RAG de conexiones

Santiago Rivas Betancourt - 61007
Juan Ignacio Matilla - 60459
Franco Panighini - 61258

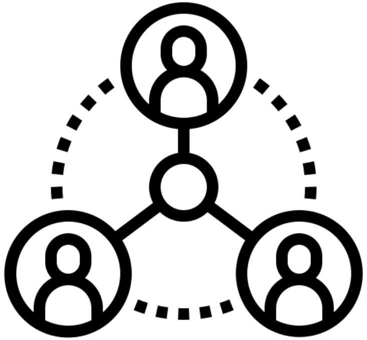
16 de Mayo 2025

Problema a resolver

El desafío de las conexiones significativas en eventos

En los eventos hay poco tiempo y mucha gente, esto lleva a que se pierdan oportunidades.

Objetivo: Generar las conexiones de calidad a gran escala, comparando el perfil de una persona con todos los participantes del evento.



Cómo lo hacemos ahora?

Hay un agente conversacional que investiga interes y caracteristicas de la persona, luego genera un "*dossier*" que es un resumen de la conversación.

Luego de que se generaron todos los "*dossiers*", se le pasa un prompt a un LLM pidiendo que haga matches y se appendean todos los "*dossiers*" debajo.

Porque RAG?

La manera actual funciona bien mientras se intenten matchear pocas personas entre sí, pero al momento de escalar este sistema para eventos más masivos, no se puede incluir todo en el prompt directamente ya que se queda sin cantidad tokens permitidos por el LLM.

Además al aumentar tanto la context windows en el prompt el LLM empieza a perder calidad en los matches que genera.

Desafíos encontrados

Desafíos encontrados

- Generación apropiada de data sintética mediante una API:
    - 500 Dossiers (Markdown files)
    - 500 Research (HTML)
- Testear cada cambio es costoso, tanto en tiempo como en dinero
- Comprobar que cada métrica evaluada este efectivamente siendo bien calculada

Decisiones de implementación

Stack de alto nivel

- Ingesta multi-fuente: Markdown dossiers + HTML crudo
- Vector Store: ChromaDB + Filtros de metadatos
- LLM: GPT-4o-mini
- Embeddings: HuggingFaceEmbeddings
- TextSplitter: TokenTextSplitter

Datos sintéticos

Porqué datos sintéticos?

- Necesitamos volumen y diversidad antes de tener usuarios reales.
- Permite simular variedad de perfiles y proyectos.
- Permite experimentar con distintos parámetros sin exponer datos reales.

Generación de los datos sintéticos

Dossiers

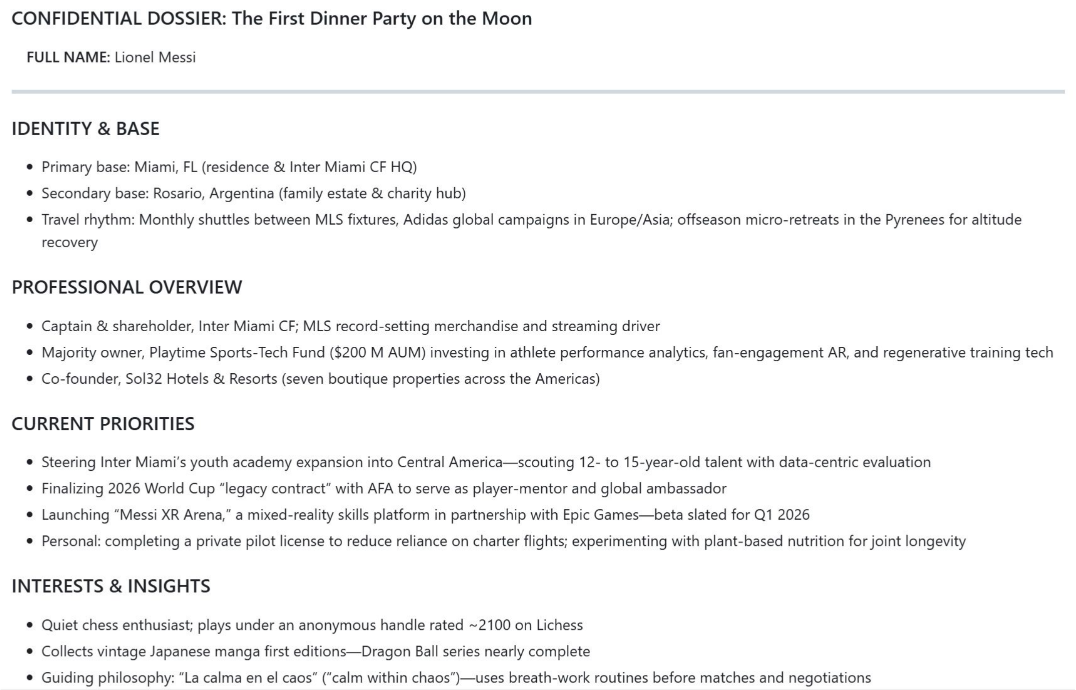
- Se creó un prompt que estructura la salida para que los dossiers generados sean similares, pero dando ’libertad creativa’ para que el contenido sea acorde al perfil de cada participante.

HTML - Public Information

- Se creó un prompt que hace *’scraping’* de la web y genera un archivo HTML con formato similar a un perfil de Wikipedia.
- El HTML contiene enlaces,tablas y distintas secciones. Toda la información está en la public web.

Dossiers

- Summarization generada por un LLM de una conversación con un agente sobre interés y preferencias, formato Markdown:



CONFIDENTIAL DOSSIER: The First Dinner Party on the Moon

FULL NAME: Lionel Messi

IDENTITY & BASE

- Primary base: Miami, FL (residence & Inter Miami CF HQ)
- Secondary base: Rosario, Argentina (family estate & charity hub)
- Travel rhythm: Monthly shuttles between MLS fixtures, Adidas global campaigns in Europe/Asia; offseason micro-retreats in the Pyrenees for altitude recovery

PROFESSIONAL OVERVIEW

- Captain & shareholder, Inter Miami CF; MLS record-setting merchandise and streaming driver
- Majority owner, Playtime Sports-Tech Fund ($200 M AUM) investing in athlete performance analytics, fan-engagement AR, and regenerative training tech
- Co-founder, Sol32 Hotels & Resorts (seven boutique properties across the Americas)

