Query-CNAT: Uma Ferramenta de Busca Semântica Híbrida para Bancos de Dados Relacionais Médicos

Denilson José do Bom Jesus Silva De Lima[1], **Rui Nóbrega de Pontes Filho**[1], **Allan Miller Silva Lima**[1], **Denis Mayr Lima Martins**[2], **Fernando Buarque de Lima Neto**[1]

[1]Departamento de Engenharia da Computação – Universidade de Pernambuco (UPE) Recife – PE – Brasil

[2]Departamento de Computação e Matemática – Universidade de São Paulo (USP) Ribeirão Preto – SP – Brasil

{djbjsl, rnpf, amsl, fbln}@ecomp.poli.br, martins.denis@usp.br

Abstract. *Retrieving accurate information from medical databases is challenging due to clinical specificity. While LLMs mitigate syntactic rigidity, their probabilistic nature risks hallucinations in strict contexts. This paper demonstrates Query-CNAT, an extractive hybrid semantic search tool that translates medical intent into deterministic tabular subsets. Combining specialized biomedical embeddings (BioBERT-PT and BioWordVec) with evolutionary optimization, the application retrieves precise relational data without relying on external generative APIs. Operating locally, it ensures privacy compliance (LGPD). The demonstration illustrates its interface and effectiveness over an epidemiological database, outperforming lexical baselines by isolating clinically relevant data and purging administrative metadata.*

Resumo. *A recuperação de informações em bancos de dados médicos é desafiadora devido à especificidade clínica. Embora LLMs mitiguem a rigidez sintática, sua natureza probabilística apresenta riscos de alucinações. Este artigo demonstra o Query-CNAT, uma ferramenta extrativa de busca semântica híbrida que traduz intenções médicas em subconjuntos tabulares. Combinando embeddings biomédicos especializados (BioBERT-PT e BioWordVec) e otimização evolutiva, a aplicação recupera dados precisos sem depender de APIs generativas externas. Operando localmente, garante conformidade com a LGPD. A demonstração*[1] *ilustra sua eficácia em um banco de dados epidemiológico real.*

1. Introdução

A recuperação de informações em Sistemas Gerenciadores de Bancos de Dados Relacionais (SGBDR) na área da saúde esbarra na lacuna semântica entre a linguagem clínica (marcada por polissemia e jargões) e os esquemas rígidos de armazenamento [Zhao et al. 2026]. Quando médicos formulam consultas por termos abrangentes, como "Colesterol", motores legados baseados em heurísticas lexicais (e.g., corres-

[1]Vídeo da demonstração: https://youtu.be/4aj0NBHoRD8 — Repositório: https://github.com/denilsonbomjesus/query-cnat

pondência exata ou TF-IDF/BM25) falham em identificar associações biológicas subjacentes (como "HDL" ou "LDL"), especialmente se estas estiverem sob nomenclaturas administrativas sintéticas no banco (e.g., `lipid_panel_v2`) [Kalyan et al. 2021].

Embora Grandes Modelos de Linguagem (LLMs) mitiguem a rigidez sintática, sua natureza probabilística arrisca alucinações em contextos clínicos [Gao et al. 2023, Ji et al. 2023]. Além disso, a dependência de APIs em nuvem fere regulamentações de privacidade (e.g., LGPD e HIPAA) [US DHHS 1996, Brasil 2018]. Em cenários que exigem auditoria e proveniência exata — a capacidade de rastrear o registro original que embasou a inferência —, métodos puramente gerativos tornam-se obstáculos arquitetônicos e jurídicos.

Este trabalho demonstra o **Query-CNAT**, uma ferramenta de busca semântica híbrida e puramente extrativa que traduz intenções médicas diretamente para tabelas relacionais. Integrando *embeddings* (BioWordVec [Zhang et al. 2019], BioBERT [Lee et al. 2020]) e otimização evolutiva, o sistema retorna resultados ranqueados com um filtro automático que purga metadados administrativos, reduzindo a sobrecarga cognitiva. Por operar *on-premise*, garante conformidade com a LGPD e viabiliza pesquisas exploratórias seguras sem recorrer a APIs externas.

2. Arquitetura da Aplicação

A arquitetura do Query-CNAT foi projetada sob o paradigma de *Bi-Encoders* desacoplados, visando garantir a recuperação de dados em frações de segundo sem a necessidade de varreduras exaustivas (*sequential scans*) no banco de dados em tempo de execução. Conforme ilustrado na Figura 1, o fluxo operacional da ferramenta é estruturado em duas fases principais: a **Fase Offline** (Vetorização Estática) e a **Fase Online** (Expansão e Calibração Dinâmica). A aplicação foi desenvolvida em Python, orquestrando as bibliotecas de aprendizado de máquina e otimização heurística em um *pipeline* unificado.

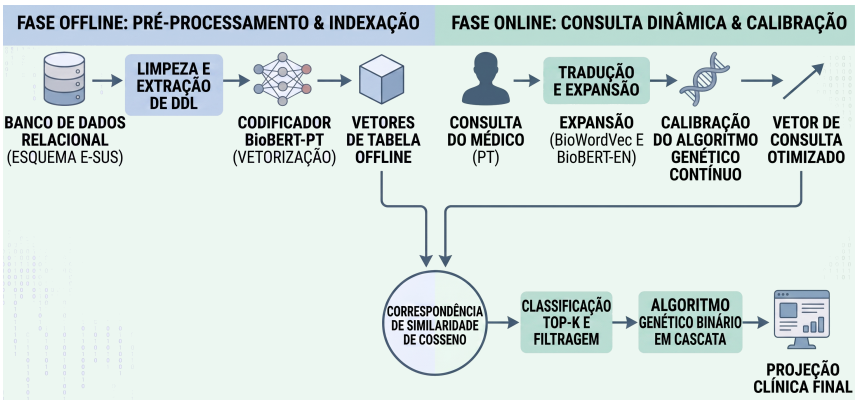

Figura 1. Diagrama arquitetural do Query-CNAT: vetorização estática (BioBERT-PT) na Fase Offline e expansão semântica com calibração evolutiva (BioWordVec + AGs) na Fase Online.

Na **Fase Offline**, o sistema realiza um pré-processamento estrutural para evitar gargalos de performance. O Query-CNAT analisa a Linguagem de Definição de Dados (DDL) do SGBDR, higieniza os metadados (removendo prefixos como `vw_` ou `tb_`) e codifica a semântica estrutural através do modelo **BioBERT-PT** [Schneider et al. 2020],

pré-treinado no *corpus* biomédico brasileiro. Como essa codificação densa é gerada de forma estática e salva em *cache*, o peso computacional do modelo fica restrito a essa etapa inicial, mitigando restrições de *hardware* on-premise nos centros de saúde durante a operação ativa do sistema.

Na **Fase Online**, as consultas em texto livre do usuário são inicialmente traduzidas via *Seq2Seq* para alinhar o jargão local aos *embeddings* médicos globais (MIMIC-III [Johnson et al. 2016]). O sistema emprega o **BioWordVec** [Zhang et al. 2019] — que incorpora o conhecimento ontológico estruturado do *MeSH* — para extrair sinônimos médicos latentes, cuja similaridade contextual é validada pelo **BioBERT-EN** [Lee et al. 2020] ainda no espaço semântico em inglês. Em seguida, esses termos expandidos são retrotraduzidos e filtrados pela máscara contextual do BioBERT-PT. Essa expansão resolve a lacuna semântica, mas introduz um desafio de otimização não-diferenciável para equilibrar os sinais topológicos e contextuais. Para resolver essa calibração sem introduzir vieses manuais ou pesos fixos, o Query-CNAT mapeia as variáveis por meio de um pipeline evolutivo bimodal disposto em subcamadas operacionais.

