

RECUPERAÇÃO MULTIMODAL DE MATERIAIS DIDÁTICOS COM GEMINI EMBEDDING 2: UMA AVALIAÇÃO NO ENSINO SUPERIOR

Modalidade: Artigo Completo

Kawan Antonio Machado*;
Lucas Senzaki;

* kawan.machado@alunos.bpkedu.com.br

RESUMO: Este trabalho avalia o Gemini Embedding 2, que representa texto, imagem, vídeo e áudio em um mesmo espaço vetorial, na recuperação cross-modal de materiais da disciplina de Modelagem Matemática e Otimização da faculdade Donaduzzi. O corpus reúne slides da disciplina e vídeos e áudios de outras fontes (250 itens), com 33 consultas textuais e uma linha de base por transcrição. As abordagens não diferiram significativamente (nDCG@10 de 0,804 e 0,832). O modelo aproximou slides e vídeos do mesmo conceito cerca de três vezes acima do acaso, mas organizou-se principalmente por formato e fonte. Em páginas de texto denso, o texto foi superior, o que sugere indexação híbrida.

Palavras-chave: busca semântica; aprendizado de representações; tecnologia educacional; inteligência artificial na educação; transcrição automática de fala.

1 INTRODUÇÃO

A diversificação dos materiais didáticos, que combinam textos, imagens, slides, vídeos e áudios, exige sistemas de recuperação de informação capazes de lidar com diferentes modalidades. Os métodos tradicionais, baseados em consultas textuais ou em transcrições, descartam a informação visual e sonora, enfraquecem o alinhamento entre o que é exibido nos slides e o que é dito pelo professor e têm dificuldade com linguagem técnica e sequências longas (LEE et al., 2022; LI et al., 2025).

Em materiais com pouco texto, como slides e quadros de vídeo, a representação puramente textual fornece pistas semânticas insuficientes e reduz a precisão da recuperação. O problema é crítico porque falhas nessa etapa não são corrigidas pelas fases posteriores de um sistema de busca ou de geração aumentada por recuperação (LEE et al., 2022; CIRILLO et al., 2026).

Estudos recentes têm explorado modelos de *embeddings* multimodais baseados em inteligência artificial generativa, que alinham conteúdos de diferentes modalidades. Destaca-se o Gemini Embedding 2, da Google DeepMind, que representa texto, imagem, vídeo e áudio em um espaço vetorial unificado e alcança desempenho estado da arte em recuperação unimodal, cross-modal e multimodal, com desempenho competitivo frente a alternativas de código aberto (SHANBHOGUE et al., 2026; CIRILLO et al., 2026).

Na educação, a integração de modalidades já reduziu alucinações e aumentou a acurácia em relação a abordagens textuais, em soluções como grafos de conhecimento disciplinar, segmentação de gravações de aula e detecção de elementos em slides (ZHAO; WANG; LU, 2025; SAPENA; ONAINDIA, 2026; MANIYAR et al., 2025). Ainda assim, os *benchmarks* usados para avaliar esses modelos são genéricos, e o uso de modelos multimodais na educação ainda é emergente (KÜCHEMANN et al., 2025).

Diante disso, este trabalho avalia a aplicabilidade do Gemini Embedding 2 na recuperação multimodal de materiais didáticos da disciplina de Modelagem Matemática e Otimização da faculdade Donaduzzi. O corpus combina os slides da disciplina com vídeos de aula e áudios de outras fontes sobre os mesmos conceitos. Nos testes de recuperação cross-modal, uma consulta em uma modalidade recupera conteúdos em outra (imagens de slides, segmentos de vídeo ou de áudio).

Os pares consulta:conteúdo relevantes são anotados como referência (*ground truth*), e o desempenho é medido por Recall@k e nDCG@k. O objetivo é verificar se o modelo aproxima conteúdos pelo conceito que abordam, independentemente da fonte e do formato, e, com isso, contribuir para sistemas de busca mais integrados.

2 REFERENCIAL TEÓRICO

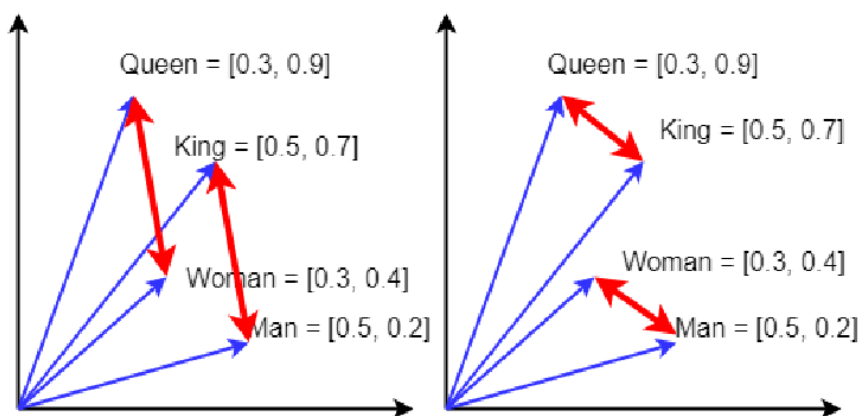
Os conteúdos didáticos podem ser representados em vários formatos, como slides, vídeos, áudios e textos, que raramente são tratados de forma integrada por mecanismos de busca tradicionais. Modelos de *embeddings* multimodais permitem representá-los em um mesmo espaço vetorial, aproximando conteúdos semanticamente equivalentes ainda que venham de fontes distintas. Esta seção apresenta os fundamentos dessa abordagem, a recuperação multimodal tradicional e o Gemini Embedding 2 (SHANBHOGUE et al., 2026).

2.1 *Embeddings* e espaços vetoriais

Para que um computador processe texto, imagem ou áudio, é necessário transformar esses dados em representações numéricas (RUSSELL; NORVIG, 2021). Um *embedding* é uma representação vetorial de um objeto (palavra, frase, imagem, vídeo) em um espaço de alta dimensão, na qual a posição do vetor codifica informações semânticas. Objetos semelhantes ficam em regiões próximas do espaço, e objetos diferentes, afastados (MIKOLOV et al., 2013; DEVLIN et al., 2019).

Formalmente, um *embedding* é um vetor $\mathbf{v} \in \mathbb{R}^d$, em que d é a dimensão do espaço, tipicamente entre 768 e 3072 em modelos modernos. A similaridade entre dois objetos pode ser medida pelo cosseno do ângulo entre seus vetores. A Figura 1 ilustra um exemplo clássico, em que $\mathbf{E}(\text{queen}) \approx \mathbf{E}(\text{king}) + \mathbf{E}(\text{woman}) - \mathbf{E}(\text{man})$ (MIKOLOV et al., 2013; PENNINGTON; SOCHER; MANNING, 2014).

Figura 1 – Espaço de *embeddings* arbitrário para as palavras *king* (rei), *queen* (rainha), *man* (homem) e *woman* (mulher), mostrando que as diferenças vetoriais dos termos masculinos em relação aos femininos são semelhantes



Fonte: adaptado de Sutor Junior et al. (2019).

A qualidade de um *embedding* depende de como ele foi treinado. Word2Vec e GloVe geram *embeddings* estáticos, com um único vetor por palavra, independente do contexto. Modelos contextualizados, como o BERT, atribuem à mesma palavra vetores diferentes conforme a frase, o que permite distinguir “manga” (fruta) de “manga” (parte da camisa), capacidade importante para a recuperação de informação (MIKOLOV et al., 2013; PENNINGTON; SOCHER; MANNING, 2014; DEVLIN et al., 2019; VIEIRA, 2026).

2.2 Recuperação multimodal tradicional

A recuperação multimodal tradicional usa codificadores separados para cada modalidade. Modelos como CLIP e ALIGN treinam, de forma contrastiva, um codificador de texto e outro de imagem para que pares correspondentes fiquem próximos no espaço vetorial. Essa abordagem, chamada *late fusion*, é eficiente entre duas modalidades, mas não lida bem com entradas intercaladas nem explora as interações entre modalidades (RADFORD et al., 2021; JIA et al., 2021; SHANBHOGUE et al., 2026).

Na educação, há desafios adicionais. O *dataset* MLP mostra que slides de aula misturam texto, diagramas, equações e imagens, e que o alinhamento entre a fala do professor e o conteúdo escrito é fraco. Modelos cross-modais como PVSE e PCME foram propostos para esse alinhamento fraco, mas ficam muito abaixo do desempenho de estudantes na recuperação entre figuras e textos (LEE et al., 2022; SONG; SOLEYMANI, 2019; CHUN et al., 2021).

Sistemas de geração aumentada por recuperação (RAG) aplicados à educação dependem de uma recuperação eficaz para evitar alucinações. Ao integrar texto, slides e vídeos em um grafo de conhecimento disciplinar, o MDKAG reduziu a alucinação em até 23% e aumentou a acurácia em 11% (ZHAO; WANG; LU, 2025). Esses sistemas, porém, ainda dependem de transcrição (ASR) e de reconhecimento óptico de caracteres (OCR), que introduzem erros e perdem informações não textuais (ZHAO; WANG; LU, 2025; SHANBHOGUE et al., 2026).

2.3 O Gemini Embedding 2

Em maio de 2026, o Google DeepMind anunciou o Gemini Embedding 2, modelo de *embedding* multimodal nativo que representa texto, imagem, vídeo e áudio em um único espaço vetorial. Diferentemente dos modelos de *late fusion*, ele é inicializado a partir do Gemini e ajustado com treinamento multitarefa em múltiplos estágios, o que captura as interações entre modalidades e permite gerar um único *embedding* para entradas intercaladas (SHANBHOGUE et al., 2026).

A principal inovação é dispensar a conversão para modalidades intermediárias: o modelo processa áudio e vídeo diretamente, sem transcrição ou legendas. Os autores relatam que, no *benchmark* MSEB, o áudio nativo alcançou MRR@10 de 73,99, contra 70,40 da abordagem com ASR, com ganho de 5,01 pontos entre línguas. Segundo eles, a representação direta preserva informações acústicas perdidas na transcrição, como ênfase e ambiguidades fonéticas (SHANBHOGUE et al., 2026).

O modelo alcança desempenho estado da arte em *benchmarks* como MSCOCO (62,9 R@1), Vatex (68,8 NDCG@10), MTEB Multilingual (69,9) e MTEB Code (84,0) (SHANBHOGUE et al., 2026). Uma avaliação independente confirma sua superioridade em recuperação densa frente a modelos de código aberto, com latência maior (231,6 ms contra 31 ms), que afeta apenas a codificação das consultas, já que os documentos são codificados uma única vez, na indexação (CIRILLO et al., 2026).

2.4 Aplicações educacionais e a lacuna de pesquisa

Na educação, representar diferentes modalidades em um mesmo espaço vetorial permite recuperar um slide a partir de uma consulta em áudio, ou um trecho de vídeo a partir de uma consulta textual, sem depender de transcrições ou legendas que introduzem erros. Isso é especialmente relevante para materiais com pouco texto, como slides e quadros de vídeo (SHANBHOGUE et al., 2026; LEE et al., 2022).

Apesar desse potencial, os *benchmarks* usados para avaliar o Gemini Embedding 2 são genéricos (MSCOCO, Vatex, MTEB, BEIR) e não incluem material didático real. O *dataset* MLP é uma exceção, mas contém apenas slides em inglês e não avalia modelos nativos. Faltam, portanto, avaliações desses modelos em cenários educacionais concretos, especialmente no ensino superior brasileiro (SHANBHOGUE et al., 2026; CIRILLO et al., 2026; LEE et al., 2022; KÜCHEMANN et al., 2025).

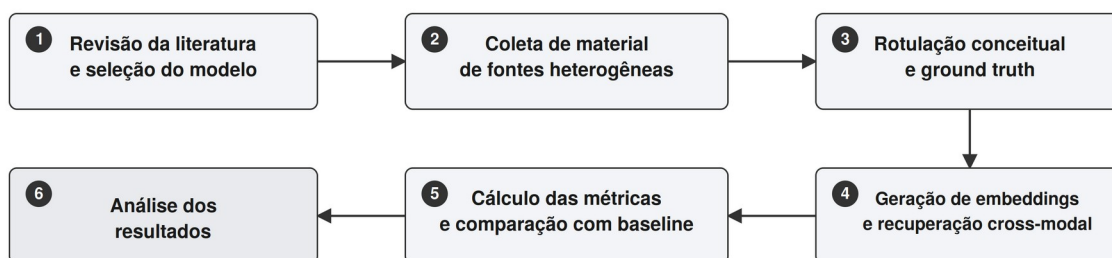
3 PROCEDIMENTOS METODOLÓGICOS

Este trabalho combinou pesquisa bibliográfica, por meio de uma revisão sistemática da literatura (RSL), e pesquisa experimental, que avaliou a aplicabilidade do Gemini Embedding 2 à recuperação cross-modal de materiais didáticos no ensino superior. A questão central foi se o modelo organiza, em um mesmo espaço vetorial, conteúdos de uma disciplina vindos de fontes, autores e formatos distintos, aproximando-os pelo conceito, e não pela origem ou modalidade.

Adotou-se a hipótese de que, se o modelo representa o conteúdo por sua semântica, itens que tratam do mesmo conceito permanecem próximos no espaço vetorial, ainda que sejam um slide, um segmento de vídeo ou um trecho de áudio, e independentemente de quem os produziu.

O trabalho foi organizado em seis etapas, apresentadas na Figura 2: (i) revisão sistemática da literatura e seleção do modelo; (ii) coleta de material de fontes heterogêneas; (iii) rotulação conceitual e *ground truth*; (iv) geração de *embeddings* e recuperação cross-modal; (v) cálculo das métricas e comparação com a linha de base; e (vi) análise dos resultados.

Figura 2 – Etapas metodológicas da pesquisa



Fonte: elaborado pelos autores (2026).

Na etapa (i), a RSL fundamentou o estudo e justificou a escolha do Gemini Embedding 2. Na (ii), reuniram-se os slides da disciplina e vídeos e áudios de terceiros sobre os mesmos tópicos. Na (iii), cada item recebeu um rótulo conceitual, que definiu os itens relevantes para cada consulta. Na (iv), os itens foram codificados na mídia nativa e no texto, e as consultas foram executadas por similaridade do cosseno.

Na etapa (v), calcularam-se $\text{Recall}@k$ e $\text{nDCG}@k$ nas duas condições e testaram-se as diferenças. Na (vi), examinaram-se o desempenho por formato e a vizinhança dos itens no espaço vetorial, para verificar a hipótese.

3.1 Revisão sistemática da literatura

A RSL seguiu um protocolo definido para tornar a seleção dos estudos transparente e reproduzível, orientada pela questão: como modelos de *embedding* multimodais têm sido aplicados e avaliados na recuperação de informação, em especial de materiais didáticos? A busca foi feita no arXiv e no Google Acadêmico, organizada em três eixos, cada um com sua expressão de busca:

E1: "Gemini Embedding 2" OR "Gemini Embedding"

E2: ("multimodal" OR "cross-modal") AND ("retrieval" OR "embedding" OR "foundation model") AND ("education" OR "lecture" OR "slides")

E3: "multimodal information retrieval" OR "multimodal document retrieval"

O eixo E1 reúne trabalhos que apresentam ou avaliam o modelo estudado; o E2, estudos sobre modelos multimodais na educação; e o E3, *frameworks* de recuperação multimodal de documentos. Foram incluídos estudos de 2022 a 2026, em inglês ou português, com texto completo acessível e revisados por pares ou publicados como *preprints* em repositórios reconhecidos, que atendessem a pelo menos um eixo.

Foram excluídos duplicados, textos em outros idiomas e trabalhos sem relação com recuperação de informação ou com modelos multimodais, exceto avaliações do próprio Gemini Embedding 2. A seleção ocorreu em três fases: leitura de títulos e resumos dos resultados de cada eixo, leitura completa dos estudos pré-selecionados e aplicação dos critérios de inclusão e exclusão.

O processo resultou em oito estudos: dois no eixo E1 (SHANBHOUE et al., 2026; CIRILLO et al., 2026), cinco no E2 (LEE et al., 2022; ZHAO; WANG; LU, 2025; SAPENA; ONAINDIA, 2026; MANIYAR et al., 2025; KÜCHEMANN et al., 2025) e um no E3 (LI et al., 2025). O número de resultados em cada fase não foi registrado, o que limita a reprodutibilidade da seleção.

As referências desses estudos forneceram as obras seminais anteriores a 2022 citadas na fundamentação teórica. Um trabalho acadêmico sobre *embeddings* para o português brasileiro (VIEIRA, 2026) foi incluído à parte do protocolo, como apoio conceitual.

3.2 Materiais e preparação dos dados

O corpus foi construído a partir da disciplina de Modelagem Matemática e Otimização da faculdade Donaduzzi, que abrange cálculo numérico e pesquisa operacional. A base principal foram os slides da disciplina (96 páginas de cinco aulas), com formulações matemáticas, funções objetivo, restrições e representações gráficas.

Para as demais modalidades, foram reunidos materiais de terceiros sobre os mesmos tópicos: dez videoaulas de seis canais do YouTube, sobre programação linear, método simplex, dualidade, programação inteira e otimização não linear, e dois episódios do podcast P.O.Cast, sobre pesquisa operacional e simulação. O uso de fontes heterogêneas é parte do desenho experimental: interessa verificar se o modelo aproxima conteúdos equivalentes de autores, formatos e estilos diferentes.

Cada página de slide foi convertida em imagem (150 dpi). Os vídeos foram recodificados em até 480p e divididos em segmentos de até 110 s, e os áudios, em segmentos de até 170 s, conforme os limites da API. Vídeo e áudio foram codificados na modalidade original, sem conversão para texto; a transcrição, feita com o Whisper large-v3-turbo, serviu apenas à linha de base e à rotulação.

O corpus principal totalizou 250 itens (96 páginas de slide, 112 segmentos de vídeo e 42 de áudio), dos quais 240 têm ao menos 30 caracteres de texto. Para a análise de robustez, usou-se um corpus ampliado de 416 itens, que acrescenta apostilas, listas de exercícios e um *notebook* da disciplina, três artigos em português, 74 páginas de um livro-texto de métodos numéricos em inglês e três arquivos de código.

3.3 Rotulação, consultas e *ground truth*

Adotou-se uma taxonomia de 13 conceitos que cobre os tópicos do material (por exemplo, método da bissecção, interpolação polinomial, método simplex e dualidade), além da classe *outro*, para itens sem conceito identificável, como capas e trechos de conversa fora do tema. Esses itens permanecem no corpus como distratores e nunca são relevantes.

Os rótulos foram sugeridos pelo modelo *gemini-3.1-flash-lite* a partir da mídia e do texto de cada item. No corpus principal, todas as sugestões foram revisadas manualmente por um dos autores, com concordância de 89,6% (95,8% nos slides, 95,5% nos vídeos e 59,5% no podcast). No corpus ampliado, os itens adicionais mantiveram o rótulo automático ou o da seção de origem.

Foram elaboradas 39 consultas em texto, três por conceito, formuladas como perguntas de um aluno da disciplina. Um item foi considerado relevante quando seu rótulo coincide com o conceito da consulta (relevância binária), qualquer que seja a fonte ou o formato. Na análise principal, foram usadas 33 consultas: as de erros numéricos e de interpolação polinomial foram excluídas por não haver itens desses conceitos no corpus principal.

Esse critério mede a recuperação no nível do tema: um item é relevante se trata do conceito da consulta, ainda que não responda à pergunta formulada. Ele foi adotado porque a hipótese diz respeito ao agrupamento por conceito entre formatos, mas não avalia se o conteúdo recuperado atende à necessidade informacional específica de cada consulta.

3.4 Tecnologias e avaliação

A avaliação centrou-se no Gemini Embedding 2 (SHANBHOGUE et al., 2026), com vetores de 768 dimensões. Cada item foi codificado em uma chamada própria à API, com os prefixos de tarefa recomendados pela documentação, e a recuperação usou a similaridade do cosseno. A implementação foi feita em Python 3.12, com a biblioteca *google-genai*, o *ffmpeg* para a segmentação e o *mlx-whisper* para a transcrição, em execução local.

Compararam-se duas condições: na cross-modal, cada item foi representado pelo vetor da sua mídia nativa; na linha de base textual, pelo vetor do seu texto (texto do slide ou transcrição). No cenário A, com todos os itens, os que não têm texto foram postos no fim do *ranking* da linha de base, já que um sistema textual não os recupera; no cenário B, de controle, esses itens foram removidos das duas condições.

As métricas principais foram Recall@k e nDCG@k, para $k \in \{1, 3, 5, 10\}$, e as complementares, Precision@k, MRR e MAP. As condições foram comparadas consulta a consulta pelo teste de Wilcoxon pareado bicaudal, com intervalo de confiança de 95% da diferença média por *bootstrap* (10.000 reamostragens) e correção de Holm sobre os 28 testes, com $\alpha = 0,05$.

Para verificar se o espaço vetorial se organiza por conceito, e não por formato ou fonte, calculou-se a fração dos dez vizinhos mais próximos de cada item que compartilha o conceito, o formato ou a fonte, comparada ao esperado ao acaso, com significância por permutação (1.000 permutações). A análise foi repetida com vizinhos restritos a outro formato. Uma condição híbrida exploratória usou, para cada item, a maior similaridade entre o vetor de mídia e o de texto.

Uma execução exploratória, com o corpus ampliado e rótulos automáticos, antecedeu a análise principal. O plano da análise principal, restrita ao corpus de slides, vídeos e podcast, foi registrado antes do cálculo das métricas finais, sem alteração das consultas, dos valores de k ou dos parâmetros do modelo. O corpus ampliado é reportado como análise de robustez.

3.5 Aspectos éticos

A pesquisa usou apenas materiais didáticos, sem coleta de dados, questionários ou testes com pessoas, o que dispensou a submissão a Comitê de Ética em Pesquisa. O material foi empregado com finalidade acadêmica, preservados os créditos de autoria.

4 ANÁLISE DOS RESULTADOS

Os resultados seguem a proposta da introdução: primeiro, o desempenho da recuperação cross-modal por Recall@ k e nDCG@ k ; em seguida, a recuperação em cada formato de material; por fim, a verificação de se o modelo aproxima os conteúdos pelo conceito, independentemente da fonte e do formato. A análise principal usou 250 itens e 33 consultas; o corpus ampliado aparece ao final, como robustez.

4.1 Desempenho da recuperação cross-modal

A Tabela 1 apresenta Recall@ k e nDCG@ k no cenário A. Nos dez primeiros resultados, as abordagens tiveram desempenho próximo: nDCG@10 de 0,804 no cross-modal e 0,832 na linha de base, com 13 consultas favoráveis a cada uma e sete empates, e Recall@10 de 0,487 e 0,475. Nenhuma comparação apresentou diferença significativa após a correção de Holm.

Nas primeiras posições, a linha de base teve valores mais altos (nDCG@1 de 0,970 contra 0,818), e o intervalo de confiança da diferença exclui o zero em nDCG@3 e nDCG@5. O teste pareado, porém, não indica diferença, pois a maioria das consultas empata e a média depende de poucas delas (no nDCG@3, quatro favoráveis ao cross-modal e nove à linha de base).

Tabela 1 – Recall@ k e nDCG@ k na análise principal (250 itens, 33 consultas, cenário A)

Métrica	Cross-modal	Linha de base	Diferença [IC 95%]	p (Holm)
Recall@1	0,059	0,071	-0,012 [-0,029; 0,003]	1,00
Recall@3	0,166	0,209	-0,042 [-0,086; -0,007]	1,00
Recall@5	0,269	0,311	-0,042 [-0,097; 0,003]	1,00
Recall@10	0,487	0,475	0,012 [-0,050; 0,077]	1,00
nDCG@1	0,818	0,970	-0,152 [-0,303; 0,000]	1,00
nDCG@3	0,818	0,955	-0,137 [-0,251; -0,032]	1,00
nDCG@5	0,805	0,901	-0,095 [-0,193; -0,006]	1,00
nDCG@10	0,804	0,832	-0,029 [-0,112; 0,051]	1,00

Fonte: dados da pesquisa (2026). Nota: diferença = cross-modal menos linha de base, com IC 95% por *bootstrap*; p corrigido por Holm.

As métricas complementares seguiram o mesmo padrão (MRR de 0,878 e 0,985; MAP de 0,706 e 0,683), e o cenário B, restrito a itens com texto, reproduziu esses resultados (nDCG@10 de 0,803 contra 0,836). Por conceito, em caráter descritivo, o cross-modal teve nDCG@10 maior em

isolamento de raízes, programação inteira, método simplex e diferenciação numérica, e menor em bissecção, Newton-Raphson e pesquisa operacional.

4.2 Recuperação por formato de material

A Tabela 2 mostra a participação de cada formato no top-10 e os itens relevantes recuperados. No cross-modal, a participação acompanhou a proporção no corpus (38,2% de slides, 46,1% de vídeo e 15,8% de podcast, contra 38,4%, 44,8% e 16,8%). Na linha de base, o podcast ficou sub-representado (9,4%). O cross-modal recuperou mais relevantes em vídeo (126 contra 115) e podcast (23 contra 21), e a linha de base, em slides (103 contra 98).

Tabela 2 – Participação de cada formato no top-10 e itens relevantes recuperados no top-10 (corpus principal, cenário A)

Formato	Corpus	Participação no top-10		Acertos no top-10	
		Cross-modal	Linha de base	Cross-modal	Linha de base
Slide	38,4%	38,2%	37,9%	98	103
Vídeo	44,8%	46,1%	52,7%	126	115
Podcast	16,8%	15,8%	9,4%	23	21

Fonte: dados da pesquisa (2026).

4.3 Aproximação por conceito, fonte e formato

A Tabela 3 verifica se os conteúdos se aproximam pelo conceito no espaço vetorial. Os dez vizinhos mais próximos de cada item compartilham o conceito sete vezes mais que o esperado ao acaso (80,2% contra 11,3%), mas compartilham ainda mais o formato (96,1%) e a fonte (91,4%). Como trechos consecutivos de um vídeo ou episódio tratam do mesmo conceito, parte dessa concordância vem da proximidade entre itens da mesma fonte.

Com os vizinhos restritos a outro formato, a fração que compartilha o conceito fica em torno de 20%, cerca de três vezes acima do acaso, entre slides e vídeos, nos dois sentidos, e do podcast para os slides. O par vídeo e podcast não pôde ser avaliado, pois não há conceitos em comum entre eles, e o par slide e podcast tem sobreposição mínima (2,2% esperado ao acaso).

Tabela 3 – Fração dos dez vizinhos mais próximos que compartilham cada atributo no espaço vetorial, comparada ao esperado ao acaso

Âncora	Vizinhos	Observado	Acaso	p
<i>Todos os itens</i>				
Mesmo conceito	todos	0,802	0,113	0,001
Mesmo formato	todos	0,961	0,374	0,001
Mesma fonte	todos	0,914	0,218	0,001
<i>Mesmo conceito, vizinhos de outro formato</i>				
Slide	vídeo	0,223	0,078	0,001
Vídeo	slide	0,234	0,077	0,001
Podcast	slide	0,178	0,051	0,011
Slide	podcast	0,024	0,022	1,000

Fonte: dados da pesquisa (2026). Nota: p obtido por teste de permutação.

4.4 Análise de robustez

No corpus ampliado (416 itens, 39 consultas), a linha de base foi superior após a correção de Holm em Recall@3, @5 e @10 e em nDCG@3, @5 e @10 (nDCG@10 de 0,712 contra 0,853; diferença de -0,141, IC 95% [-0,221; -0,064]). A diferença concentra-se nas páginas de texto denso: no top-10, as páginas do livro em inglês somaram 41 acertos na linha de base e nenhum no cross-modal, e as dos artigos, 28 e 1.

Nos slides, vídeos e podcast, o cross-modal manteve mais acertos (83, 124 e 23, contra 59, 106 e 21). A condição híbrida exploratória obteve nDCG@10 de 0,846 no corpus principal; no ampliado, coincidiu com a linha de base, porque a similaridade entre textos foi sempre maior que a similaridade entre texto e mídia.