CURRENT PRIORITIES

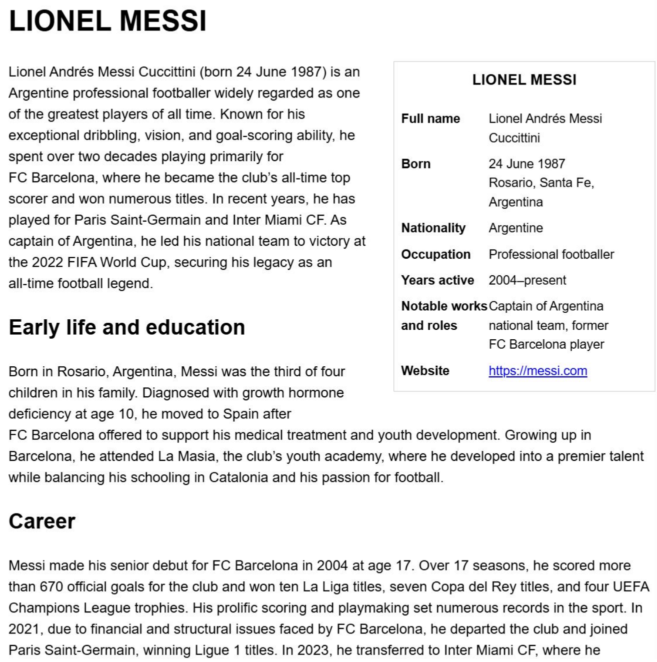
- Steering Inter Miami's youth academy expansion into Central America—scouting 12- to 15-year-old talent with data-centric evaluation
- Finalizing 2026 World Cup "legacy contract" with AFA to serve as player-mentor and global ambassador
- Launching "Messi XR Arena," a mixed-reality skills platform in partnership with Epic Games—beta slated for Q1 2026
- Personal: completing a private pilot license to reduce reliance on charter flights; experimenting with plant-based nutrition for joint longevity

INTERESTS & INSIGHTS

- Quiet chess enthusiast; plays under an anonymous handle rated ~2100 on Lichess
- Collects vintage Japanese manga first editions—Dragon Ball series nearly complete
- Guiding philosophy: "La calma en el caos" ("calm within chaos")—uses breath-work routines before matches and negotiations

Public information

- Representa información pública que esté disponible en internet del entrevistado, para poder aumentar más el conocimiento de cada persona, formato HTML:



LIONEL MESSI

Lionel Andrés Messi Cuccittini (born 24 June 1987) is an Argentine professional footballer widely regarded as one of the greatest players of all time. Known for his exceptional dribbling, vision, and goal-scoring ability, he spent over two decades playing primarily for FC Barcelona, where he became the club's all-time top scorer and won numerous titles. In recent years, he has played for Paris Saint-Germain and Inter Miami CF. As captain of Argentina, he led his national team to victory at the 2022 FIFA World Cup, securing his legacy as an all-time football legend.

LIONEL MESSI	
Full name	Lionel Andrés Messi Cuccittini
Born	24 June 1987 Rosario, Santa Fe, Argentina
Nationality	Argentine
Occupation	Professional footballer
Years active	2004–present
Notable works and roles	Captain of Argentina national team, former FC Barcelona player
Website	https://messi.com

Early life and education

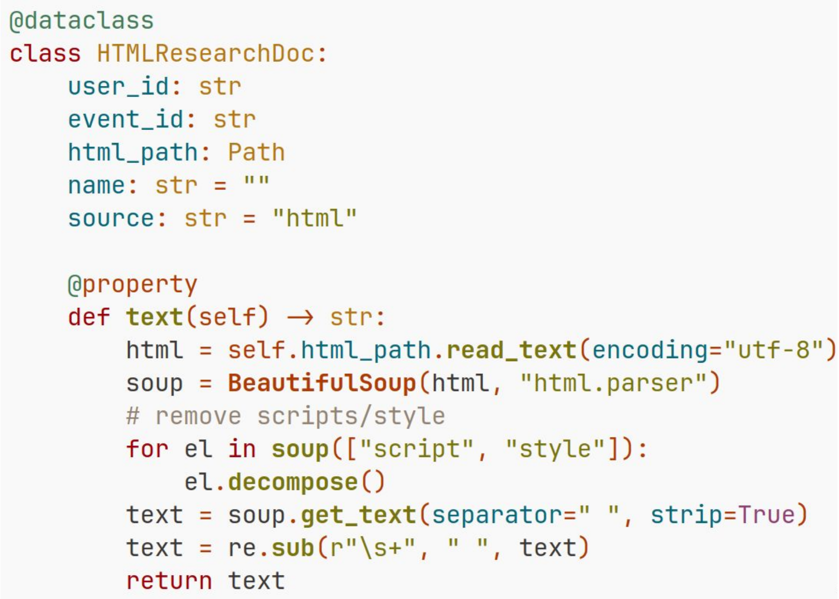
Born in Rosario, Argentina, Messi was the third of four children in his family. Diagnosed with growth hormone deficiency at age 10, he moved to Spain after FC Barcelona offered to support his medical treatment and youth development. Growing up in Barcelona, he attended La Masia, the club's youth academy, where he developed into a premier talent while balancing his schooling in Catalonia and his passion for football.

Career

Messi made his senior debut for FC Barcelona in 2004 at age 17. Over 17 seasons, he scored more than 670 official goals for the club and won ten La Liga titles, seven Copa del Rey titles, and four UEFA Champions League trophies. His prolific scoring and playmaking set numerous records in the sport. In 2021, due to financial and structural issues faced by FC Barcelona, he departed the club and joined Paris Saint-Germain, winning Ligue 1 titles. In 2023, he transferred to Inter Miami CF, where he

Preprocesamiento HTML

- HTML se preprocesa mediante BeautifulSoup para convertirlo a texto plano, eliminando elementos script y style.



```python
@dataclass
class HTMLResearchDoc:
    user_id: str
    event_id: str
    html_path: Path
    name: str = ""
    source: str = "html"

    @property
    def text(self) -> str:
        html = self.html_path.read_text(encoding="utf-8")
        soup = BeautifulSoup(html, "html.parser")
        # remove scripts/style
        for el in soup(["script", "style"]):
            el.decompose()
        text = soup.get_text(separator=" ", strip=True)
        text = re.sub(r"\s+", " ", text)
        return text
```