2.1. Otimização Evolutiva: Mecanismos e Seleção de Características

O motor de busca resolve o balanceamento de relevância dos termos através de uma cascata de **Algoritmos Genéticos (AGs)** [Mirjalili et al. 2020, Katoch et al. 2021], gerenciados via biblioteca *PyGAD* [Gad 2023]. Para viabilizar a execução sub-segundo em ambientes de infraestrutura local modesta, a heurística atua de forma complementar e sequencial em duas frentes:

1. **Ranqueamento Dinâmico (AG Contínuo):** Navega no espaço de busca vetorial para calibrar os pesos de relevância dos sinônimos. Para mitigar o desvio semântico (*semantic drift*), o algoritmo opera sobre um espaço altamente restrito, limitando o pool de sinônimos candidatos de forma estrita a um intervalo de $3 \leq N \leq 20$ termos dominantes. Cada solução é codificada como um cromossomo contínuo $W = [w_1, ..., w_n]$ com $w_i \in [0, 1]$, ajustando os componentes geométricos para reter o foco na intenção clínica central. Essa calibração é guiada diretamente por uma função de aptidão *(fitness)* que incorpora os scores contextuais do BioBERT-PT [Schneider et al. 2020], impedindo que sinônimos periféricos causem desvio semântico.
2. **Seleção de Colunas Clínicas (AG Binário):** Após o ranqueamento e isolamento das tabelas candidatas, um segundo AG opera localmente como um filtro de exibição estrutural. Codificado como um cromossomo binário $X \in \{0, 1\}^k$, o algoritmo avalia a assinatura de cada coluna para decidir sua retenção (1) ou purgação (0). A função de aptidão (*fitness*) penaliza ativamente subconjuntos grandes por meio de um regularizador de parcimônia, aprendendo a remover chaves identificadoras (`uuid`, `id_paciente`), flags e metadados de sessão (`session_token`). O resultado projeta um Phenotype tabular limpo focado puramente no conteúdo biológico essencial.

3. Demonstração Prática e Funcionalidades

A interface gráfica do Query-CNAT foi desenvolvida utilizando o *framework* Streamlit, focado em interatividade reativa e processamento rápido on-premise por meio de *ca-*

che em memória que mantém os modelos prontos para inferência imediata. O fluxo de demonstração inicia-se com o médico inserindo um termo clínico livre (e.g., "colesterol"). Por meio de controles deslizantes na interface, o especialista define de forma clara e parametrizada as restrições do sistema, estipulando o limiar (*threshold*) mínimo de similaridade de cosseno aceito para a validação dos sinônimos e o número máximo de tabelas exibidas no *ranking*.

Conforme ilustrado na Figura 2, a tela principal expõe os resultados de forma transparente (*Explainable AI*): exibe-se os sinônimos sugeridos (e.g., "triglicerídeo", "ldlc") acompanhados da importância percentual exata calculada pelo AG Contínuo, garantindo a auditoria do termo raiz. À direita, exibe-se o ordenamento determinístico das tabelas correspondentes.

Pesos da Consulta Otimizados pelo AG

O AG aprendeu a importância de cada termo para esta busca:

	Termo Candidato	Peso (Importância)
0	colesterol	36.11%
10	não esterificado	14.32%
14	contendo apob	11.57%
3	lipoproteína de densidade muito baixa	8.57%
13	vldl-triacilglicerol	6.54%
12	quilomícron	3.76%
1	triglicerídeo	3.49%
5	t-col	3.16%
15	prebeta-hdl	2.15%
18	alfa-hdl	1.89%

Ranking Final das Tabelas

Tabelas mais relevantes, ordenadas pela consulta otimizada.

	Tabela	Similaridade (Pontuação)
0	tb_exame_colesterol_ldl	0.8308
1	tb_exame_colesterol_hdl	0.8150
2	ta_exame_colesterol_hdl	0.8032
3	tb_exame_colesterol_total	0.7905
4	tb_tipo_glicemia	0.7748
5	tb_exame_triglicerideos	0.7632
6	tb_exame_creatina_serica	0.7568
7	ta_exame_triglicerideos	0.7535
8	ta_exame_hemoglobina_glicada	0.7504
9	tl_cds_pic	0.7378

Figura 2. Interface do sistema. À esquerda, os pesos evolutivos dos sinônimos expandidos; à direita, o *ranking* das tabelas relacionais retornadas para a consulta "colesterol".

A funcionalidade mais significativa ocorre ao expandir uma tabela específica: o AG Binário intercepta a DDL e executa um *de-noise* estrutural, ocultando colunas administrativas e retornando exclusivamente um quadro clínico limpo. Isso entrega os dados necessários sem exigir que o médico compreenda o paradigma relacional complexo, mitigando de forma robusta a fadiga cognitiva.

4. Resultados Preliminares

A arquitetura foi avaliada empiricamente no esquema do banco epidemiológico público brasileiro (E-SUS) — contendo 1.100 tabelas e mais de 9.600 colunas —, utilizando a consulta clínica "dislipidemia" como desafio semântico. A eficácia do Query-CNAT foi contrastada com o *baseline* BM25. As métricas (Tabela 1) refletem o desempenho de recuperação extrativa durante o estudo de caso demonstrado.

A validação quantitativa utilizou um subconjunto cardiovascular focado em métricas de "Colesterol" (relacionado a "dislipidemia") como **proxy** de teste. O **Ground Truth** foi construído via metodologia de supervisão fraca, utilizando um dicionário clínico extraído das diretrizes AHA/ACC de 2018 [Grundy et al. 2019] por meio de **prompt** ontológico. Os termos foram normalizados e convertidos em expressões regula-

Tabela 1. Avaliação contra o *Baseline* Lexical sob *Proxy* Cardiovascular.

Métrica	Baseline BM25	Query-CNAT (Fase 1)	Query-CNAT (Fase 2)
Precision@5	0.00	**0.60**	**0.60**
Precision@10	0.00	**0.40**	**0.40**
F1-Score	0.00	**0.62**	0.28

res para um classificador automatizado, que rotulou as tabelas alvo com base na densidade semântica dos metadados do esquema, mitigando o viés de anotação manual.

O Query-CNAT atingiu **Precision@5 de 0.60** na Fase 1 (Expansão Semântica), superando a vulnerabilidade do BM25 na ausência do termo exato de busca (0.00) [Zhang et al. 2019]. Essa alta precisão inicial é vital no contexto de saúde para evitar a fadiga do usuário. O declínio observado no *F1-Score* na Fase 2 (0.28) decorre de uma escolha de projeto deliberada: o filtro de parcimônia do AG Binário penaliza subconjuntos grandes para purgar colunas administrativas irrelevantes. Sacrifica-se a revocação periférica total sobre o universo de 1.100 tabelas para priorizar a legibilidade clínica imediata. Essa abordagem puramente extrativa mitiga as alucinações e as falhas de integridade em esquemas massivos mapeadas em modelos generativos clássicos de *Text-to-SQL* como o DIN-SQL [Pourreza e Rafiei 2023]. Além disso, enquanto a implantação de LLMs locais (e.g., LLaMA-3) para inferência estruturada exige infraestrutura de alto custo com GPUs corporativas dedicadas, o Query-CNAT consome recursos mínimos por operar sobre catálogos vetoriais em *cache*, tornando a busca semântica viável mesmo em servidores hospitalares modestos.

5. Conclusões e Trabalhos Futuros

A arquitetura do Query-CNAT demonstra a viabilidade de integrar *embeddings* biomédicos especializados (BioBERT e BioWordVec) com heurísticas evolutivas para a recuperação semântica de dados em esquemas relacionais de saúde. A validação na base do E-SUS evidenciou que a aplicação mapeia com sucesso a intenção clínica livre para as restrições tabulares, superando as abordagens lexicais tradicionais (BM25) na precisão imediata. Ao adotar uma metodologia estritamente extrativa, a ferramenta reduz os riscos de alucinações comuns aos LLMs generativos, garantindo a rastreabilidade determinística exigida em ambientes críticos, como na medicina. Além disso, a sua implantação local está alinhada aos requisitos de privacidade da LGPD. Trabalhos futuros incluem investigar a automação dos hiperparâmetros do sistema via ferramentas de AutoML e o desenvolvimento de gatilhos de banco de dados para re-vetorização síncrona.

Agradecimentos

Este trabalho foi realizado com o apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001. Os autores agradecem o apoio financeiro do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e da Fundação de Amparo à Ciência e Tecnologia do Estado de Pernambuco (FACEPE).

Declaração de Uso de IA Generativa

Declara-se o uso da ferramenta Google Gemini estritamente para o auxílio na formatação e na revisão linguística do manuscrito. A responsabilidade integral pelo conteúdo permanece com os autores.

Referências

Zhao, W.X., et al.: A survey of large language models. Frontiers of Computer Science **20**(12), 2012627 (2026). doi:10.1007/s11704-026-60308-3

Kalyan, K., Rajasekharan, A., Sangeetha, S.: AMMUS: a survey of transformer-based pretrained models in natural language processing. arXiv preprint arXiv:2108.05542 (2021).

Gad, A.F.: PyGAD: an intuitive genetic algorithm Python library. Multimedia Tools Appl. (2023). doi:10.1007/s11042-023-17167-y

Grundy, S.M., et al.: 2018 AHA/ACC guideline on the management of blood cholesterol. Journal of the American College of Cardiology **73**(24), e285–e350 (2019). doi:10.1016/j.jacc.2018.11.003

Mirjalili, S., Dong, J., Lewis, A.: Nature-Inspired Optimizers: Theories, Literature Reviews and Applications. Springer (2020). doi:10.1007/978-3-030-12127-3

Gao, Y., et al.: Retrieval-augmented generation for large language models: a survey. arXiv preprint arXiv:2312.10997 (2023).

Johnson, A.E.W., et al.: MIMIC-III, a freely accessible critical care database. Scientific Data **3**, 160035 (2016). doi:10.1038/sdata.2016.35

Ji, Z., et al.: Survey of hallucination in natural language generation. ACM Comput. Surv. **55**(12), 1–38 (2023). doi:10.1145/3571730

Zhang, Y., et al.: BioWordVec, improving biomedical word embeddings with subword information and MeSH. Scientific Data **6**(1), 52 (2019). doi:10.1038/s41597-019-0055-0

Lee, J., et al.: BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics **36**(4), 1234–1240 (2020). doi:10.1093/bioinformatics/btz682

Schneider, E., et al.: BioBERTpt - A Portuguese Neural Language Model for Clinical Named Entity Recognition. In: Proc. Clinical Natural Language Processing Workshop. pp. 65–72 (2020). doi:10.18653/v1/2020.clinicalnlp-1.7

Katoch, S., Chauhan, S.S., Kumar, V.: A review on genetic algorithm: past, present, and future. Multimedia Tools Appl. **80**, 8091–8126 (2021). doi:10.1007/s11042-020-10139-6

Pourreza, M., Rafiei, D.: DIN-SQL: decomposed in-context learning of text-to-SQL with self-correction. In: Adv. Neural Inf. Process. Syst. vol. 36, pp. 36339–36348 (2023).

U.S. Dept. Health Human Serv.: HIPAA of 1996. Pub. L. 104-191 (1996).

Presidência da República do Brasil: LGPD, Lei nº 13.709. Diário Oficial da União (2018).