ARS VETERINARIA, Jaboticabal, SP, v.42, n.2, 040-048, 2026. ISSN 2175-0106
DOI: http://dx.doi.org/10.15361/2175-0106.2026v42n2p40-48

SOPHIALLM: DESENVOLVIMENTO E VALIDAÇÃO DE UM MODELO DE LINGUAGEM ESPECIALIZADO PARA RACIOCÍNIO CLÍNICO, DIAGNÓSTICO E EDUCAÇÃO EM MEDICINA VETERINÁRIA DE PEQUENOS ANIMAIS

SOPHIALLM: DEVELOPMENT AND VALIDATION OF A SPECIALIZED LANGUAGE MODEL FOR CLINICAL REASONING, DIAGNOSIS AND EDUCATION IN SMALL ANIMAL VETERINARY MEDICINE

F. S. BARROS[1*]; A. C. MINATO[2]; R. O. SANTOS JUNIOR[2]; L. M. SOUZA[3]; C. F. M. MANSANO[4]; E. C. A. SIMÊNCIO[5]; L. A. M. PEREIRA[4]; M. A. A. BELO[4]; I. C. SILVA[4]

RESUMO

A SophiaLLM é um modelo de linguagem especializado em medicina veterinária de pequenos animais (cães e gatos), desenvolvido integralmente em português brasileiro para apoiar raciocínio clínico, diagnóstico, educação continuada veterinária e recuperação de conhecimento científico. Diferentemente de modelos generalistas, a SophiaLLM foi projetada exclusivamente para o domínio veterinário, buscando maior precisão e menor incidência de alucinações. O sistema combina um modelo de linguagem com aproximadamente 109,5 milhões de parâmetros a uma arquitetura de Retrieval-Augmented Generation (RAG), permitindo que o conhecimento científico seja armazenado e atualizado separadamente dos pesos do modelo. Sua base de conhecimento é composta por aproximadamente 500 milhões de tokens (cerca de 6,94 GB) de literatura veterinária, dezenas de milhares de casos clínicos, um grafo de conhecimento contendo 454 doenças, 680 entidades e 9.546 relações clínicas, além de um motor de raciocínio clínico capaz de gerar hipóteses diagnósticas, diagnósticos diferenciais e recomendações de exames complementares. A arquitetura foi desenvolvida para reproduzir o processo de raciocínio utilizado por médicos-veterinários, correlacionando sinais clínicos, exames laboratoriais, exames de imagem e histórico do paciente para apoiar a tomada de decisão clínica. A separação entre conhecimento factual e raciocínio permite maior auditabilidade, atualização contínua e redução de erros gerados por inteligência artificial. A SophiaLLM também atua como uma assistente inteligente para clínicas veterinárias, permitindo o registro de informações por voz, transcrição automática de consultas, interpretação de exames, documentos e notas fiscais por meio de visão computacional. Integrada aos sistemas de gestão de clínicas veterinárias, através de API, preenche prontuários, registra procedimentos, auxilia no controle financeiro e automatiza tarefas administrativas. A SophiaLLM representa uma iniciativa pioneira para a criação de uma inteligência artificial veterinária especializada, demonstrando que modelos menores, combinados com bases de conhecimento estruturadas e mecanismos de recuperação de informação, podem oferecer desempenho relevante para aplicações clínicas, educacionais e científicas na medicina veterinária de cães e gatos.

PALAVRAS-CHAVE: Decoder-only. Grafo de conhecimento. Modelo de linguagem de grande escala. Português brasileiro. Raciocínio clínico. RAG.