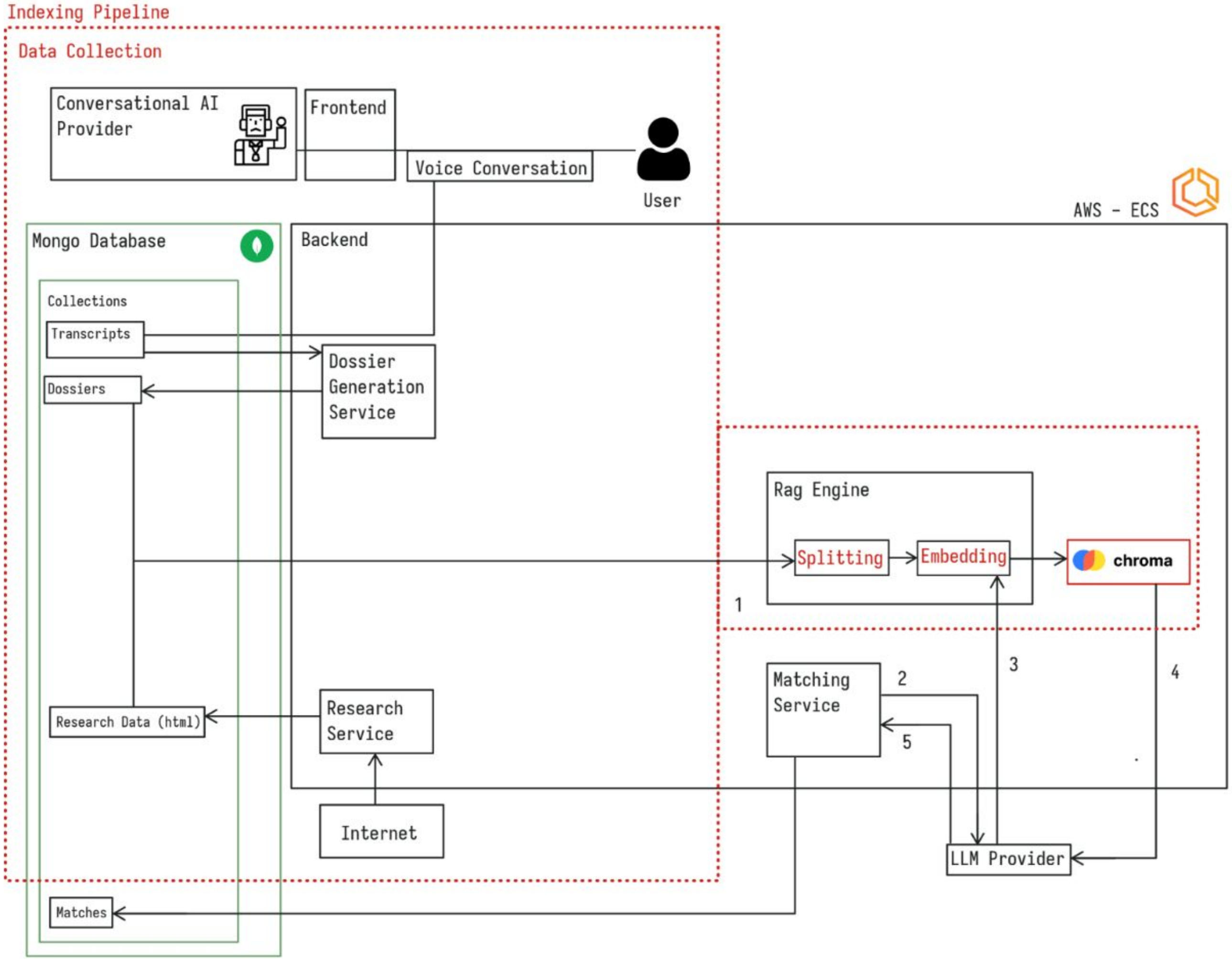
Preprocesamiento Markdown

- Se elimina el título que contienen todos los dossiers:

```
### CONFIDENTIAL DOSSIER: The First Dinner Party on the Moon
```

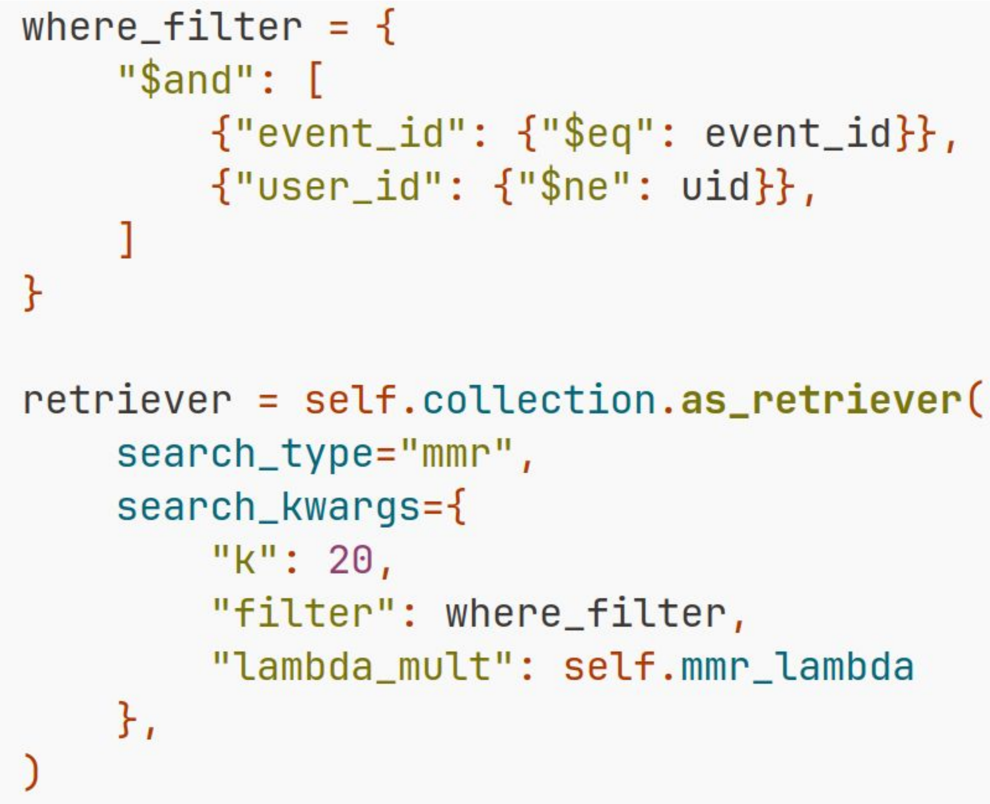
- En este caso son datos sintéticos, pero con datos reales se debería hacer lo mismo por el formato en el que se generan

Arquitectura


Indexing Pipeline
Data Collection
Conversational AI Provider
Frontend
Voice Conversation
User
AWS - ECS
Mongo Database
Backend
Collections
Transcripts
Dossiers
Dossier Generation Service
Rag Engine
Splitting
Embedding
chroma
1
Matching Service
2
3
4
5
Research Data (html)
Research Service
Internet
LLM Provider
Matches

Técnica de Advance RAG

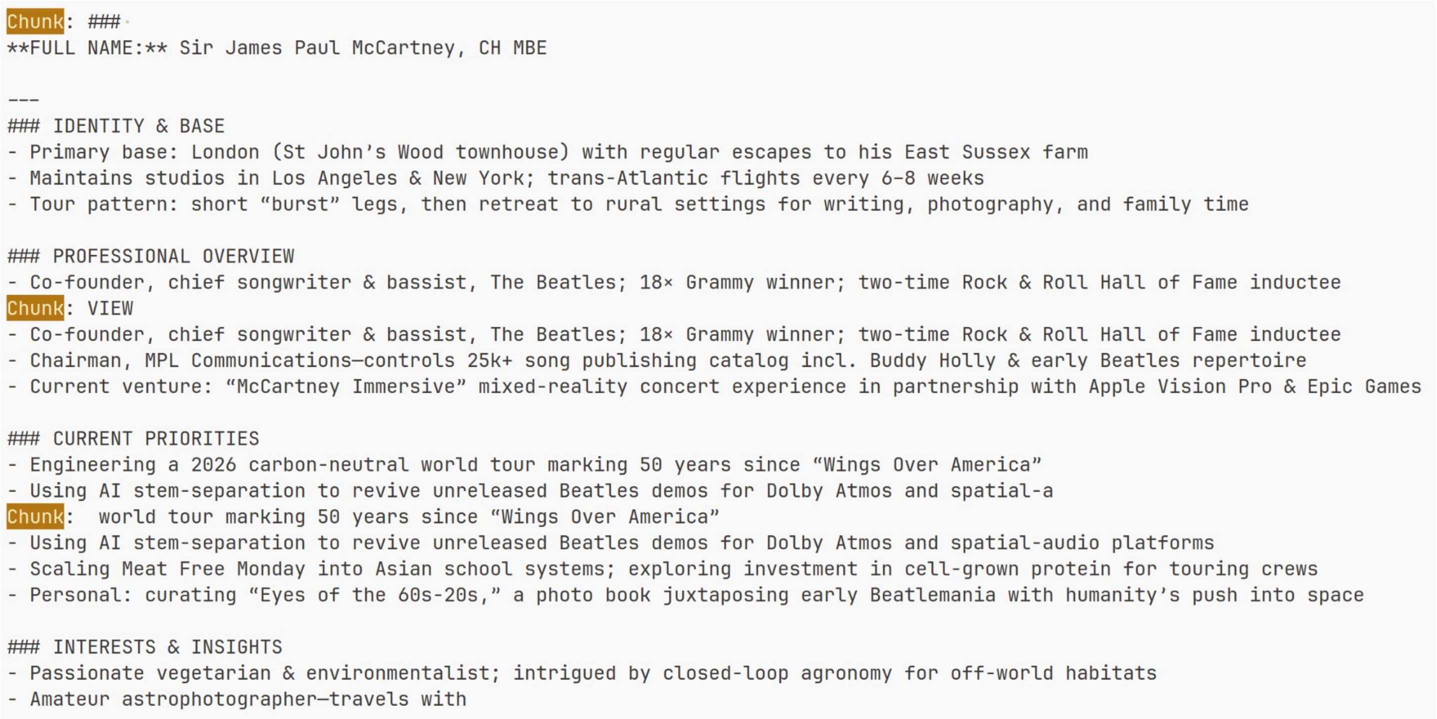
Filtrado por metadata



```
where_filter = {
    "$and": [
        {"event_id": {"$eq": event_id}},
        {"user_id": {"$ne": uid}},
    ]
}

retriever = self.collection.as_retriever(
    search_type="mmr",
    search_kwargs={
        "k": 20,
        "filter": where_filter,
        "lambda_mult": self.mmr_lambda
    },
)
```

Ejemplo

Chunking e indexado



```
Chunk: ###
**FULL NAME:** Sir James Paul McCartney, CH MBE

---
### IDENTITY & BASE
- Primary base: London (St John’s Wood townhouse) with regular escapes to his East Sussex farm
- Maintains studios in Los Angeles & New York; trans-Atlantic flights every 6–8 weeks
- Tour pattern: short “burst” legs, then retreat to rural settings for writing, photography, and family time

### PROFESSIONAL OVERVIEW
- Co-founder, chief songwriter & bassist, The Beatles; 18× Grammy winner; two-time Rock & Roll Hall of Fame inductee
Chunk: VIEW
- Co-founder, chief songwriter & bassist, The Beatles; 18× Grammy winner; two-time Rock & Roll Hall of Fame inductee
- Chairman, MPL Communications—controls 25k+ song publishing catalog incl. Buddy Holly & early Beatles repertoire
- Current venture: “McCartney Immersive” mixed-reality concert experience in partnership with Apple Vision Pro & Epic Games

### CURRENT PRIORITIES
- Engineering a 2026 carbon-neutral world tour marking 50 years since “Wings Over America”
- Using AI stem-separation to revive unreleased Beatles demos for Dolby Atmos and spatial-a
Chunk:  world tour marking 50 years since “Wings Over America”
- Using AI stem-separation to revive unreleased Beatles demos for Dolby Atmos and spatial-audio platforms
- Scaling Meat Free Monday into Asian school systems; exploring investment in cell-grown protein for touring crews
- Personal: curating “Eyes of the 60s-20s,” a photo book juxtaposing early Beatlemania with humanity’s push into space

### INTERESTS & INSIGHTS
- Passionate vegetarian & environmentalist; intrigued by closed-loop agronomy for off-world habitats
- Amateur astrophotographer—travels with
```

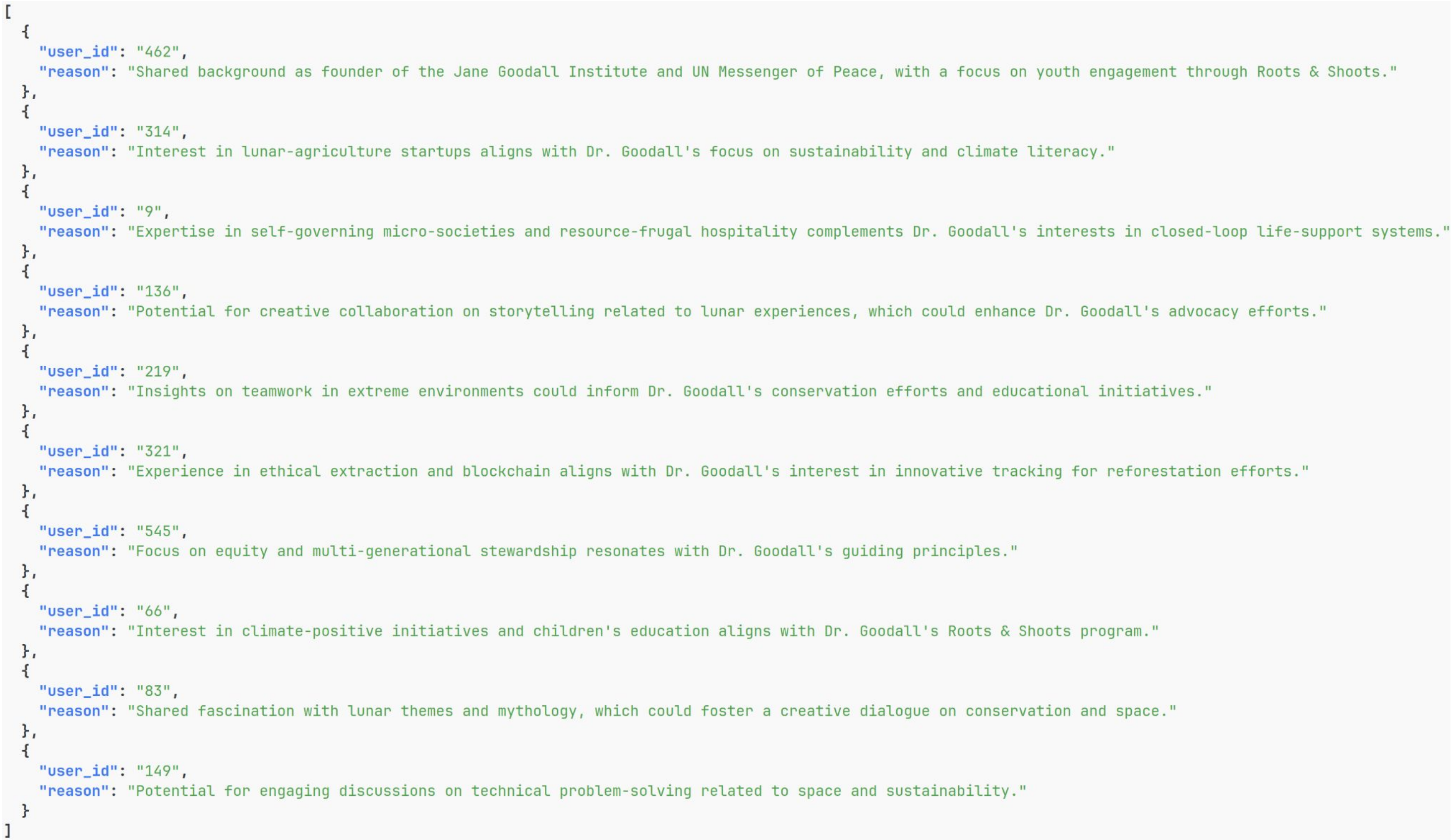
```
chunk size    = 128
chunk overlap = 32
```

Prompt

```
You are an expert networking assistant.
Given the following context chunks about other participants in the
same event, return a JSON list of the top 10 people that should
match with user '1' (Sir James Paul McCartney, CH MBE).
This is the dossier of 1:

### CONFIDENTIAL DOSSIER: The First Dinner Party on the Moon

**FULL NAME:** Sir James Paul McCartney, CH MBE
---

### IDENTITY & BASE

- Primary base: London (St John’s Wood townhouse) with regular
escapes to his East Sussex farm
- Maintains studios in Los Angeles & New York; trans-Atlantic
flights every 6–8 weeks
- Tour pattern: short “burst” legs, then retreat to rural settings
for writing, photography, and family time

. . .
```

```
Each item must be an object {"user_id": str, "reason": str}.
Only output valid JSON without markdown fences.

Context (JSON):
[
  {
    "user_id": "488",
    "name": "Liu Zihong",
    "chunk": "6 g\n\n### NETWORKING & OPPORTUNITIES\n- Warm intros desired: SpaceX
Starship materials team; Axiom Space habitat architects; Sony Pictures Immersive for
cross-reality storytelling\n- Open to co-funding student-led lunar experiments featuring
flexible sensors; offers factory time and mentorship\n- Best engagement: invite opinion
on how “soft tech” changes hard environments—he loves white-board napkin sketches over
dessert\n\n### ADDITIONAL NOTES\n- Avoid rehashing Royole’s delayed 2021 IPO; frame
questions around future resilience, not past liquidity crunch\n- Prefers concise stats
first, philosophy second; will dis"
  },
  {
    "user_id": "461",
    "name": "James Tour",
    "chunk": "NETWORKING & OPPORTUNITIES\n- High-value intros: NASA’s in-situ materials
team (Dr. Violet Zhang), SpaceX Starship payload procurement, and Mubadala Ventures’
advanced materials fund\n- Helpful resources: lunar dust simulant access, micro-gravity
furnace time slots, co-marketing with luxury watchmakers using Tour-Lab carbon
nano-shells\n- Engagement tips: open with music gear or superbikes; pivot to “How would
you build a bridge on the Moon?”—he’ll dive in enthusiastically\n\n### ADDITIONAL
NOTES\n- Avoid long evolutionary origin debates; he finds them draining in social
settings"
  },
  {
    "user_id": "55",
    "name": "Chance the Rapper",
    "chunk": "partial gravity\n\n### NETWORKING & OPPORTUNITIES\n- Warm intros sought:
SpaceX Inspiration4 media team (for joint charity broadcast), ESA astronaut Dr. Samantha
Cristoforetti (mental-health insights), Nike Spacewear R&D lead\n- Could mentor emerging
creators in Planta’s network on independent monetization models\n- Best engagement: Ask
him to freestyle using lunar-inspired word prompts; segue into map collection or
parabolic-flight stories for deeper connection\n\n### ADDITIONAL NOTES\n- Sensitive:
Prefers to avoid lengthy discussions about major-label deals; values his indie status\n-
Unusual fact prompt: He is mastering the"
  },

. . .
```

