

Sobreviviendo la era de los LLMs

Raúl Ramos Pollán

Prof. Universidad de Antioquia

AI Lead at FDL NASA/ESA



ccc Claude C Compiler

AI

Research

Economic Futures

Commitments ▾

Learn ▾

News

Try Claude



Engineering at Anthropic



Building a C compiler with a team of parallel Claudes

Published Feb 05, 2026

We tasked Opus 4.6 using agent teams to build a C Compiler, and then (mostly) walked away. Here's what it taught us about the future of autonomous software development.

Breakthrough 1: Large Language Models

- I studied French during a year and now
- I like the Titanic movie but

history (h):

Alice painted her house ?

next element:

$$P_{\theta}(? = \text{brown} | h) = 0.2$$

$$P_{\theta}(? = \text{beige} | h) = 0.1$$

$$P_{\theta}(? = \text{red} | h) = 0.05$$

$$P_{\theta}(? = \text{because} | h) = 0.09$$

$$P_{\theta}(? = \text{with} | h) = 0.08$$

Generation samples:

- Alice painted her house → brown → and → then → took → a → rest
- Alice painter her house → because → she → was → bored

Transformers - Self attention

Attention Is All You Need

Ashish Vaswani*
Google Brain
avaswani@google.com

Noam Shazeer*
Google Brain
noam@google.com

Niki Parmar*
Google Research
nikip@google.com

Jakob Uszkoreit*
Google Research
usz@google.com

Llion Jones*
Google Research
llion@google.com

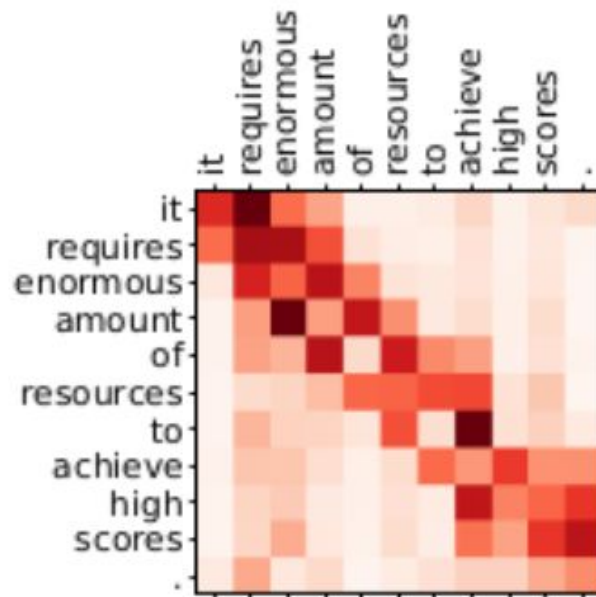
Aidan N. Gomez* †
University of Toronto
aidan@cs.toronto.edu

Lukasz Kaiser*
Google Brain
lukaszkaier@google.com

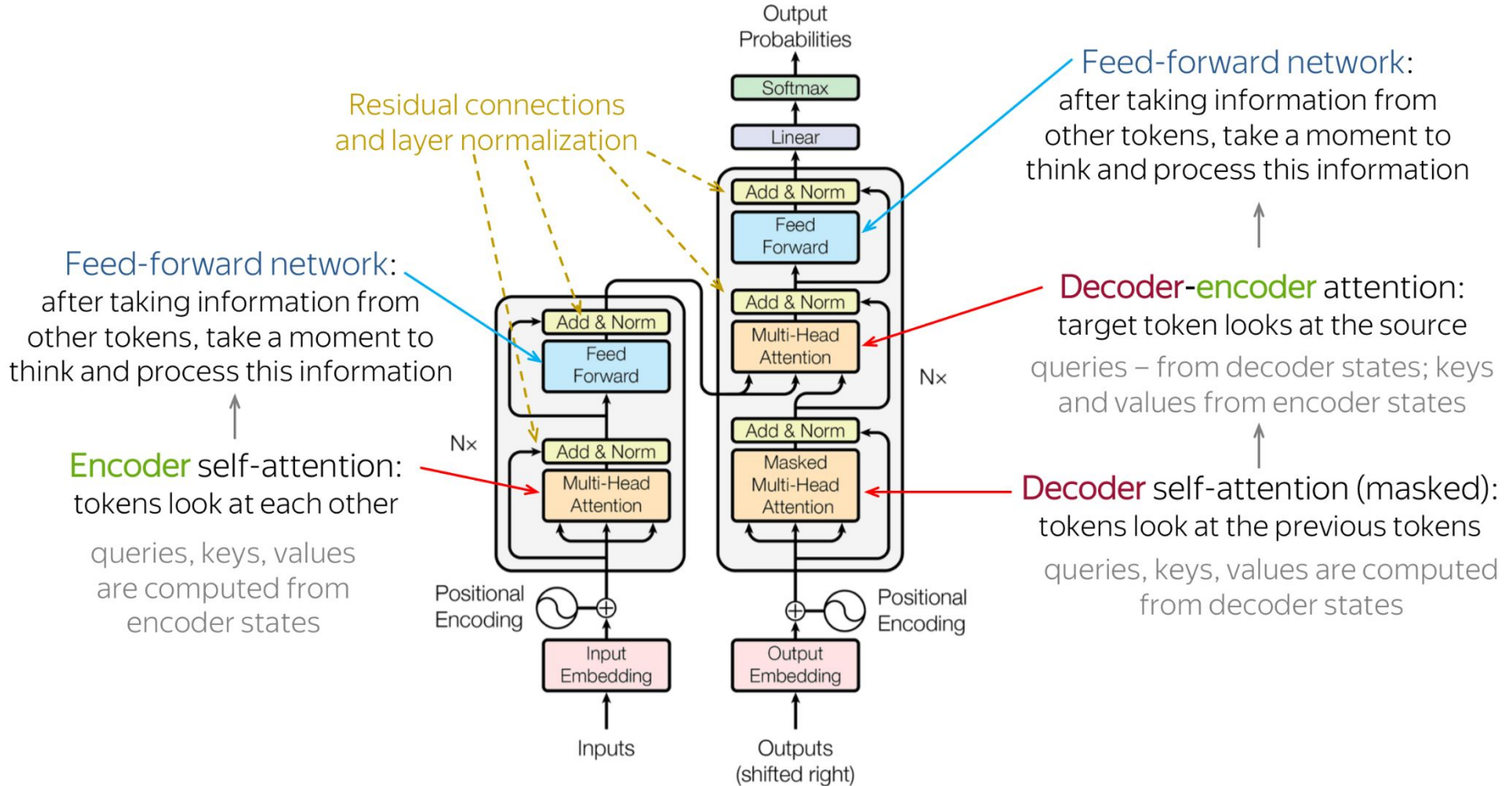
Illia Polosukhin* †
illia.polosukhin@gmail.com

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.



Transformers



Breakthrough 2: Aligning LLMs with user behaviour

Why. Alignment problem → Supervised Finetuning

User question:

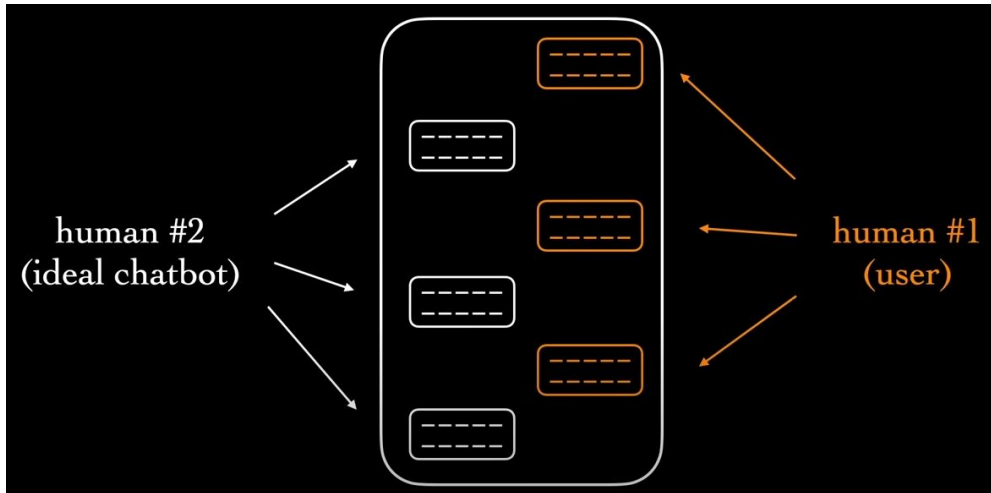
explain how bubble sort algorithm works

Model answer (only with pretraining next word)

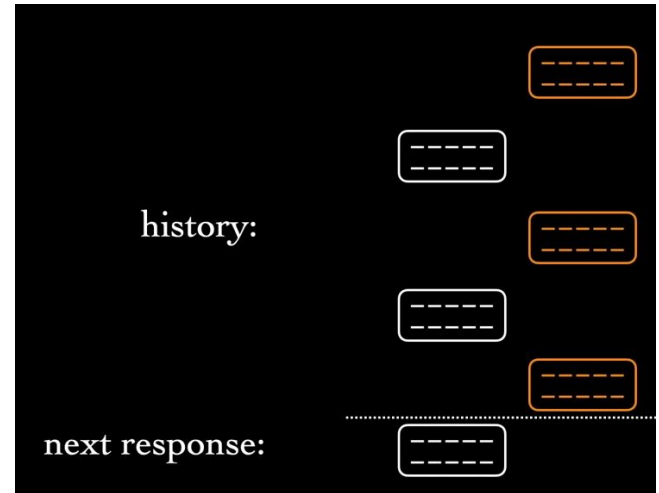
explain how the merge sort algorithm works.

Problem: Predicting next word represents a very large class of tasks within which the one we want is a subset. The task is underspecified and maybe there are subjectively preferred answers.

Generate data with humans



Train dataset



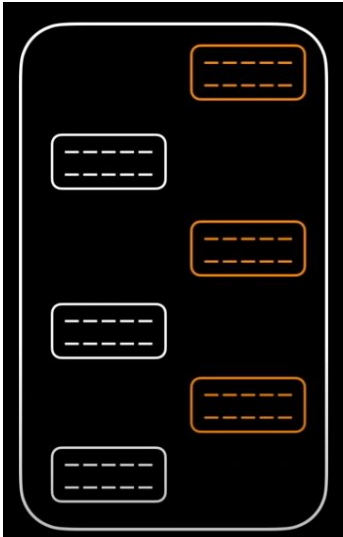
GPT3 - Reinforcement learning

Why. Distribution shift → Reinforcement learning

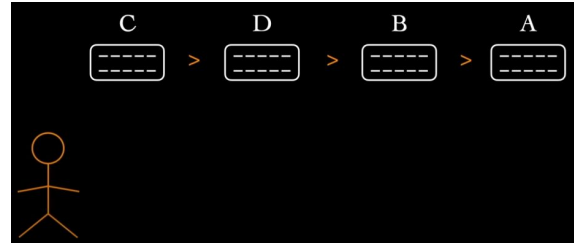
The responses from the supervised learning stage were not producing answers 100% satisfactory
Training with experts but inference is made by users

How to solve it: RL

Supervised model outputs



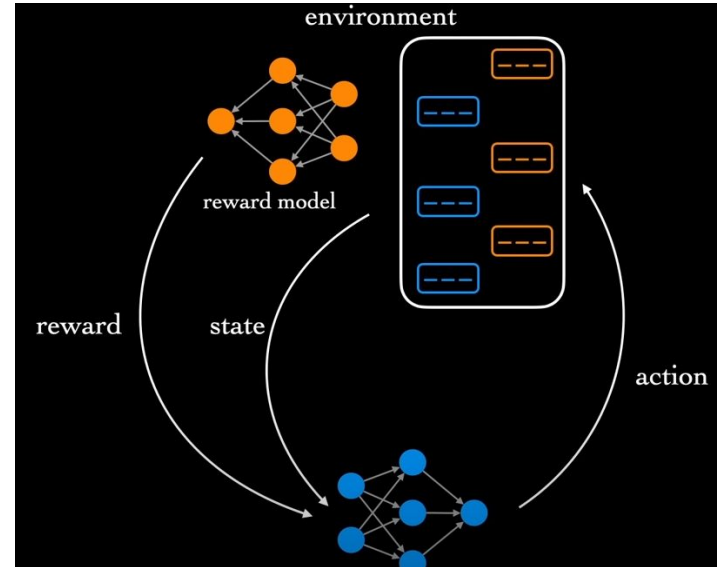
Humans to rank it



Challenge: How to measure reward
→ a *reward model* learns to reward !!!

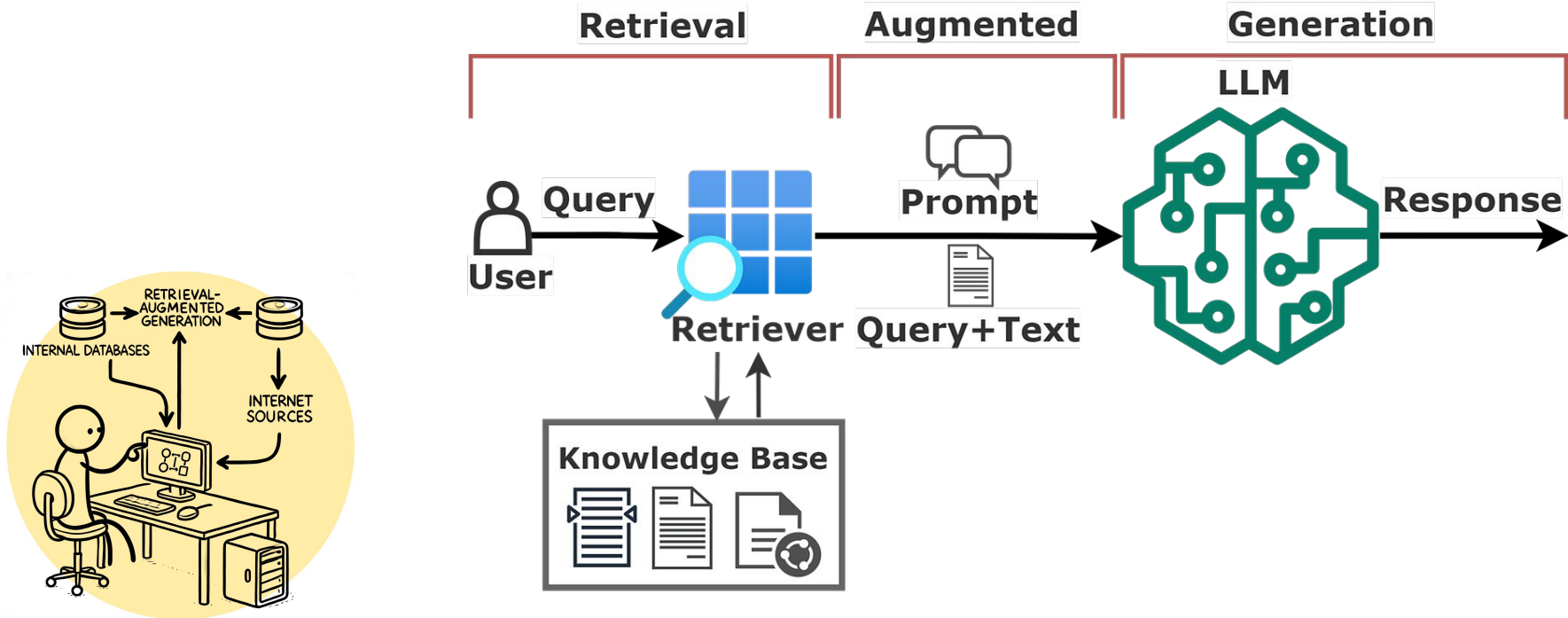
simulating human behaviour trained
on user preference data (scores)

Model to learn the ranking



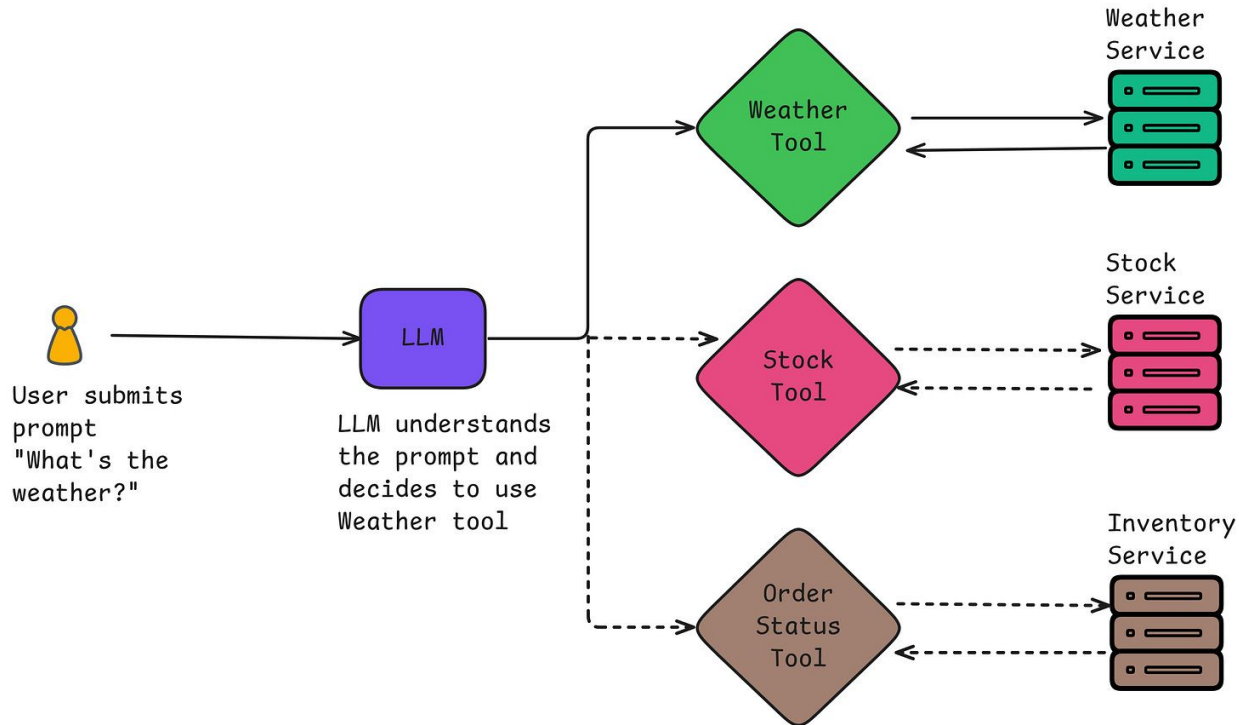
Breakthrough 3: Large context, RAG

Problem. Grounding, Hallucinations, Cutoff knowledge, Proprietary knowledge



Breakthrough 4: Calling external tools

System managing the LLM does several rounds



Calling functions

```
# Define the tool's schema
weather_tool = {
    "name": "get_current_weather",
    "description": "Get the current weather in a given location",
    "parameters": {
        "type": "object",
        "properties": {
            "location": {
                "type": "string",
                "description": "The city and state, e.g. San Francisco,
            }
        },
        "required": ["location"]
    }
}
```

```
from google import genai
from google.genai import types

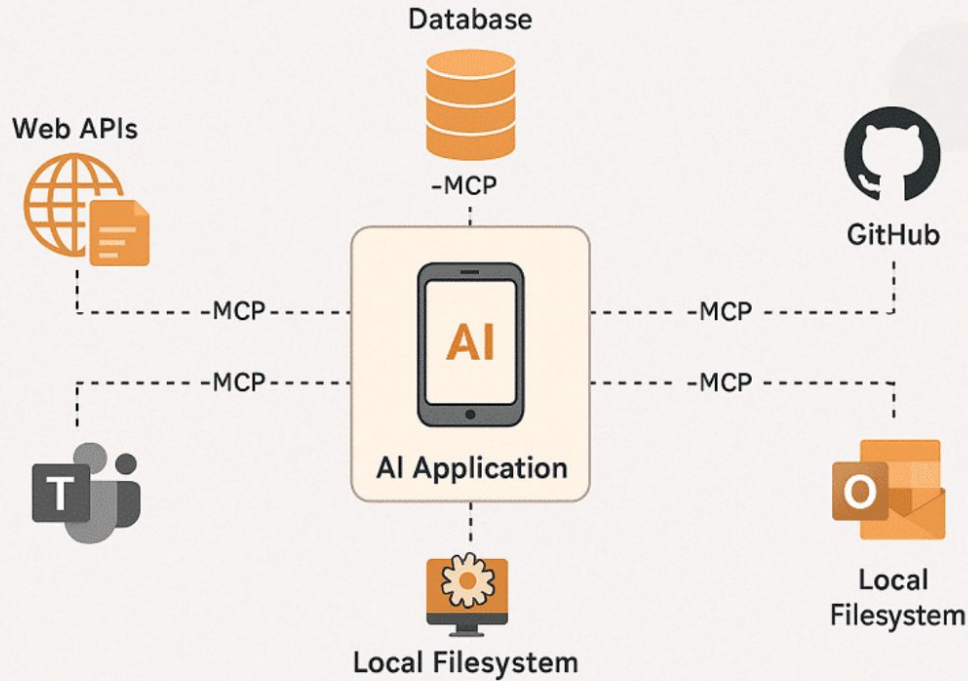
client = genai.Client(api_key="YOUR_API_KEY")
model_id = "gemini-2.0-flash" # Supports latest function calling featu

# Provide the tool to the model
config = types.GenerateContentConfig(
    tools=[types.Tool(function_declarations=[weather_tool])]
)

# Send a prompt that requires the tool
response = client.models.generate_content(
    model=model_id,
    contents="What's the weather like in London?",
    config=config
)
```

Model Context Protocol

What is MCP ?



Server Implementations

Note

We now have a [web-based directory](#) that is synced with the repository.

- [Aggregators](#)
- [Art & Culture](#)
- [Architecture & Design](#)
- [Browser Automation](#)
- [Biology Medicine and Bioinformatics](#)
- [Cloud Platforms](#)
- [Code Execution](#)
- [Coding Agents](#)
- [Command Line](#)
- [Communication](#)
- [Customer Data Platforms](#)
- [Databases](#)
- [Data Platforms](#)
- [Delivery](#)
- [Developer Tools](#)
- [Data Science Tools](#)
- [Data Visualization](#)
- [Embedded system](#)
- [Education](#)
- [E-Commerce](#)
- [Environment & Nature](#)

- [TheLunarCompany/iunar#mcpX](#) - MCPX is a production-ready, open-source gateway to manage MCP servers at scale—centralize tool discovery, access controls, call prioritization, and usage tracking to simplify agent workflows.
- [thinkchainai/mcpbundles](#) - MCP Bundles: Create custom bundles of tools and connect providers with OAuth or API keys. Use one MCP server across thousands of integrations, with programmatic tool calling and MCP UI for managing bundles and credentials.
- [tigranbs/mcgravity](#) - A proxy tool for composing multiple MCP servers into one unified endpoint. Scale your AI tools by load balancing requests across multiple MCP servers, similar to how Nginx works for web servers.
- [toadlyBroodle/satring](#) - Discover and compare L402 + x402 paid API services from satring.com, the best curated Lightning and USDC API directory.
- [VeriTeknik/pluggedin-mcp-proxy](#) - A comprehensive proxy server that combines multiple MCP servers into a single interface with extensive visibility features. It provides discovery and management of tools, prompts, resources, and templates across servers, plus a playground for debugging when building MCP servers.
- [ViperJuice/mcp-gateway](#) - A meta-server for minimal Claude Code tool bloat with progressive disclosure and dynamic server provisioning. Exposes 9 stable meta-tools, auto-starts Playwright and Context7, and can dynamically provision 25+ MCP servers on-demand from a curated manifest.
- [WayStation-ai/mcp](#) - Seamlessly and securely connect Claude Desktop and other MCP hosts to your favorite apps (Notion, Slack, Monday, Airtable, etc.). Takes less than 90 secs.
- [wegotdocs/open-mcp](#) - Turn a web API into an MCP server in 10 seconds and add it to the open source registry: <https://open-mcp.org>
- [whiteknightonhorse/APIbase](#) - Unified API hub for AI agents with 56+ tools across travel (Amadeus, Sabre), prediction markets (Polymarket), crypto, and weather. Pay-per-call via x402 micropayments in USDC.
- [rplyryan/x402-discovery-mcp](#) - Runtime discovery layer for x402-payable APIs. Agents discover and route to pay-per-call x402 endpoints by capability, get quality-ranked results with trust scores (0-100), and pay per query via x402. Includes MCP server, Python SDK, and CLI (npm install -a x402scoutl).

Breakthrough 5: LLM call flows → aka “Agents”

```
creative_prompt = ChatPromptTemplate.from_template(" You are a poetic storyteller. Write a story about: {input}")
tech_prompt = ChatPromptTemplate.from_template(" You are a senior software engineer. Provide a code-focused answer for: {input}")

creative_chain = creative_prompt | model | StrOutputParser()
tech_chain = tech_prompt | model | StrOutputParser()

# 2. Define the Router (The Assessment Step)
router_prompt = ChatPromptTemplate.from_template(
    "Classify the following user input as either 'creative' or 'technical'. "
    "Respond with only one word: 'creative' or 'technical'. \n\nInput: {input}"
)
router_chain = router_prompt | model | StrOutputParser()

# 3. Define the Routing Logic
def route(info):
    if "technical" in info["topic"].lower():
        return tech_chain
    else:
        return creative_chain

# 4. Combine into a single chain
full_chain = {"topic": router_chain, "input": lambda x: x["input"]} | RunnableLambda(route)

# 5. Test it
print(full_chain.invoke({"input": " How do I reverse a linked list in Python?"}))
# Output: [Technical/Code explanation]

print(full_chain.invoke({"input": " Write a haiku about a lonely robot."}))
# Output: [Creative/Poetic response]
```

Breakthrough 5: LLM call flows → aka "Agents"


EXPLORING OPEN-SOURCE AI AGENT FRAMEWORKS


- AutoGen
- LangGraph
- Goog AI
- Semantic Kernel
- OpenAI Agents
- Google ADK
- Hugging Face Smol Agents
- PydanticAI
- LlamaIndex
- DSPy
- Composio



Breakthrough 5: LLM call flows → aka “Agents”

The Core Agentic Loop

Most autonomous systems follow this cycle to move from reactive responses to proactive intelligence:  Facebook +3

- **Observe:** The agent collects data from its environment (sensors, logs, user inputs, or APIs).
- **Orient (or Reason/Plan):** The agent synthesizes this data, applying internal models or LLMs to understand the context and identify the best path forward.
- **Decide:** Using a decision engine (e.g., an LLM or logic gate), the agent selects a specific action or tool to execute.
- **Act:** The agent performs the chosen action (e.g., calling an API, writing code, or sending a message) and observes the outcome, starting the loop over.  Boston Consulting Group +7

What might this mean for science

What might this mean for science

Google Scholar

agentic science



Articles

About 8,540 results (0.08 sec)

Any time

Since 2026

Since 2025

Since 2022

Custom range...

Sort by relevance

Sort by date

Any type

Review articles

include patents

include citations

Create alert

ASTER--**Agentic Science** Toolkit for Exoplanet Research

[PDF] arxiv.org

E Panek, A Roman, G Shukla, L Pagliaro... - arXiv preprint arXiv ..., 2026 - arxiv.org

... Rather than replacing established **scientific** methods, such systems have the potential to ...
-suited field for the development of **agentic scientific** workflows. Modern studies rely on complex ...

☆ Save Cite Cited by 1 Related articles All 2 versions

Toward ultra-long-horizon **agentic science**: Cognitive accumulation for machine learning engineering

[PDF] arxiv.org

X Zhu, Y Cai, Z Liu, B Zheng, C Wang, R Ye... - arXiv preprint arXiv ..., 2026 - arxiv.org

... to validated knowledge and abstract wisdom, we establish a theoretical basis for **agentic science**, enabling agents to sustain coherent reasoning over extended **scientific** workflows. ...

☆ Save Cite Cited by 5 Related articles All 2 versions

Agentic AI and the rise of in silico team **science** in biomedical research

B Li, AK Saini, JG Hernandez, JH Moore - Nature Biotechnology, 2026 - nature.com

... **Agentic AI** systems have been developed for various ... contribute to the development of **agentic AI** systems. We highlight ... **agentic AI** systems to advance collaborative **scientific** research. ...

☆ Save Cite Cited by 3 Related articles All 3 versions

[HTML] **Agentic material science**

[HTML] oaepublish.com

C Li, N Ran, J Liu - Journal of Materials Informatics, 2026 - oaepublish.com

... **science** towards a new stage of agent-driven. This article systematically reviews the progress of AI agents in material **science**... of AI agents with material **science**, thereby accelerating the ...

☆ Save Cite Cited by 1 Related articles

Higher-Order Knowledge Representations for **Agentic Scientific** Reasoning


[PDF] arxiv.org


IA Stewart, MJ Buehler - arXiv preprint arXiv:2601.04878, 2026 - arxiv.org

... We further demonstrate that equipping **agentic** systems with hypergraph ... **agentic** reasoning

What might this mean for science


Core Tasks in Agentic Science


Agentic science streamlines the research lifecycle through several key autonomous functions: 

- **Literature Mining & Synthesis:** Instead of manual reviews, agents scan vast datasets to extract insights, validate literature evidence, and identify knowledge gaps automatically.
- **Hypothesis Generation:** Agents propose novel research directions by identifying hidden correlations across large, complex datasets.
- **Experimental Design & Execution:** Agents can autonomously design protocols, set variables, and interface with laboratory robotics or computational simulators to run experiments without human intervention.
- **Iterative Refinement:** Based on feedback from experimental outcomes, agents refine their models and strategies in real-time, creating a continuous discovery loop.  arXiv +3

What might this mean for science

Core Tasks in Agentic Science


Agentic science streamlines the research lifecycle through several key autonomous functions: 

- **Literature Mining:** Automates the search and analysis of vast scientific literature to identify knowledge gaps.
- **Hypothesis Generation:** Identifies hidden patterns and relationships in data to propose new hypotheses.
- **Experimental Design & Execution:** Agents can autonomously design protocols, set variables, and interface with laboratory robotics or computational simulators to run experiments without human intervention.
- **Iterative Refinement:** Based on feedback from experimental outcomes, agents refine their models and strategies in real-time, creating a continuous discovery loop. 

coding

In Earth Observation

show me flood risk areas in valencia, spain

<> Full Response  Orchestration

✓ Created plan

Agent Plan

1. **Retrieval:** Fetch the polygon geometry for Valencia, Spain (locality) using the Administrative Boundaries tool to define the area of interest.
2. **Retrieval:** Identify flood-prone regions within the Valencia boundary using Earth Engine's flood identification tool, utilizing methods such as historical surface water occurrence and floodplain mapping to assess long-term susceptibility.
3. **Retrieval:** Retrieve any currently predicted or forecasted flood impact areas for the Valencia region using the Flood Forecasting API to identify immediate risks.



✓ Executed plan

I have analyzed the flood risk areas in Valencia, Spain, by examining both historical surface water occurrence and current flood forecasts.

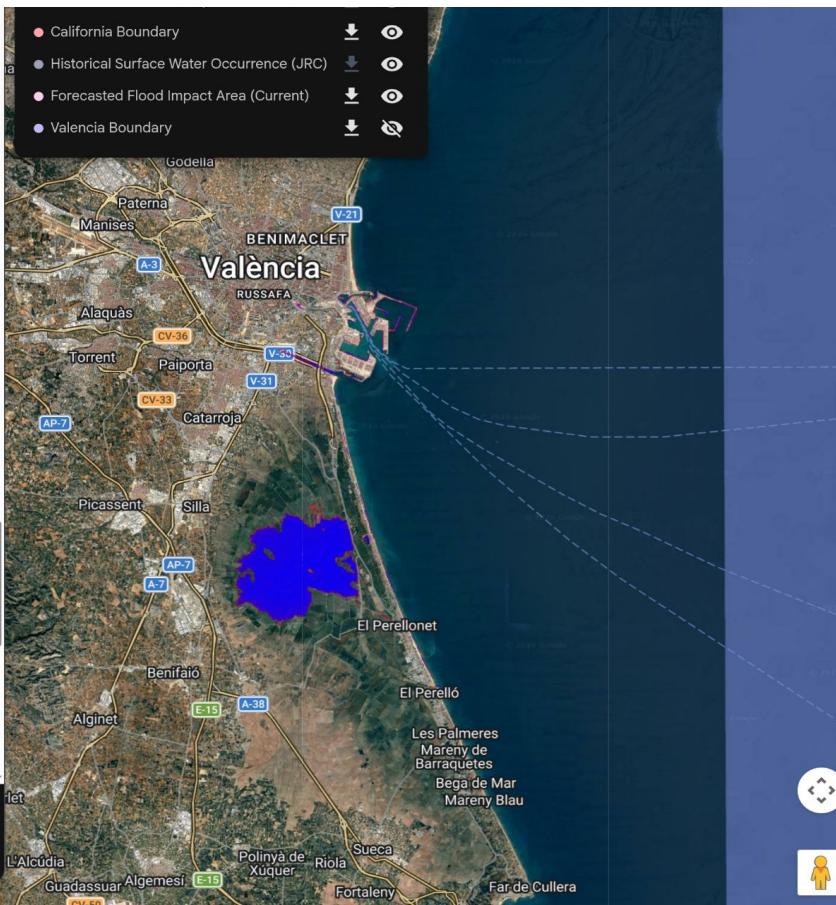
1. Historical Flood Risk (Long-Term Susceptibility)

Using the **JRC Global Surface Water dataset**, I have mapped areas that have historically been covered by water between 1984 and 2021. This provides a reliable indicator of flood-prone regions, including natural floodplains and areas with poor drainage.

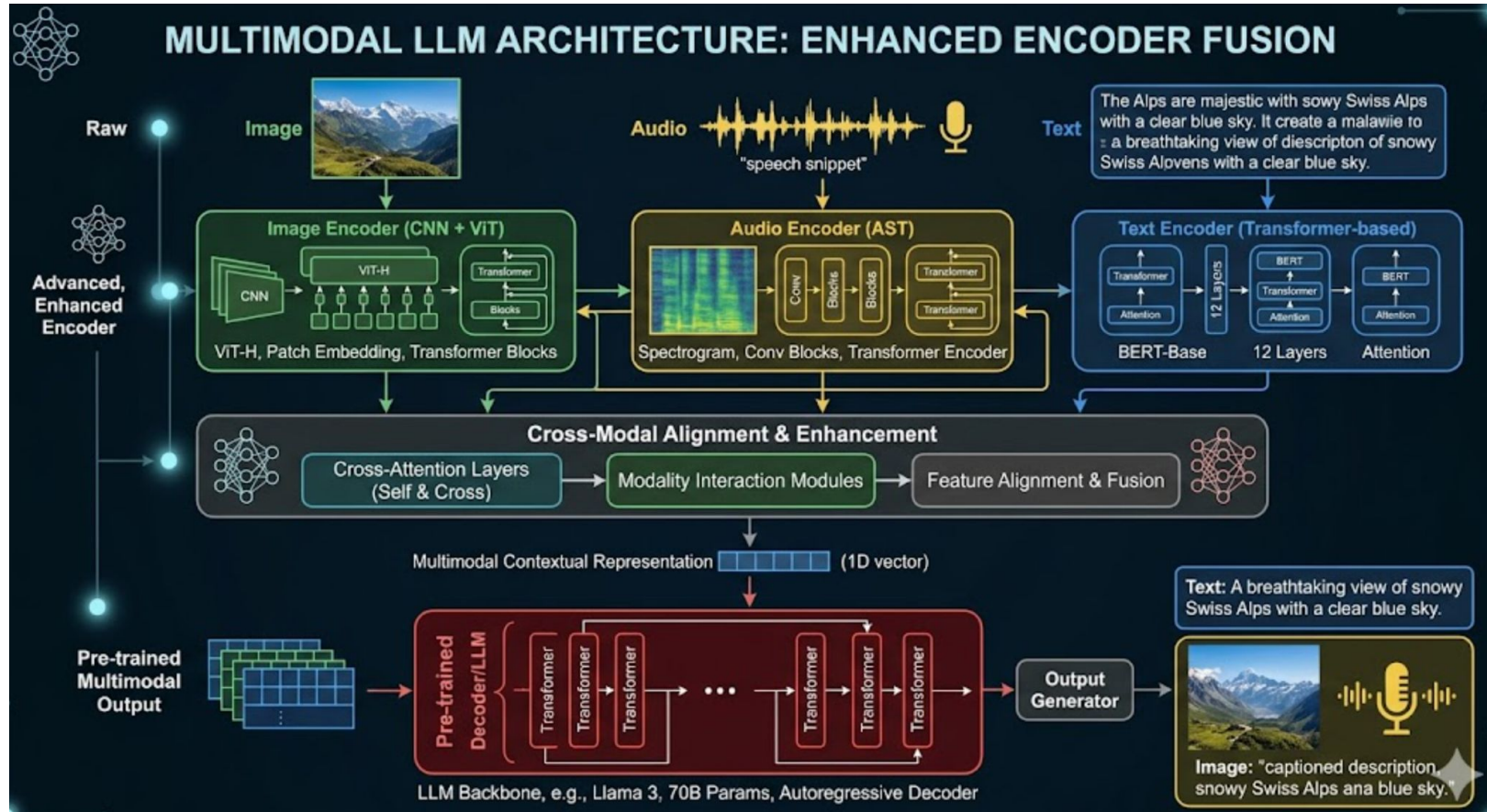
- **Map Layer:** "Historical Surface Water Occurrence (JRC)" has

 Enter prompt 

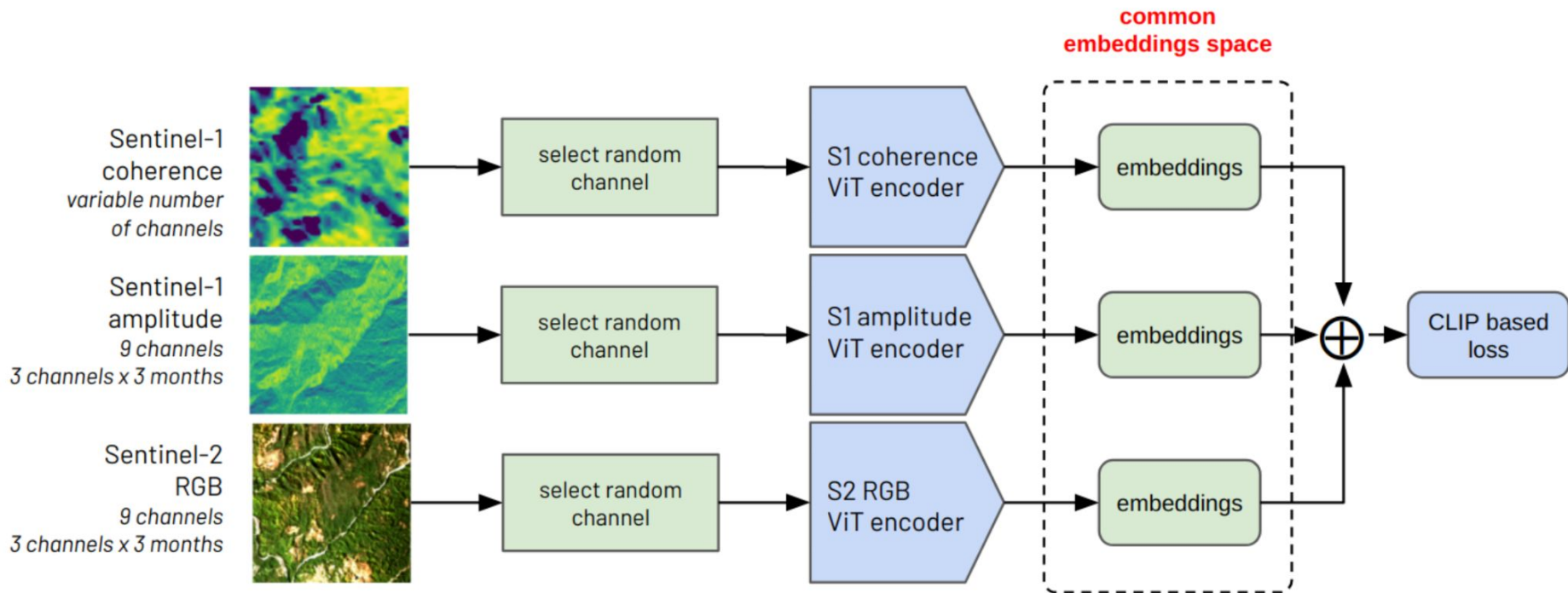
Generative AI experiment · 2026-04-08-a8e4045



Scientific instrumentation



Scientific instrumentation



Validation / Evaluation

**How do I trust an LLM/Agent suggestion
so as to invest effort in pursuing it?**

Validation / Evaluation – Sensitivity Analysis

question

Determine the surface area of Lago Grande in Southern Italy

output variables

- **lago_grande_surface_area** (km²): The surface area of Lago Grande in Southern Italy, measured in square kilometers

input variables (in the code)

scale:

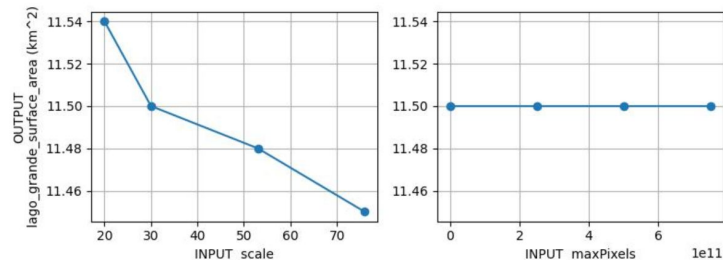
- **meaning:** This variable defines the spatial resolution (in meters) at which the reduction operation calculates the total water surface area.
- **values range:** [10, 100]
- **estimated impact:** MEDIUM
- **impact justification:** The scale dictates the pixel size during the region reduction process. Altering this value forces Earth Engine to resample the binary water mask. This changes how boundary/edge pixels are included or excluded from the geometry, causing slight variations in the computed total surface area.

maxPixels:

- **meaning:** This constant sets the upper limit on the number of pixels allowed to be processed by the reducer to prevent memory/computation overloads.
- **values range:** [10000000.0, 1000000000000.0]
- **estimated impact:** LOW
- **impact justification:** This acts strictly as a computational safeguard. Modifying this threshold does not change the mathematical outcome or precision of the final surface area calculation; it only dictates whether the execution succeeds or aborts.

sensitivity with respect to variables in the code

output vars in y-axes, input vars in x-axes



Validation / Evaluation – Factuality checking

issue 1: Region of Interest (ROI) Accuracy -->

Can the bounding box coordinates be verified against authoritative lake boundary data (e.g., from ISTAT, Italian environmental agencies, or high-resolution satellite imagery)?

Are there known seasonal or long-term fluctuations in Lago Grande's extent that could affect the ROI?

issue 2: JRC Global Surface Water Dataset Limitations -->

Is a 50% occurrence threshold appropriate for Lago Grande, or should it be adjusted based on local hydrological patterns?

Are there higher-resolution or more recent datasets (e.g., Sentinel-2) that could improve accuracy for this specific lake?

issue 3: Pixel Area Calculation and Projection -->

Should the analysis reproject the data to a local equal-area coordinate system (e.g., UTM zone 33N) to minimize area calculation errors?

Are there known issues with `ee.Image.pixelArea()` for small lakes at this latitude?

issue 4: Temporal Representativeness -->

Are there recent (post-2021) changes in Lago Grande's surface area due to climate or anthropogenic factors?

Would using a shorter time window (e.g., 2010–2020) or a specific season improve accuracy?

issue 5: Edge Effects and Mixed Pixels -->

Would a subpixel water detection method (e.g., using Sentinel-2 NDWI) significantly improve area estimates for Lago Grande?

Are there known issues with JRC GSW's edge detection for small volcanic lakes?

issue 6: Validation Against Ground Truth -->

Are there published or official surface area measurements for Lago Grande that could serve as ground truth?

What is the typical error margin for JRC GSW-derived areas in this region?

issue 7: Scale and `maxPixels` Parameters -->

Is 30-meter scale appropriate, or should it match a specific output projection?

Could a smaller `maxPixels` value improve efficiency without affecting accuracy?

Assessment of Coordinate Accuracy and Buffer Radius for Lago Grande

1. Coordinate Accuracy (15.6111°E, 40.9306°N)

- **Verification:** The provided coordinates (15.6111°E, 40.9306°N) are **approximately correct** for **Lago Grande** (Basilicata, Italy). However, they represent a point **near the lake's southeastern shore**.
 - **Centroid Estimate:** Based on open-source geospatial data (e.g., OpenStreetMap, Sentinel-2 image 40.9320°N (shifted ~500m northwest).
 - **Impact:**
 - The current point is **within the lake's boundary**, so the buffer-based selection will still capture the lake.
 - However, using a **non-centroid point** may bias the `filterBounds()` operation if multiple waters are present.
- **Recommendation:** Use the **centroid coordinates** (15.6050°E, 40.9320°N) to ensure the buffer is not excluding or including unrelated features.

2. Buffer Radius (2000 meters)

- **Context:**
 - Lago Grande has a **surface area of ~0.38 km²** (varies seasonally) and a **maximum diameter of ~700 meters**.
 - Lago Piccolo (the smaller twin lake) lies **~600 meters to the west**.
 - A **2000-meter buffer** is **excessively large** for this analysis:
 - It captures **both Monticchio lakes** and potentially other small water bodies in the region (e.g., Lago Piccolo).
 - The JRC Global Surface Water dataset may include **temporary or seasonal water bodies** within the buffer.
- **Optimal Buffer Size:**
 - A **1000-meter buffer** would suffice to isolate Lago Grande while excluding Lago Piccolo.
 - For higher precision, use a **polygon boundary** (e.g., from OpenStreetMap or manual digitization) instead of a buffer.

3. Additional Risks

- **JRC GSW Dataset Limitations:**
 - The `max_extent` band (used in the code) shows **historical maximum water extent**, which may include areas that are not currently water.
 - Lago Grande's water level fluctuates seasonally; the JRC dataset might overestimate its area if the lake is full.
- **Vector Conversion Artifacts:**
 - `reduceToVectors()` with `scale=30` (Landsat resolution) may produce **jagged polygons** for small features.

Visual Confirmation (Expected Output)

If the current code runs successfully:

- The **red polygon** (Lago Grande) should align with the **western basin** of the two Monticchio lakes.
- If the red polygon **includes both lakes**, the buffer is too large.
- If the polygon is **offset from the blue water mask**, the coordinates may need adjustment.

RECOMMENDATION update_coordinates

explanation :: Replace the point coordinates with Lago Grande's centroid to ensure accuracy.

```
code_snippet ::
# Updated centroid coordinates for Lago Grande
lon, lat = 15.6050, 40.9320 # Centroid (approx.)
point = ee.Geometry.Point([lon, lat])
```

RECOMMENDATION reduce_buffer_radius

explanation :: Reduce the buffer to 1000 meters to exclude Lago Piccolo and other nearby features.

```
code_snippet ::
# Smaller buffer to isolate Lago Grande
roi = point.buffer(1000) # 1 km radius
```

RECOMMENDATION add_manual_polygon_filter

explanation :: Optionally, use a predefined polygon (e.g., from OpenStreetMap) to filter out other water bodies.

```
code_snippet ::
# Example: Replace buffer with a manual polygon (uncomment if using)
# lago_grande_polygon = ee.Geometry.Polygon([[
#   [15.600, 40.930], [15.605, 40.935], # Adjust vertices to match lake boundary
#   [15.610, 40.928], [15.600, 40.925]
# ]])
# roi = lago_grande_polygon
```

RECOMMENDATION refine_vectorization_scale

explanation :: Use a finer scale (e.g., 10m) for smoother lake boundaries if high-resolution data is available.

Parallels with Dot Com bubble in the 90s

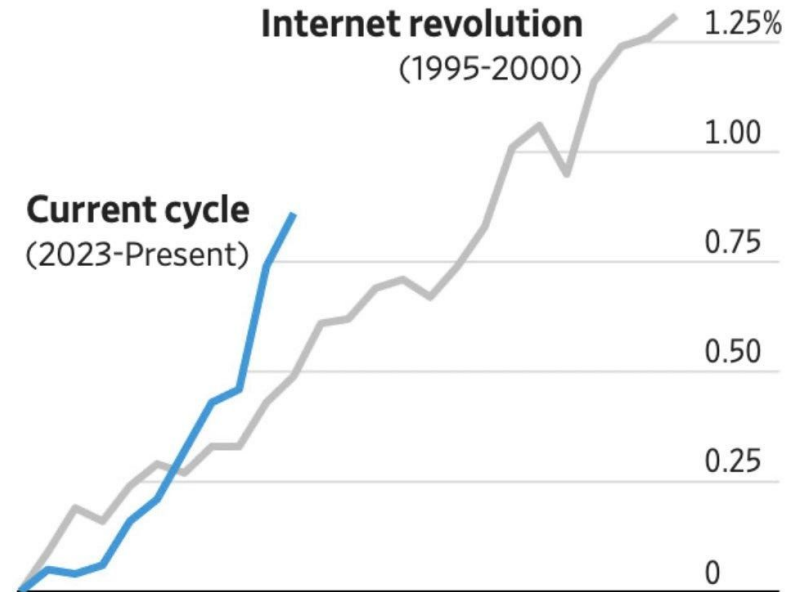


Dotcom Bubble

[ˈdɑːt-,kəm ˈbæ-bəl]

A rapid rise in U.S. technology stock equity valuations fueled by investments in internet-based companies during the late 1990s.

Investment in AI and internet as a percentage of GDP, during the...



Source: Citi Research

A.G.I.

LLMs are “just” an statistical model

raul.ramos@udea.edu.co

THnx