SUMMARY

SophiaLLM is a language model specialized in small animal veterinary medicine (dogs and cats), developed entirely in Brazilian Portuguese to support clinical reasoning, diagnosis, continuing veterinary education and scientific knowledge retrieval. Unlike generalist models, SophiaLLM was designed exclusively for the veterinary domain, seeking greater accuracy and a lower incidence of hallucinations. The system combines a language model with approximately 109.5 million parameters with a Retrieval-Augmented Generation (RAG) architecture, allowing scientific knowledge to be stored and updated separately from the model weights. Its knowledge base comprises approximately 500 million tokens (about 6.94 GB) of veterinary literature, tens of thousands of clinical cases, a knowledge graph containing 454 diseases, 680 entities and 9,546 clinical relations, in addition to a clinical reasoning engine capable of generating diagnostic hypotheses, differential diagnoses and recommendations for complementary examinations. The architecture was developed to reproduce the reasoning process used by veterinarians, correlating clinical signs, laboratory tests, imaging examinations and patient history to support clinical decision-making. The separation between factual knowledge and reasoning allows greater auditability, continuous updating and the reduction of errors generated by artificial intelligence. SophiaLLM also acts as an intelligent assistant for veterinary clinics, enabling voice-based information recording, automatic transcription of consultations, and interpretation of exams, documents and invoices through computer vision. Integrated into clinic systems via API, it fills in medical records, registers procedures, assists in financial control and automates administrative tasks. SophiaLLM represents a pioneering initiative for the creation of specialized veterinary artificial intelligence, demonstrating that smaller models, combined with structured knowledge bases and information retrieval mechanisms, can offer relevant performance for clinical, educational and scientific applications in the veterinary medicine of dogs and cats.

KEY-WORDS: Brazilian Portuguese. Clinical reasoning. Decoder-only. Knowledge graph. Large language model. RAG.

[1] Docente do Departamento de Clínica Médica, Universidade Brasil (UB), Descalvado, São Paulo, Brasi
[2] Docente do Curso de Medicina Veterinária, Universidade Brasil (UB), Descalvado, São Paulo, Brasil.
[3] Laboratório de Parasitologia Veterinária, Universidade Brasil (UB), Descalvado, São Paulo, Brasil.
[4] Programa de Pós-Graduação em Produção Animal, Universidade Brasil (UB), Descalvado, SP, Brasil.
[5] Departamento Administrativo, Universidade Brasil (UB), Descalvado, São Paulo, Brasil.
Autor para correspondência: fbarros.medvet@gmail.com***

INTRODUÇÃO

A medicina veterinária de pequenos animais é uma área de crescente demanda no Brasil, terceiro maior mercado pet do mundo (BRASIL, 2024). Apesar dessa relevância, o acesso a ferramentas de suporte clínico baseadas em inteligência artificial em português brasileiro permanece extremamente limitado. Os principais modelos generalistas: GPT-4, Claude e Gemini, são treinados predominantemente em inglês e apresentam lacunas críticas quando interrogados em português sobre protocolos clínicos, farmacologia e diagnóstico veterinário.

Modelos de linguagem especializados em medicina humana têm avançado significativamente. O Med-PaLM (SINGHAL et al., 2023) demonstrou desempenho competitivo em questões no estilo do USMLE; BioMedLM (BOLTON et al., 2022) e BioGPT (LUO et al., 2022) exploraram domínios biomédicos em inglês. Na área veterinária, o VetBERT (HUR et al., 2020) aplicou BERT à classificação de registros clínicos veterinários, e trabalhos recentes utilizaram GPT-4 via prompting para a realização de consultas veterinárias. Contudo, até o momento, nenhum trabalho publicado descreve uma LLM decoder-only desenvolvida do zero especificamente para medicina veterinária de pequenos animais, tampouco a integração de um grafo de conhecimento veterinário a um motor de raciocínio diferencial.

No âmbito de modelos decoder-only, a arquitetura GPT-2 (RADFORD et al., 2019) estabeleceu a base do paradigma, e o LLaMA (TOUVRON et al., 2023) popularizou o uso de RMSNorm, RoPE e SwiGLU em escala. Do ponto de vista arquitetural, a SophiaLLM foi concebida como uma plataforma escalável, permitindo sua evolução progressiva entre diferentes versões (V1, V2 e V3) sem alterações estruturais significativas na interação com o usuário. Para isso, foi adotada uma arquitetura baseada em Retrieval-Augmented Generation (RAG) com interface padronizada por meio de tokens especiais, garantindo compatibilidade futura e atualização contínua da base de conhecimento sem a necessidade de retreinamento completo do modelo. O presente trabalho descreve o desenvolvimento e a validação da arquitetura da SophiaLLM, contemplando a verificação da contagem de parâmetros do modelo, testes automatizados do processo de treinamento e auditoria dos principais componentes do sistema, com o objetivo de demonstrar a viabilidade técnica de uma inteligência artificial veterinária especializada em português brasileiro.