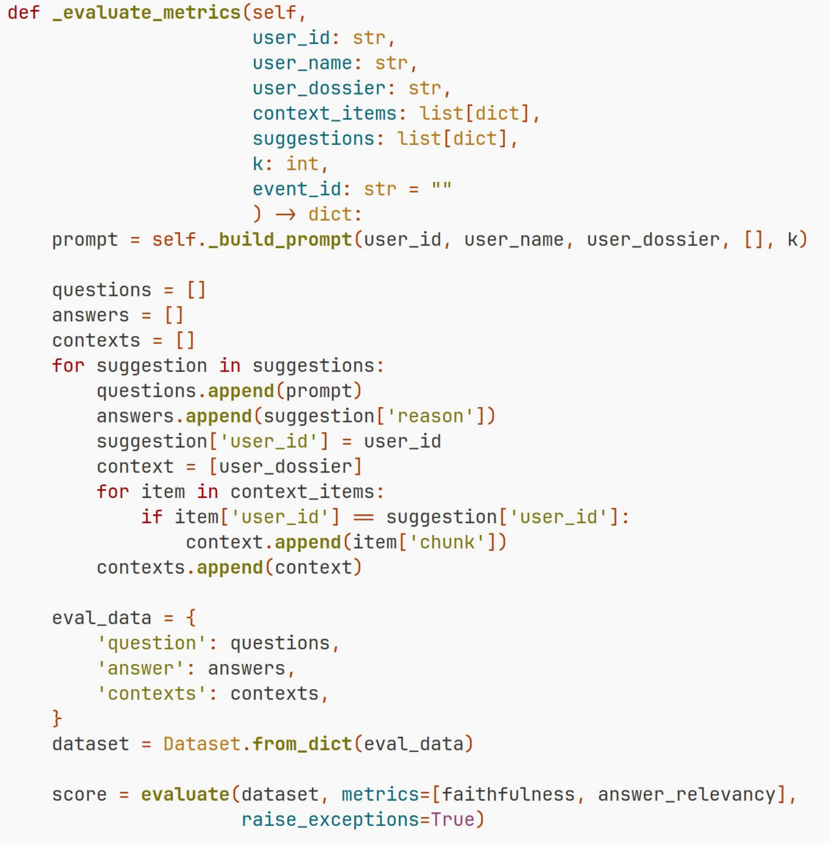
Respuesta



```
[
  {
    "user_id": "462",
    "reason": "Shared background as founder of the Jane Goodall Institute and UN Messenger of Peace, with a focus on youth engagement through Roots & Shoots."
  },
  {
    "user_id": "314",
    "reason": "Interest in lunar-agriculture startups aligns with Dr. Goodall's focus on sustainability and climate literacy."
  },
  {
    "user_id": "9",
    "reason": "Expertise in self-governing micro-societies and resource-frugal hospitality complements Dr. Goodall's interests in closed-loop life-support systems."
  },
  {
    "user_id": "136",
    "reason": "Potential for creative collaboration on storytelling related to lunar experiences, which could enhance Dr. Goodall's advocacy efforts."
  },
  {
    "user_id": "219",
    "reason": "Insights on teamwork in extreme environments could inform Dr. Goodall's conservation efforts and educational initiatives."
  },
  {
    "user_id": "321",
    "reason": "Experience in ethical extraction and blockchain aligns with Dr. Goodall's interest in innovative tracking for reforestation efforts."
  },
  {
    "user_id": "545",
    "reason": "Focus on equity and multi-generational stewardship resonates with Dr. Goodall's guiding principles."
  },
  {
    "user_id": "66",
    "reason": "Interest in climate-positive initiatives and children's education aligns with Dr. Goodall's Roots & Shoots program."
  },
  {
    "user_id": "83",
    "reason": "Shared fascination with lunar themes and mythology, which could foster a creative dialogue on conservation and space."
  },
  {
    "user_id": "149",
    "reason": "Potential for engaging discussions on technical problem-solving related to space and sustainability."
  }
]
```

Métricas

Faithfulness y Relevance



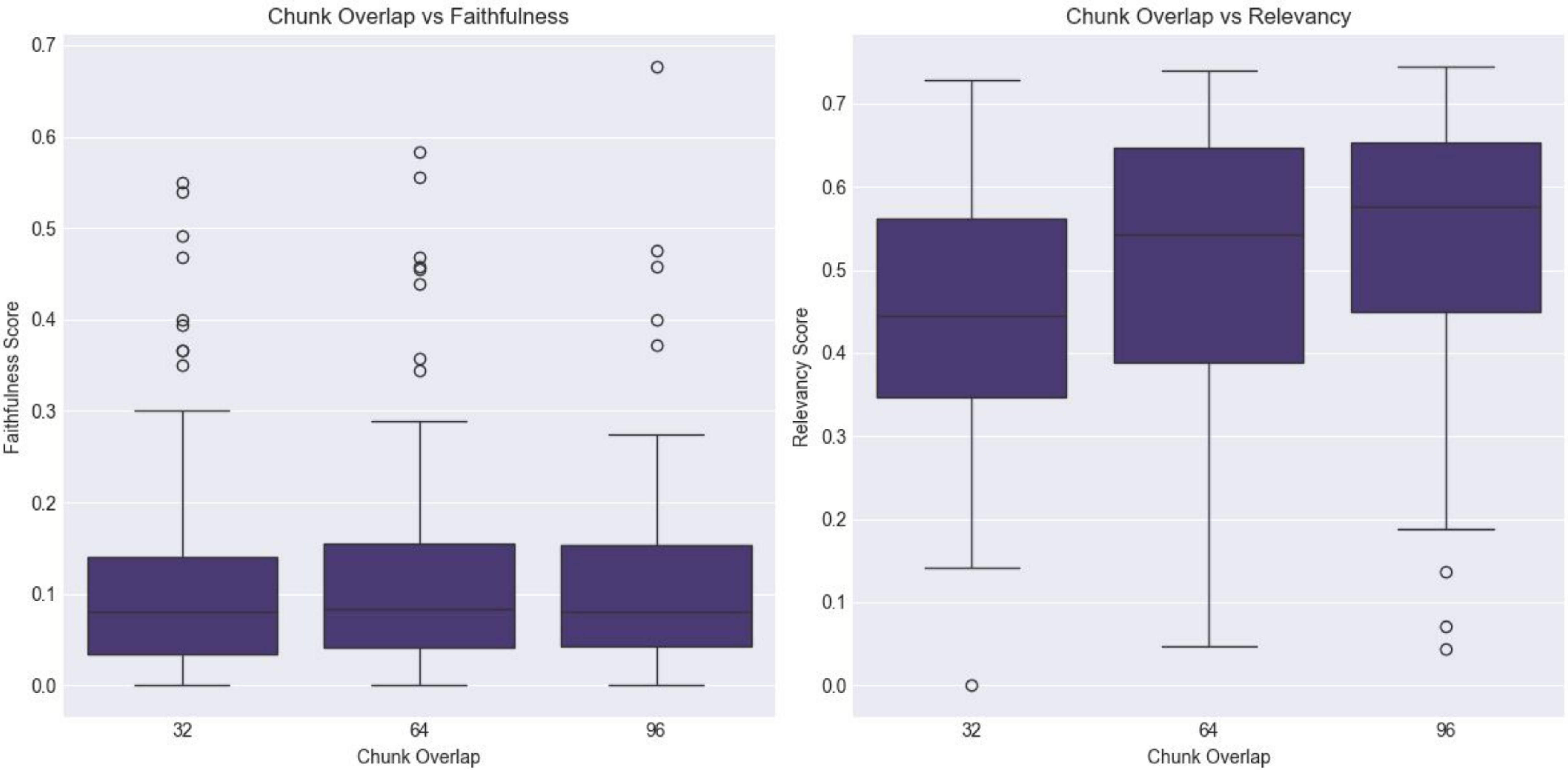
```python
def _evaluate_metrics(self,
                      user_id: str,
                      user_name: str,
                      user_dossier: str,
                      context_items: list[dict],
                      suggestions: list[dict],
                      k: int,
                      event_id: str = ""
                      ) -> dict:
    prompt = self._build_prompt(user_id, user_name, user_dossier, [], k)

    questions = []
    answers = []
    contexts = []
    for suggestion in suggestions:
        questions.append(prompt)
        answers.append(suggestion['reason'])
        suggestion['user_id'] = user_id
        context = [user_dossier]
        for item in context_items:
            if item['user_id'] == suggestion['user_id']:
                context.append(item['chunk'])
        contexts.append(context)

    eval_data = {
        'question': questions,
        'answer': answers,
        'contexts': contexts,
    }
    dataset = Dataset.from_dict(eval_data)

    score = evaluate(dataset, metrics=[faithfulness, answer_relevancy],
                     raise_exceptions=True)
```

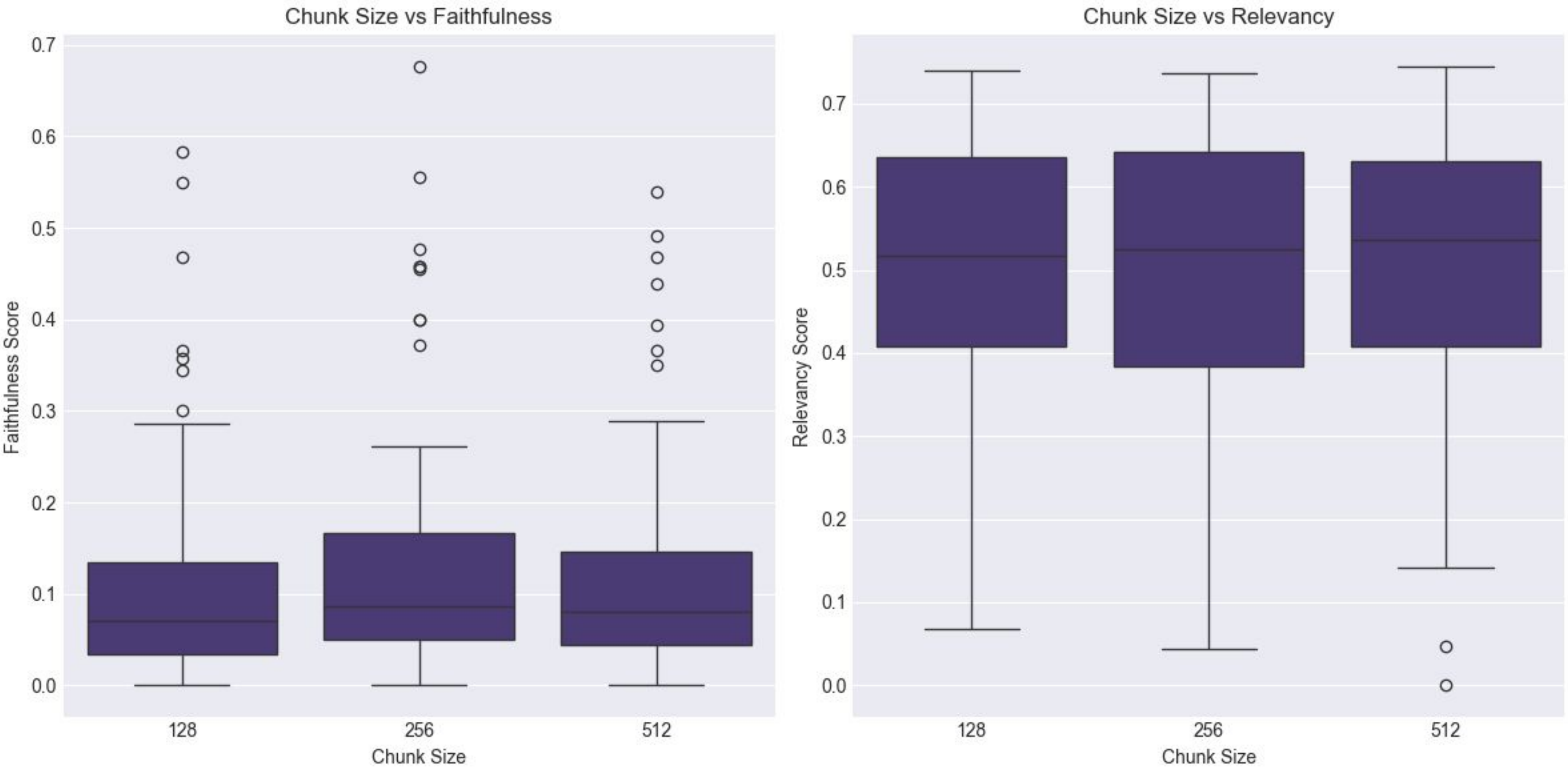
- Framework: Ragas

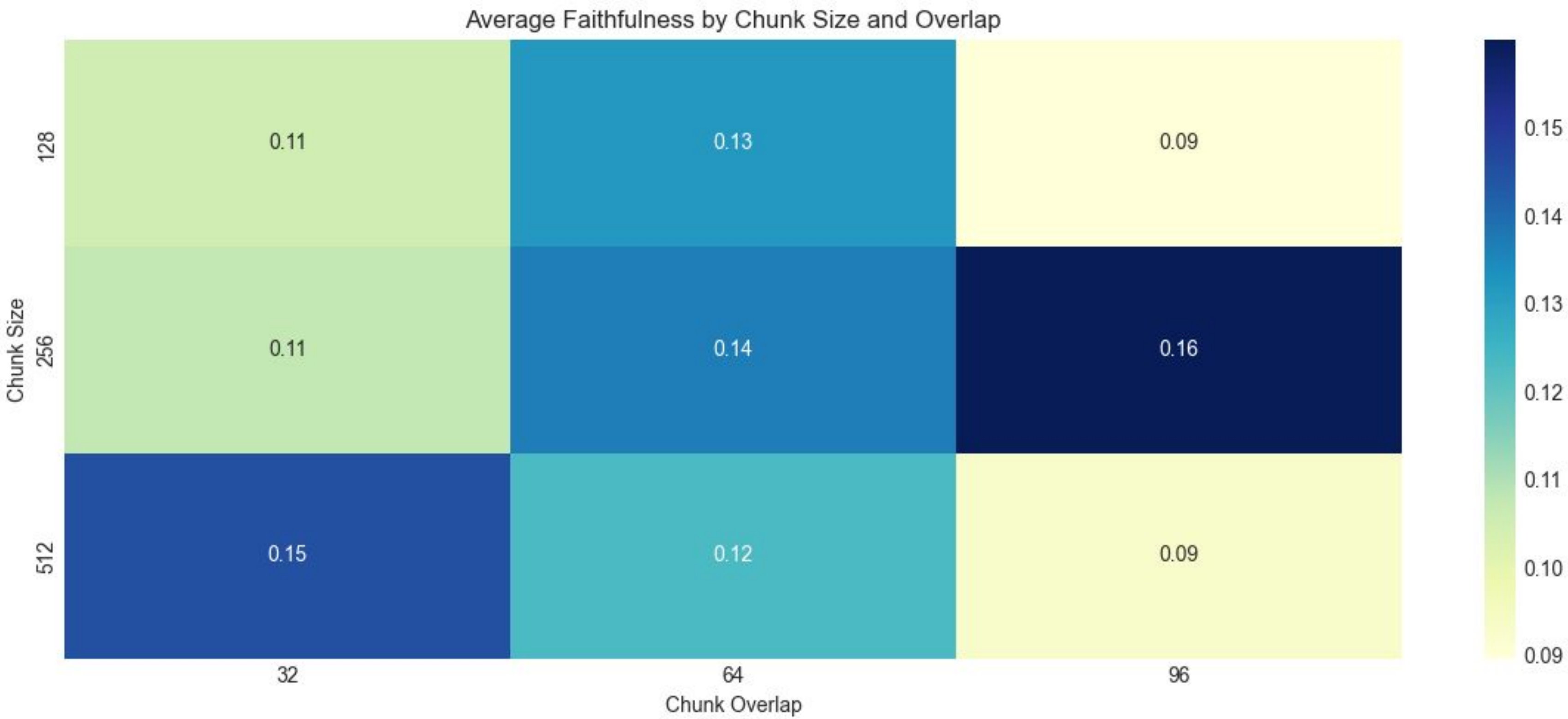
- Decidimos medir Faithfulness y Relevance con diferentes:
    - chunk sizes
    - chunk overlap
    - k

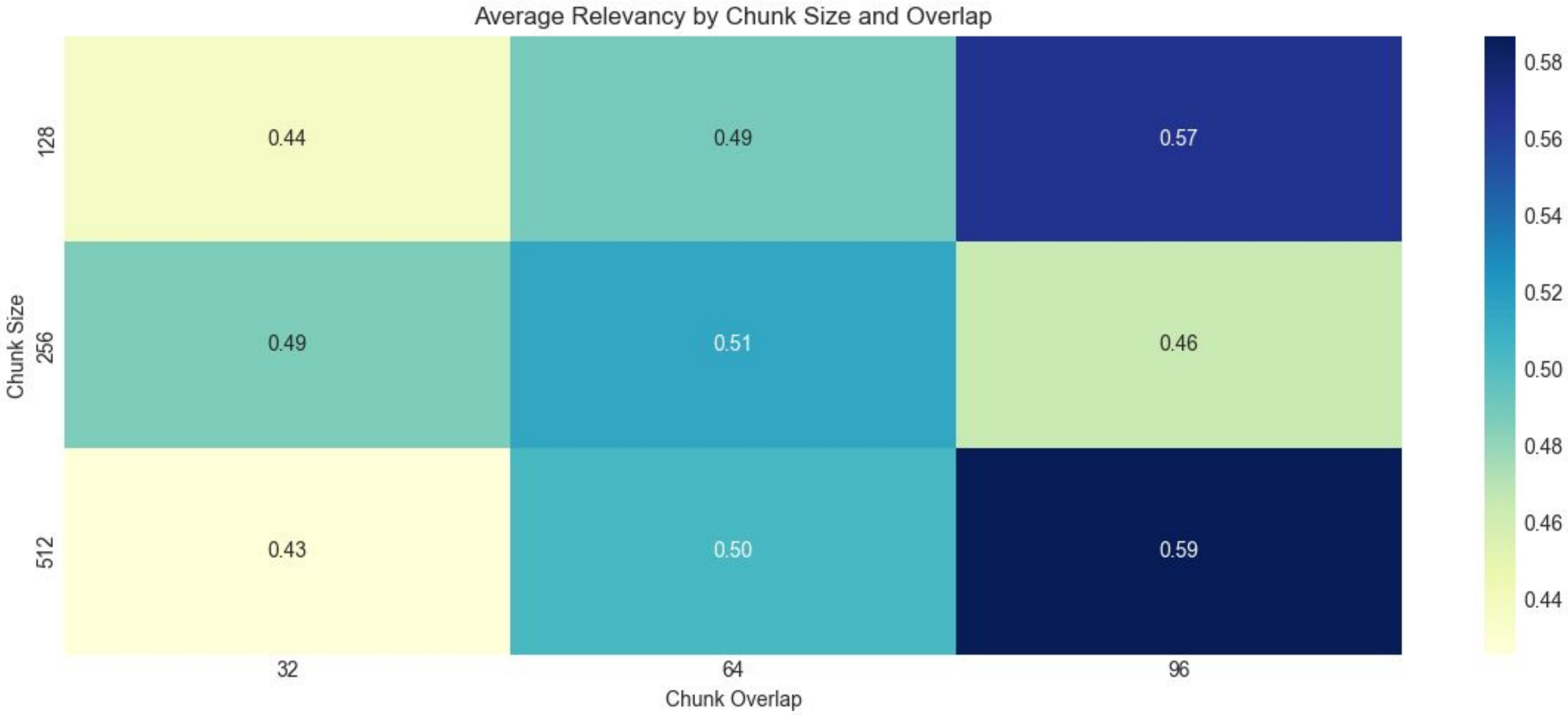
- Faithfulness:
    - No queremos razones de conexiones inventadas
    - **Dataset** no ayudó a esta métrica
- Relevance:
    - Conexiones de buena calidad

- Las *questions* son el prompt (sin los chunks)
- Las sugerencias (JSON) se pasan a string
- El *context* son los chunks relacionados con el usuario con quien se realiza la conexión
- Cada sugerencia se considera como *answer* diferente


Chunk Overlap vs Faithfulness
Faithfulness Score
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
32
64
96
Chunk Overlap
Chunk Overlap vs Relevancy
Relevancy Score
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
32
64
96
Chunk Overlap


Chunk Size vs Faithfulness
Faithfulness Score
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
128
256
512
Chunk Size
Chunk Size vs Relevancy
Relevancy Score
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
128
256
512
Chunk Size


Average Faithfulness by Chunk Size and Overlap



Average Relevancy by Chunk Size and Overlap
Chunk Size
128
256
512
Chunk Overlap
32
64
96
0.44
0.49
0.57
0.49
0.51
0.46
0.43
0.50
0.59
0.58
0.56
0.54
0.52
0.50
0.48
0.46
0.44

Faithfulness vs. Relevancy Tradeoff

- Correlación negativa entre *Faithfulness* y *Relevancy*:
  - *Faithfulness*: *chunks* más pequeños, menos *overlap*.
  - *Relevancy*: *chunks* más grandes, mayor *overlap*.

- *Faithfulness*
  - El LLM agrega palabras que afectan a esta métrica
  - Puede que esto ocurra por que el Dataset es de perfiles conocidos
  - Para casos donde los perfiles no son públicos, puede que esta métrica mejore

Conclusiones

Conclusiones

- **RAG es la mejor opción para hacer conexiones con LLMs en eventos grandes,** permite escalar sin perder calidad, algo que el enfoque tradicional no logra.
- En eventos masivos enviar todos los perfiles en un solo prompt se vuelve inviable por límites de contexto y caída de rendimiento.
- Incluso con modelos de context window ampliado, el costo crece rápidamente hay más latencia por cada sugerencia.
- **RAG mejora el foco del modelo al enviar sólo los fragmentos más relevantes eso reduce distracción,** mejora la precisión y genera mejores razones de conexión.

Gracias