4.5 Discussão

O primeiro ponto proposto era avaliar a aplicabilidade do Gemini Embedding 2 à recuperação cross-modal de materiais didáticos. No corpus da disciplina, a recuperação a partir de slides, vídeos e áudios alcançou nDCG@10 de 0,804, estatisticamente indistinguível do obtido com a transcrição dos mesmos conteúdos. Nesses termos, o modelo mostrou-se aplicável à recuperação temática: recupera material audiovisual do conceito buscado em nível comparável ao de um sistema textual, sem exigir transcrição.

O resultado não indica, contudo, ganho da representação multimodal sobre a textual, e o número de consultas não permite afirmar equivalência. A vantagem da representação nativa relatada por Shanbhogue et al. (2026) em *benchmarks* genéricos não se reproduziu aqui. Uma explicação plausível é a qualidade das transcrições: aulas com fala clara, transcritas por um modelo recente, preservam o vocabulário técnico das consultas, o que favorece a linha de base no topo.

O objetivo central era verificar se o modelo aproxima os conteúdos pelo conceito, independentemente da fonte e do formato. A hipótese previa que itens do mesmo conceito permaneceriam próximos no espaço vetorial, qualquer que fosse o formato ou o autor. Na recuperação, o resultado é favorável: as consultas recuperaram slides, vídeos e áudios na proporção em que aparecem no corpus, enquanto a linha de base sub-representou o podcast.

Na organização do espaço vetorial, o resultado é parcial. Há aproximação conceitual entre formatos, cerca de três vezes acima do acaso, mas ela atinge apenas cerca de 20% dos vizinhos, enquanto 96% compartilham o formato. A parte da hipótese que exige independência da fonte e do formato, central no estudo, não se confirmou: esses dois atributos pesam mais na vizinhança que o conceito.

A interpretação exige cautela, porque neste corpus formato e fonte estão confundidos: todos os slides vêm da disciplina, os vídeos de canais do YouTube e os áudios de um único podcast. Não é possível, portanto, separar o efeito do formato do efeito da fonte. Como toda comparação entre formatos é também entre fontes, o alinhamento observado indica que o sinal conceitual resiste à troca simultânea de ambos, ainda que enfraquecido.

Dois mecanismos podem explicar o peso do formato e da fonte. O primeiro é a separação entre modalidades no espaço vetorial: a condição híbrida mostrou que as similaridades entre texto e texto são sempre maiores que entre texto e mídia. O segundo são marcas próprias de cada fonte, como a voz do professor, o leiaute dos slides e a continuidade entre trechos de um mesmo vídeo, que aproximam itens independentemente do conceito.

Para a busca, isso implica que, em um corpus misto, a ordenação pode favorecer o formato da consulta ou da fonte mais frequente. Neste estudo, as consultas textuais distribuíram os resultados

entre os formatos na proporção do corpus, mas o equilíbrio não é garantido em outros acervos. Separar os dois efeitos exige corpora em que um mesmo formato venha de várias fontes, como vídeos de canais diferentes sobre o mesmo conceito.

Quanto à contribuição para sistemas de busca integrados, os resultados indicam onde cada representação é mais adequada. Para slides, vídeos e áudio, a representação nativa indexa todo o material em um mesmo espaço, sem transcrição e com equilíbrio entre formatos. Para páginas de texto denso, sobretudo em outro idioma, o texto foi superior. Um sistema integrado deve combinar as duas, após normalizar as similaridades, que a condição híbrida mostrou estarem em escalas diferentes.

4.6 Limitações

A baixa concordância entre a sugestão automática e a revisão no podcast (59,5%) mostra que rotular áudio conversacional é especialmente difícil. A relevância binária por conceito trata como igualmente relevantes itens de profundidades diferentes e não verifica se o item responde à pergunta, o que pode favorecer itens apenas tematicamente próximos. As 33 consultas limitam o poder estatístico, e, como as consultas foram só textuais, a recuperação a partir de imagem ou áudio não foi avaliada.

Apenas 10 dos 250 itens não tinham texto, o que deixou pouco representados os materiais em que a literatura espera maior vantagem da representação nativa. O corpus é de uma única disciplina, formato e fonte estão confundidos, e vídeos e podcast não têm conceitos em comum. A segmentação por duração fixa pode dividir explicações ao meio, e a análise principal foi definida após uma execução exploratória, sem alterar consultas nem parâmetros.

5 CONSIDERAÇÕES FINAIS

Este trabalho avaliou o Gemini Embedding 2 na recuperação cross-modal de materiais da disciplina de Modelagem Matemática e Otimização, combinando os slides da disciplina com videoaulas e podcast de outras fontes. O modelo recuperou slides, vídeos e áudios com desempenho comparável ao da recuperação por transcrição, sem depender dela, e distribuiu os resultados de forma equilibrada entre os formatos.

A hipótese de que o modelo aproxima os conteúdos pelo conceito, independentemente da fonte e do formato, foi confirmada em parte: conteúdos do mesmo conceito se aproximaram entre formatos, mas o formato e a fonte ainda pesam mais na vizinhança. Em páginas de texto denso, a representação textual foi superior, o que indica que o modelo deve ser combinado ao texto em sistemas de busca integrados.

Essas conclusões valem para a recuperação temática, com consultas textuais, em uma única disciplina, e não estabelecem equivalência entre as abordagens nem indicam se o conteúdo recuperado responde a cada pergunta. Como trabalhos futuros, propõem-se julgamentos de relevância graduados por pergunta, corpora que separem formato e fonte, consultas em imagem e áudio, outras disciplinas e estratégias híbridas com normalização das similaridades.

REFERÊNCIAS

- CHUN, S. et al. Probabilistic embeddings for cross-modal retrieval. *In: IEEE/CVF CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION*, 2021. **Proceedings** [...]. [S. l.]: IEEE, 2021. p. 8415-8424.
- CIRILLO, S. et al. Benchmarking Google Embeddings 2 against open-source models for multilingual dense retrieval and RAG systems. **arXiv preprint**, arXiv:2605.23618, 2026.
- DEVLIN, J. et al. BERT: pre-training of deep bidirectional transformers for language understanding. *In: CONFERENCE OF THE NORTH AMERICAN CHAPTER OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS*, 2019. **Proceedings** [...]. [S. l.]: ACL, 2019. p. 4171-4186.
- JIA, C. et al. Scaling up visual and vision-language representation learning with noisy text supervision. *In: INTERNATIONAL CONFERENCE ON MACHINE LEARNING*, 2021. **Proceedings** [...]. [S. l.]: PMLR, 2021. p. 4904-4916.
- KÜCHEMANN, S. et al. On opportunities and challenges of large multimodal foundation models in education. **npj Science of Learning**, v. 10, n. 1, p. 11, 2025.
- LEE, D. W. et al. Multimodal lecture presentations dataset: understanding multimodality in educational slides. **arXiv preprint**, arXiv:2208.08080, 2022. Disponível em: <https://github.com/dondongwon/MLPDataset>.
- LI, Z. et al. DocMMIR: a framework for document multi-modal information retrieval. **arXiv preprint**, arXiv:2505.19312, 2025.
- MANIYAR, S. et al. AI-generated lecture slides for improving slide element detection and retrieval. *In: DOCUMENT ANALYSIS AND RECOGNITION – ICDAR 2025*. Cham: Springer, 2025. (Lecture Notes in Computer Science, v. 16023).
- MIKOLOV, T. et al. Efficient estimation of word representations in vector space. **arXiv preprint**, arXiv:1301.3781, 2013.
- PENNINGTON, J.; SOCHER, R.; MANNING, C. D. GloVe: global vectors for word representation. *In: CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING (EMNLP)*, 2014. **Proceedings** [...]. [S. l.]: ACL, 2014. p. 1532-1543.
- RADFORD, A. et al. Learning transferable visual models from natural language supervision. *In: INTERNATIONAL CONFERENCE ON MACHINE LEARNING*, 2021. **Proceedings** [...]. [S. l.]: PMLR, 2021. p. 8748-8763.
- RUSSELL, S. J.; NORVIG, P. **Artificial intelligence: a modern approach**. 4. ed. [S. l.]: Pearson, 2021.

SAPENA, O.; ONAINDIA, E. Multimodal segmentation and labeling of lecture recordings for educational content retrieval. **Neural Computing and Applications**, v. 38, n. 7, p. 223, 2026.

SHANBHOGUE, M. et al. Gemini Embedding 2: a native multimodal embedding model from Gemini. **arXiv preprint**, arXiv:2605.27295, 2026.

SONG, Y.; SOLEYMANI, M. Polysemous visual-semantic embedding for cross-modal retrieval. *In: IEEE/CVF CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION*, 2019. **Proceedings** [...]. [S. l.]: IEEE, 2019. p. 1979-1988.

SUTOR JUNIOR, P. et al. Metaconcepts: isolating context in word embeddings. *In: IEEE CONFERENCE ON MULTIMEDIA INFORMATION PROCESSING AND RETRIEVAL (MIPR)*, 2019. **Proceedings** [...]. [S. l.]: IEEE, 2019. p. 544-549. DOI: 10.1109/MIPR.2019.00110.

VIEIRA, M. P. R. **Plasticidade e rigidez em modelos de embeddings globais e monolíngues para o Português Brasileiro**. Orientador: Pedro Henrique Lopes Silva. 2026. Monografia (Bacharelado em Ciência da Computação) – Universidade Federal de Ouro Preto, Ouro Preto, 2026.

ZHAO, X.; WANG, G.; LU, Y. MDKAG: retrieval-augmented educational QA powered by a multimodal disciplinary knowledge graph. **Applied Sciences**, v. 15, n. 16, p. 9095, 2025.