MATERIAL E MÉTODOS

Arquitetura do modelo

A SophiaLLM é um decoder-only autoregressivo construído com PyTorch 2.x, sem utilização de pesos pré-treinados. A arquitetura incorpora modificações modernas e tem aproximadamente 109,5 milhões de parâmetros. O modelo utiliza um vocabulário de 32.000 tokens representados por embeddings de dimensão 768, compartilhados com a camada de saída (weight tying). A rede é composta por 12 blocos Transformer empilhados, cada um contendo mecanismos de auto atenção com 12 cabeças, projeções residuais e camadas MLP baseadas na função de ativação SwiGLU. A arquitetura emprega codificação posicional RoPE e normalização RMSNorm, seguindo padrões modernos de modelos de linguagem de médio porte. A saída final é produzida por uma camada LM Head que gera probabilidades sobre os 32.000 tokens do vocabulário, permitindo a predição autoregressiva de texto em português e domínio veterinário (Figura 1).

Memória e custo computacional

A SophiaLLM V1 foi projetada para ser treinável em hardware de pesquisa de médio porte, mantendo uma relação favorável entre capacidade de modelagem e custo computacional. Em precisão float32, os pesos do modelo ocupam aproximadamente 418 MB, enquanto em bfloat16 esse valor é reduzido para cerca de 209 MB. Durante o treinamento com AdamW, considerando pesos, gradientes e os dois momentos do otimizador, o consumo mínimo estimado é de aproximadamente 1,7 GB de memória apenas para os parâmetros do modelo em bf16, sem contabilizar ativações intermediárias. Na prática, considerando batches moderados e o armazenamento das ativações necessárias para retropropagação, o treinamento completo requer entre 6 e 10 GB de VRAM, permitindo sua execução em GPUs de consumo amplamente disponíveis. O protocolo de treinamento adotado utiliza AdamW com taxa de aprendizado inicial de 3×10^{-4}, decaimento cossenoidal de até 3×10^{-5}, warmup de 2.000 passos, weight decay de 0,1, gradient clipping em 1,0 e batch efetivo de 128 exemplos (16×8), totalizando 100.000 passos de otimização. O custo computacional do treinamento segue a aproximação clássica para Transformers autoregressivos, resultando em aproximadamente $5{,}1 \times 10^{8}$ FLOPs por token processado (forward + backward), valor compatível com modelos decoder-only da mesma escala.

Figura 1 - Arquitetura, implementação da SophiaLLM. (a) Fluxo computacional da SophiaLLM V1, mostrando o pipeline autoregressivo composto por camada de *Embedding* (32.000 × 768), 12 blocos *decoder* Transformer, RMSNorm final e LM Head compartilhada (*weight tying*). O detalhe interno de cada bloco apresenta a sequência RMSNorm → Multi-Head Attention → conexão residual → RMSNorm → MLP SwiGLU → conexão residual. **(b)** Representação esquemática da rede neural completa, destacando as dimensões dos embeddings, as 12 cabeças de atenção, as projeções internas e a profundidade de 12 camadas, totalizando aproximadamente 109,5 milhões de parâmetros. **(c)** Tokens especiais introduzidos para integração nativa com recuperação de informação (RAG), incluindo marcadores de pergunta, contexto e resposta, além do formato de treinamento utilizado no Estágio 2 (<q>, <ctx>, <a>), permitindo que o modelo condicione suas respostas ao contexto recuperado durante a inferência.

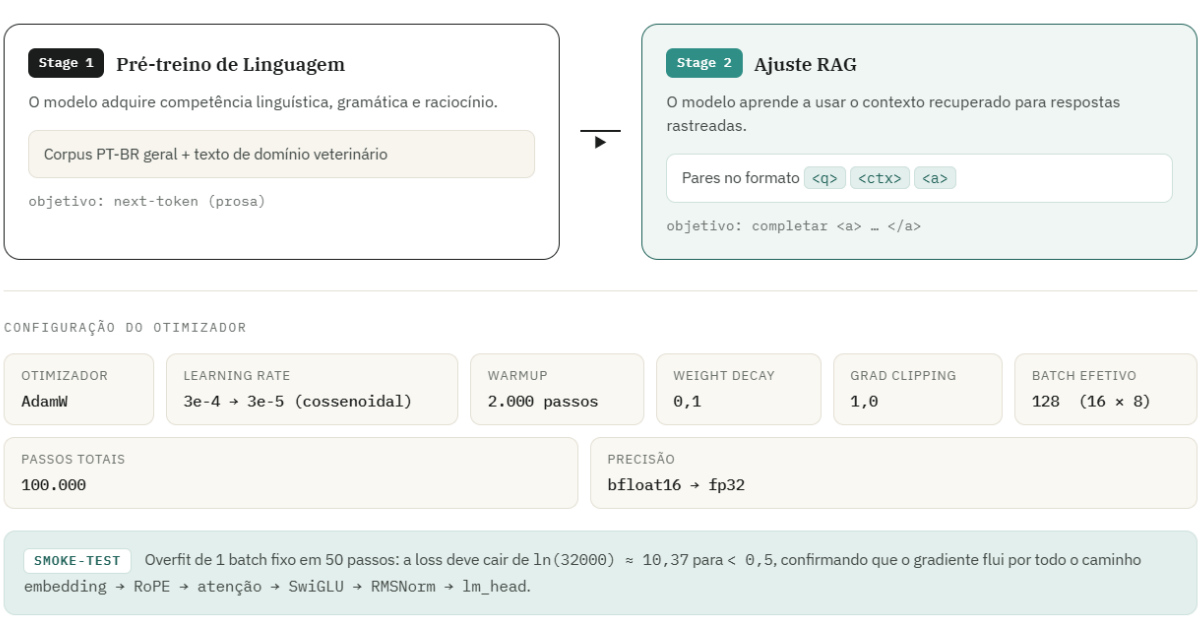

Figura 2 - Configuração de treinamento e requisitos computacionais da SophiaLLM V1.
Representação dos principais parâmetros utilizados durante o treinamento do modelo, incluindo precisão numérica, consumo estimado de memória, configuração do otimizador AdamW, cronograma de taxa de aprendizado, tamanho efetivo de batch e custo computacional aproximado para treinamento do modelo.

Tokenizador

A SophiaLLM utiliza um tokenizador SentencePiece baseado no algoritmo Unigram, com vocabulário fixo de 32.000 tokens treinado sobre um corpus composto por textos gerais em português brasileiro e documentos especializados em medicina veterinária de cães e gatos. A escolha desse tamanho de vocabulário representa um compromisso entre eficiência computacional e preservação semântica dos termos técnicos, reduzindo significativamente a fragmentação de expressões clínicas complexas como hipoadrenocorticismo, trombocitopenia e leishmaniose visceral. Os embeddings de entrada possuem dimensão 768 e são compartilhados com a camada de projeção de saída por meio da estratégia de weight tying, reduzindo o número total de parâmetros treináveis e melhorando a eficiência estatística do modelo. Além dos tokens convencionais de controle (<pad>, <unk>, <bos> e <eos>), a SophiaLLM incorpora seis tokens especiais nativos destinados à integração com o mecanismo Retrieval-Augmented Generation (RAG): <ctx>, </ctx>, <q>, </q>, <a> e (Figura 1C). Diferentemente de abordagens que utilizam marcadores textuais adicionados posteriormente ao treinamento, esses tokens fazem parte do vocabulário original e possuem identificadores fixos, garantindo estabilidade entre versões do modelo. Durante o segundo estágio de treinamento, dedicado ao ajuste RAG, as amostras seguem o formato: <bos> <q> pergunta </q> <ctx> contexto recuperado </ctx> <a> resposta <eos>. Esse esquema ensina explicitamente o modelo a distinguir pergunta, contexto recuperado e resposta, favorecendo a geração condicionada por evidências externas.

Base de conhecimento de domínio

A base de conhecimento da SophiaLLM foi construída por meio de um processo híbrido de curadoria manual e semi-automatizada, priorizando rastreabilidade, auditabilidade e relevância clínica. O corpus final totaliza 81.242 arquivos válidos, derivados de um universo inicial de 370.646 documentos processados, ocupando aproximadamente 6,94 GB em disco e correspondendo a 500,2 milhões de tokens (304,2 milhões de palavras). Todo o conteúdo foi filtrado para contemplar exclusivamente medicina veterinária de pequenos animais, com foco em cães e gatos. A estratégia de curadoria eliminou conteúdos relacionados a espécies de produção, animais silvestres e material de baixa relevância clínica. Cada documento pode ser rastreado até sua fonte primária, composta principalmente por livros-texto especializados, manuais técnicos, protocolos clínicos e literatura científica revisada por pares. Foram identificados 59.471 documentos com conteúdo relacionado a cães e 68.676 com conteúdo relacionado a gatos, resultando em volumes de tokens praticamente equivalentes, com 444,4 milhões de tokens para cães e 440,6 milhões para gatos, representando diferença inferior a 1% entre as espécies.

O conhecimento foi organizado em múltiplas camadas complementares destinadas a funções distintas durante treinamento e inferência. A primeira camada é composta por textos clínicos curados e literatura especializada. A segunda reúne conteúdos estruturados em formato de perguntas e respostas para treinamento instrucional. A terceira contempla materiais organizados para recuperação de informação via RAG. A quarta camada consiste em

conhecimento explicitamente estruturado por meio de um grafo veterinário e de um motor de raciocínio clínico. Essa separação permite que fatos clínicos permaneçam fora dos pesos da rede neural, reduzindo obsolescência do modelo e facilitando atualizações futuras sem necessidade de novo treinamento completo. A camada de prosa clínica curada reúne 527 documentos especializados, totalizando aproximadamente 449 mil tokens auditados, distribuídos em múltiplas especialidades veterinárias, com maior representatividade em Farmacologia (67.554 tokens), Antibioticoterapia em felinos (48.866 tokens), Clínica médica (38.488 tokens), Hematologia veterinária (26.153 tokens), Neurologia (17.975 tokens) e Oftalmologia (17.223 tokens).

Grafo de conhecimento e motor de raciocínio clínico

O componente de conhecimento estruturado da SophiaLLM foi implementado por meio de um grafo veterinário contendo 454 doenças, conectadas por 9.546 relações semânticas tipadas. A estrutura foi construída a partir de literatura técnica especializada e submetida a validação automática para eliminação de duplicidades e inconsistências. A análise topológica mostra que cada doença se conecta, em média, a diferentes sinais clínicos, exames diagnósticos, diagnósticos diferenciais e possíveis complicações, reproduzindo de forma explícita parte do raciocínio utilizado por médicos-veterinários durante a prática clínica. Complementando o grafo, foi desenvolvido um motor de raciocínio clínico destinado à geração de hipóteses diagnósticas e priorização de exames complementares. Foram gerados e validados aproximadamente 3.000 exemplos de treinamento clínico, nos quais cada hipótese recebe pesos de relevância baseados na força de associação entre sinais e doenças. O objetivo do motor não é emitir diagnósticos definitivos, mas auxiliar na formulação de hipóteses, sugerir exames discriminatórios e explicar o raciocínio clínico subjacente.

Procedimentos de validação

A validação formal da SophiaLLM V1 foi realizada por meio de uma suíte automatizada de testes. Os testes incluíram: (i) verificação da contagem de parâmetros (test_param_count), assegurando concordância entre os valores analíticos e implementados; (ii) validação das dimensões dos logits na passagem forward (test_forward_shape); (iii) confirmação do compartilhamento de pesos entre a camada de embeddings e a projeção de saída (test_weight_tying); (iv) consistência entre geração com e sem KV-cache (test_kv_cache_consistency); e (v) validação da expansão dinâmica do cache RoPE para sequências superiores ao comprimento de contexto inicial (test_long_generation). