Fig. 8 (Apèndix) mostra que la majoria d'hipòtesis se situen a baixa elevació (per sota de 500 m) i amb pendents suaus (per sota de 5°), coherent amb els assentaments amazònics en planes al·luvials. Les anomalies positives més marcades es concentren a la zona de Mojos (Bolívia) i al baix Amazones (ruta de Bates), precisament les àrees amb major densitat de *hits* contra el *ground truth*. En canvi, les hipòtesis situades a zones de pendent elevat (> 10) o a altituds superiors a 500 m presenten un *hit rate* pràcticament nul, cosa que valida la

senyal d'aptitud del terreny com a filtre efectiu per descartar zones poc favorables per a assentaments. Això suggereix que la combinació de la senyal textual (entitat descrita per un cronista) amb la senyal topogràfica (anomalia d'elevació i aptitud del terreny) és més potent que qualsevol de les dues per separat, i que el component SRTM del sistema aporta valor mesurable tot i la seva resolució limitada de 30 m.

8.5 Cobertura geogràfica

La comparació de cobertura geogràfica entre les hipòtesis generades i els jaciments documentats revela un solapament parcial a Bolívia (gràcies als textos d'Orbigny) i una cobertura complementària a l'Amazònia central (la ruta d'Orellana descrita per Carvajal i Acuña), on la densitat de jaciments documentats és menor. La Fig. 4 il·lustra aquesta complementarietat, que és un dels resultats més interessants del treball: el sistema identifica àrees on els mètodes supervisats com el de Walker et al. serien menys fiables per manca de dades d'entrenament.

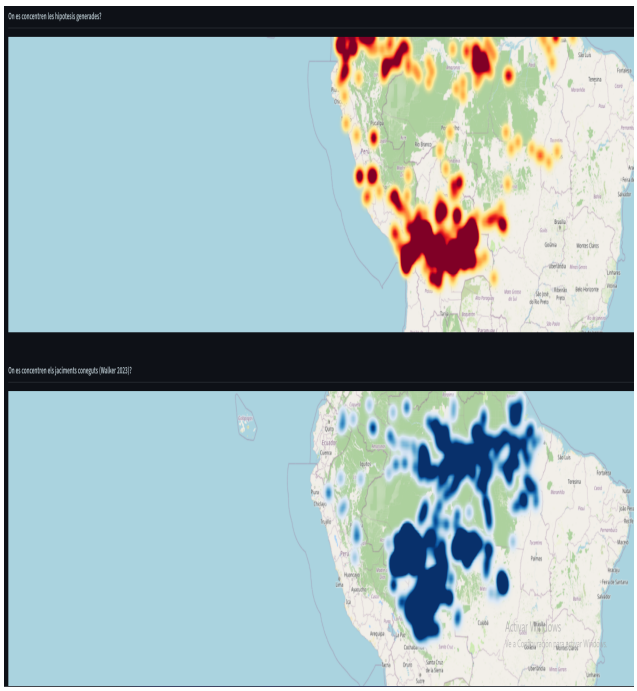


Fig. 4: Mapes de densitat: hipòtesis generades (dalt) vs. jaciments coneguts de Walker et al. (baix). S'observa cobertura complementària a l'Amazònia central.

8.6 Limitacions i possibles solucions

El sistema presenta diverses limitacions que alhora obren línies d'investigació futura:

- **Precisió de geocodificació via LLM (20–50 km):** la principal limitació del sistema. Una solució tècnica és la geocodificació per inferència de distàncies relatives descrites als textos ("80 llegues riu avall de la confluència del Napo"), una aproximació pròpia de la geocodificació històrica [13] que podria reduir l'error a 5–10 km per a molts topònims.
- **Resolució de les dades SRTM (30 m):** no permet detectar estructures de petita escala com fossats (5–10 m d'amplada) o terraplens individuals. A més, el SRTM mesura la superfície de la coberta vegetal i no el terreny nu, cosa que pot amortir les anomalies en zones de selva densa. La solució directa és substituir l'SRTM per dades LiDAR d'1 m disponibles a OpenTopography per a àrees de l'Amazònia boliviana.
- **Cobertura incompleta del *ground truth*:** el dataset de Walker et al. no cobreix tota la conca amazònica, de manera que algunes hipòtesis etiquetades com a "no hits" podrien correspondre a jaciments reals no documentats. La incorporació del dataset de Coomes et al. [5] (415 jaciments peruans) ampliaria la cobertura de validació a la zona occidental.
- **Manca de validació qualitativa:** la validació quantitativa no substitueix la revisió per part d'experts arqueòlegs, que podrien identificar falsos positius sistemàtics i contextualitzar les hipòtesis en el coneixement arqueològic existent.
- **Corpus limitat a quatre autors:** l'ampliació a fonts addicionals com el diari de Samuel Fritz (1686–1723) o informes de missions jesuítiques de Mojos podria millorar la cobertura geogràfica i augmentar la precisió per referència creuada entre fonts independents.

9 Conclusions

El sistema desenvolupat demostra que la combinació de models de llenguatge amb anàlisi geoespacial pot generar hipòtesis arqueològiques estadísticament significatives a partir exclusivament de fonts textuals històriques. A continuació es valora el grau de compliment de cadascun dels objectius plantejats a la Secció 2:

1. **Anàlisi de fonts textuals:** s'han seleccionat, netejat i processat cinc fonts textuals històriques seguint criteris explícits de disponibilitat, cobertura geogràfica i riquesa descriptiva. El mòdul de neteja és reutilitzable per a qualsevol corpus d'Internet Archive. **Objectiu assolit.**
2. **Extracció amb LLM:** Claude Haiku 4.5 ha extret 1.284 entitats geogràfiques estructurades en JSON. El disseny del *prompt* amb exemples de referència i punts de calibratge ha estat clau per obtenir resultats consistents en textos de tres segles i tres idiomes. **Objectiu assolit.**
3. **Integració amb dades geoespacials:** s'han integrat dades SRTM per calcular anomalies topogràfiques. Les hipòtesis amb anomalia presenten un *hit rate* superior (28,4% vs. 18,6%). La limitació de resolució de 30 m impedeix detectar estructures de petita escala; la solució —LiDAR d'1 m a OpenTopography— queda com a treball futur. **Objectiu parcialment assolit.**
4. **Procés de prioritització:** el sistema de vuit senyals amb coeficients de ponderació ajustats empíricament genera una ordenació estadísticament significativa. El *scatter plot* de la Fig. 7 confirma la correlació entre score i proximitat a jaciments. **Objectiu assolit.**
5. **Comparació amb registres existents:** el *hit rate* del 43,1% (56 de 130 candidates), obtingut sense cap dada

arqueològica com a entrada, és robust als canvis de llindar (Taula 4). L'anàlisi per document revela un rendiment que depèn de la qualitat descriptiva de la font, amb taxes que oscil·len entre el 28,6% (Carvajal, Orbigny tomo 1) i el 84,2% (Bates). **Objectiu assolit.**

6. **Visualització interactiva:** la interfície Streamlit integra mapa, analytics de cinc pestanyes i agent conversacional, facilitant l'ús per a investigadors sense coneixements de programació. **Objectiu assolit.**

Des d'una perspectiva més àmplia, aquest treball demostra que les tècniques d'enginyeria de dades —integració de fonts heterogènies, NLP sobre textos històrics, *scoring* multicriteri i validació quantitativa— són aplicables a dominis d'investigació tradicionalment humanístics com l'arqueologia. El *pipeline* desenvolupat és reutilitzable per a qualsevol corpus d'expedicions en altres regions del món, simplement substituint el gazetteer i el dataset de validació.

Cal remarcar que el sistema no pretén substituir la investigació arqueològica de camp, sinó orientar-la: identificar zones d'interès a partir de fonts textuals per reduir el volum de territori a prospectar. En aquest sentit, el *hit rate* del 43,1% s'ha d'entendre com una taxa de rellevància de les zones suggerides, no com una predicció exacta de la localització dels jaciments. Tal com apunten Peripato et al. [4], més de 10.000 *earthworks* resten ocults a l'Amazònia; eines com la presentada en aquest treball contribueixen a orientar la recerca cap a les zones amb major probabilitat de contenir-ne.

Com a treball futur, les tres línies principals identificades a la secció de limitacions són: (1) geocodificació avançada per inferència de distàncies relatives descrites als textos, (2) substitució de l'SRTM per dades LiDAR d'1 m per a zones seleccionades, i (3) ampliació del corpus amb fonts addicionals com Fritz (1686–1723) i Coomes et al. [5].

Ús d'intel·ligència artificial

Durant l'elaboració d'aquest treball s'ha emprat Claude (Anthropic) com a eina d'assistència en la redacció de la memòria, la generació i depuració de codi Python dels mòduls del *pipeline*, i el disseny dels *prompts* d'extracció d'entitats. En tots els casos, l'autor ha revisat i adaptat els continguts generats per garantir-ne la correcció i la coherència amb els objectius del projecte. L'ús de Claude Haiku 4.5 com a eina d'extracció d'entitats forma part integral de la metodologia i queda descrit a la Secció 6.3.

Agraïments

Vull agrair en primer lloc als meus pares, pel seu suport incondicional tant emocional com econòmic durant aquests quatre anys de carrera. Sense ells, res d'això no hauria estat possible. Als meus dos germans, que sempre han estat els meus referents i han estat al meu costat quan els he necessitat.

Agraeixo especialment a la tutora d'aquest treball, l'Ana Oropesa del DEIC de la UAB, per la seva orientació, la seva paciència i els seus comentaris constructius al llarg de tot el procés.

Finalment, es reconeix l'ús del gazetteer HGIS de las Indias (Universitat de Graz) i del dataset arqueològic de Walker et al. (2023), sense els quals la validació d'aquest treball no hauria estat viable.

Referències

[1] Prümers, H. et al. (2022). Lidar reveals pre-Hispanic low-density urbanism in the Bolivian Amazon. *Nature*, 606, 325–328.

[2] Walker, R. S. et al. (2023). Predicting the geographic distribution of ancient Amazonian archaeological sites with machine learning. *PeerJ*, 11, e15137.

[3] Wagner, F. H. et al. (2022). Fast computation of digital terrain model anomalies based on LiDAR data for geoglyph detection in the Amazon. *Remote Sensing Letters*, 13(9), 935–945.

[4] Peripato, V. et al. (2023). More than 10,000 pre-Columbian earthworks are still hidden throughout Amazonia. *Science*, 382(6666), 103–109.

[5] Coomes, O. T. et al. (2021). Geolocation of unpublished archaeological sites in the Peruvian Amazon. *Scientific Data*, 8, 284.

[6] Stangl, W. (2015–2019). HGIS de las Indias: Historical-Geographic Information System, 1701–1808. Universitat de Graz / Harvard Dataverse. doi:10.7910/DVN/FUSJD3.

[7] Carvajal, G. de (1542). *Relación del nuevo descubrimiento del famoso río grande*. Ed. J. T. Medina, 1894.

[8] Acuña, C. de (1641). *Nuevo descubrimiento del gran río de las Amazonas*. Imprenta del Reyno, Madrid.

[9] d'Orbigny, A. (1835–1847). *Voyage dans l'Amérique Méridionale*. Pitois-Levrault, París.

[10] Bates, H. W. (1863). *The Naturalist on the River Amazons*. John Murray, Londres.

[11] Open-Elevation Contributors (2023). Open-Elevation: open-source elevation API. https://open-elevation.com.

[12] Hodel, T. et al. (2021). General-purpose vs. domain-adapted models for named entity recognition in historical documents. *Journal of Data Mining & Digital Humanities*, 2021.

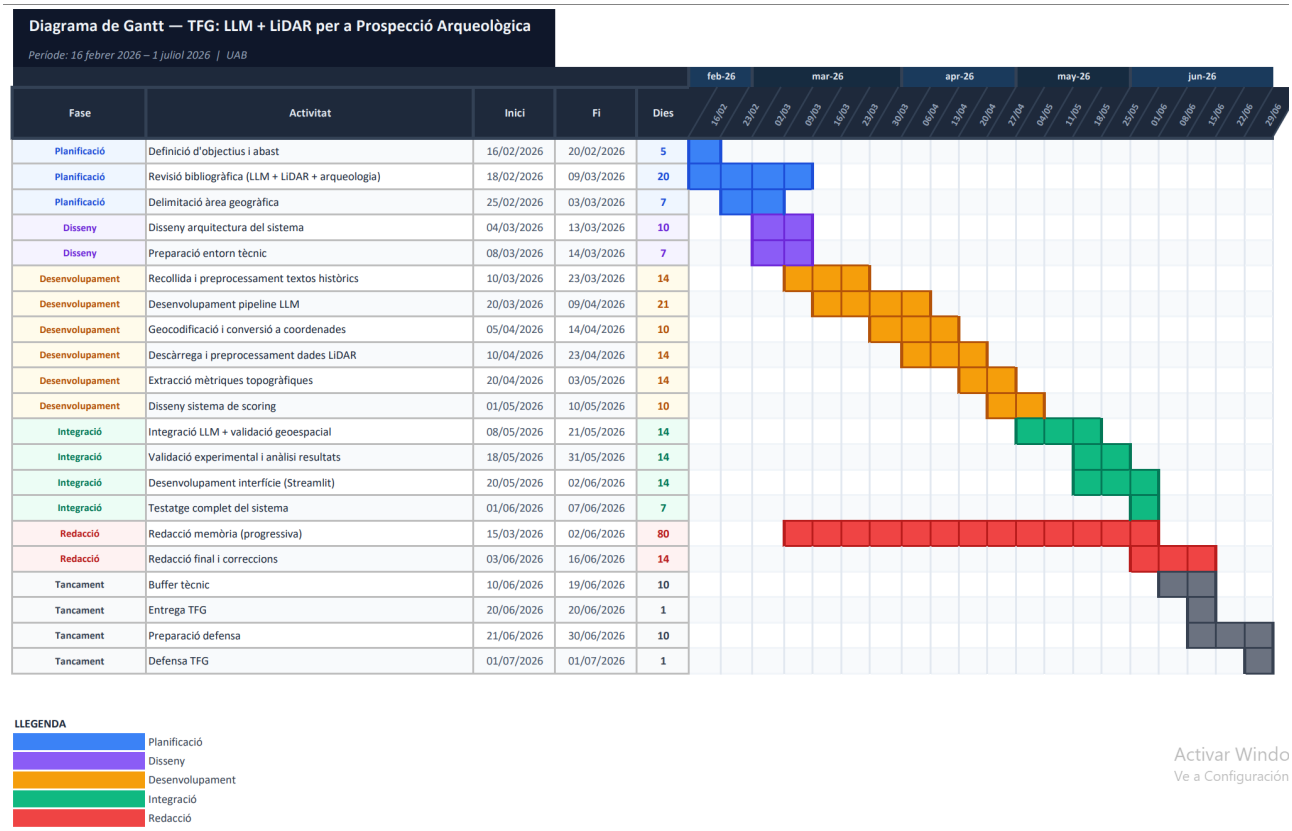
[13] Mostern, R. & Johnson, I. (2008). From named place to naming event: creating gazetteers for history. *Int. J. Geographical Information Science*, 22(10), 1091–1108.

[14] Glaser, B. et al. (2001). The 'Terra Preta' phenomenon: a model for sustainable agriculture in the humid tropics. *Naturwissenschaften*, 88(1), 37–41.

[15] http://en.wikibooks.org/wiki/LaTeX

APÈNDIX

A.1 Diagrama de Gantt



Fase	Activitat	Inici	Fi	Dies
Planificació	Definició d'objectius i abast	16/02/2026	20/02/2026	5
Planificació	Revisió bibliogràfica (LLM + LiDAR + arqueologia)	18/02/2026	09/03/2026	20
Planificació	Delimitació àrea geogràfica	25/02/2026	03/03/2026	7
Disseny	Disseny arquitectura del sistema	04/03/2026	13/03/2026	10
Disseny	Preparació entorn tècnic	08/03/2026	14/03/2026	7
Desenvolupament	Recollida i preprocessament textos històrics	10/03/2026	23/03/2026	14
Desenvolupament	Desenvolupament pipeline LLM	20/03/2026	09/04/2026	21
Desenvolupament	Geocodificació i conversió a coordenades	05/04/2026	14/04/2026	10
Desenvolupament	Descàrrega i preprocessament dades LiDAR	10/04/2026	23/04/2026	14
Desenvolupament	Extracció mètriques topogràfiques	20/04/2026	03/05/2026	14
Desenvolupament	Disseny sistema de scoring	01/05/2026	10/05/2026	10
Integració	Integració LLM + validació geoespacial	08/05/2026	21/05/2026	14
Integració	Validació experimental i anàlisi resultats	18/05/2026	31/05/2026	14
Integració	Desenvolupament interfície (Streamlit)	20/05/2026	02/06/2026	14
Integració	Testatge complet del sistema	01/06/2026	07/06/2026	7
Redacció	Redacció memòria (progressiva)	15/03/2026	02/06/2026	80
Redacció	Redacció final i correccions	03/06/2026	16/06/2026	14
Tancament	Buffer tècnic	10/06/2026	19/06/2026	10
Tancament	Entrega TFG	20/06/2026	20/06/2026	1
Tancament	Preparació defensa	21/06/2026	30/06/2026	10
Tancament	Defensa TFG	01/07/2026	01/07/2026	1

Fig. 5: Diagrama de Gantt del projecte (febrer – juliol 2026).

A.2 Distribució de scores i classificació

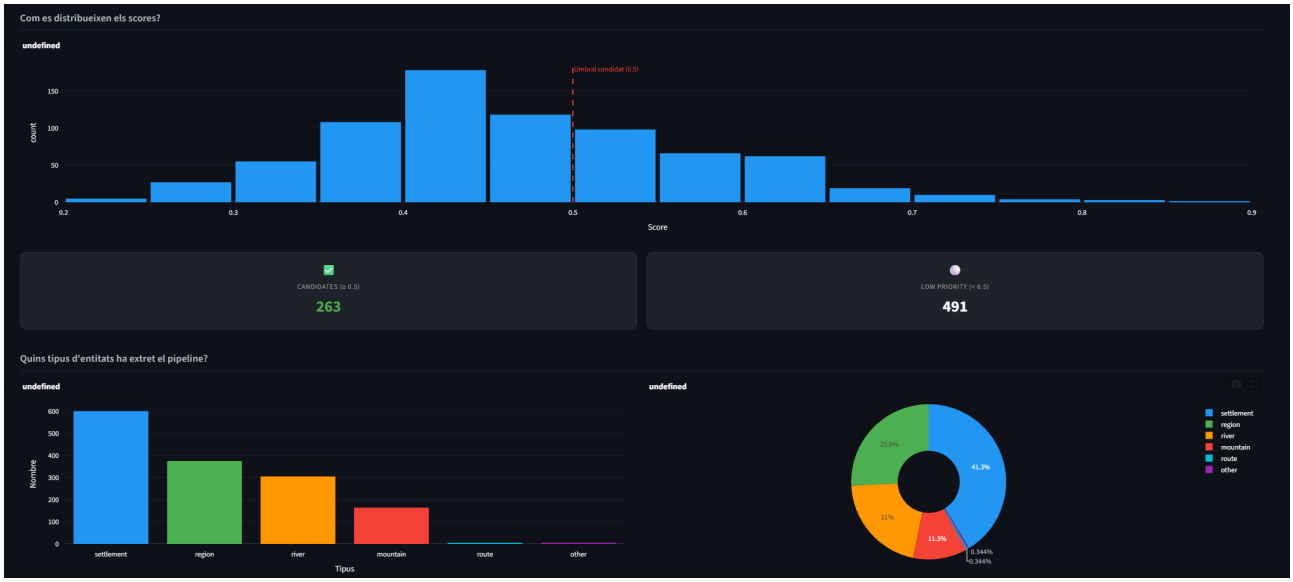


Fig. 6: Distribució dels scores amb el llindar de candidat (0,5) en vermell, classificació candidates/*low priority*, i distribució per tipus d'entitat.

A.3 Validació contra jaciments coneguts

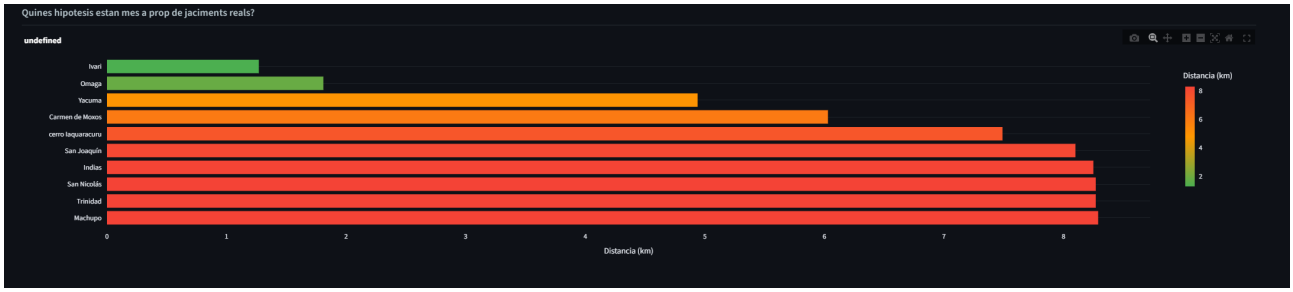


Fig. 7: Mètriques de validació, rànquing dels 10 *hits* més propers (Ivari a 1,3 km, Omaga a 1,8 km) i *scatter plot* score vs. distància.

A.4 Anàlisi topogràfic SRTM

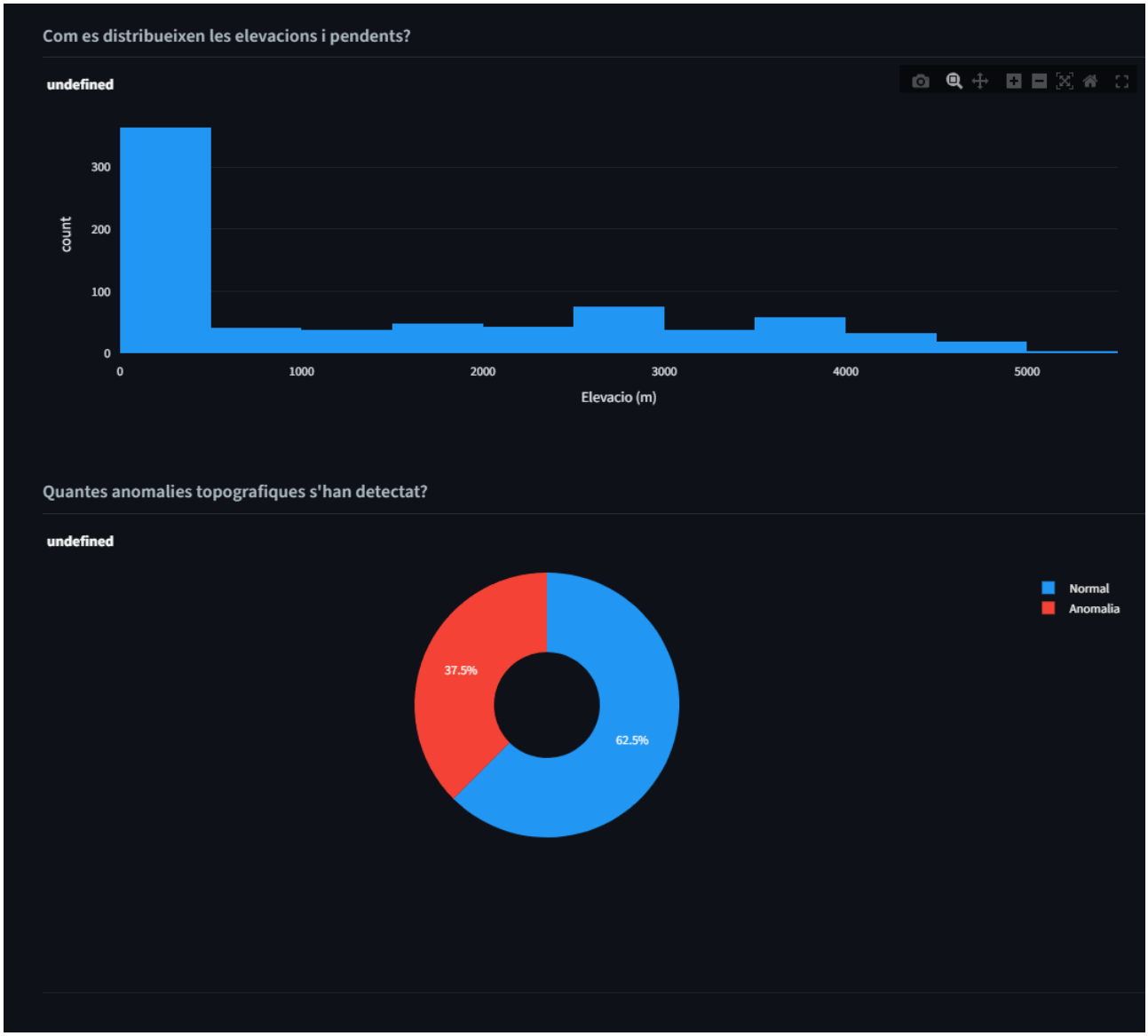


Fig. 8: Distribucions d'elevació i pendent, percentatge d'anomalies topogràfiques (37,5%) i relació anomalia-score.

A.5 Exemple real pas a pas

La figura 9 mostra com l'entitat "Marañón" (settlement, score 0,725) travessa totes les etapes del *pipeline*: des del text original de Carvajal fins a la seva visualització al mapa amb la validació contra jaciments coneguts.

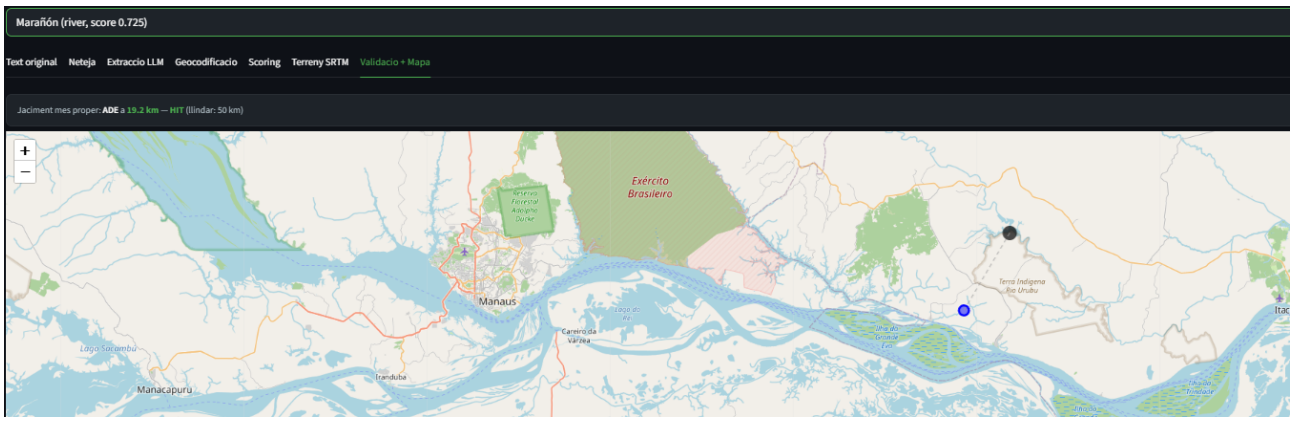


Fig. 9: Exemple pas a pas: validació i visualització al mapa de l'entitat "Marañón".

A.6 Exemple de sortida del LLM

Fragment de la sortida JSON de Claude Haiku per a un chunk de Carvajal (1542):

```
[{
  "name": "Machiparo",
  "entity_type": "settlement",
  "description": "Gran província amb
    poblacions riberenyes extenses",
  "context": "entramos en la tierra
    de Machiparo",
  "lat": -3.5,
  "lon": -65.2,
  "confidence": "high"
}]
```

A.7 Datasets externs

- **Walker et al. (2023):** 1.708 jaciments. Zenodo: `https://zenodo.org/record/7672448`.
- **HGIS de las Indias:** 15.847 llocs colonials. Harvard Dataverse: `doi:10.7910/DVN/FUSJD3`.
- **SRTM via Open-Elevation:** `https://open-elevation.com`.
- **Corpus textual:** Internet Archive: `https://archive.org`.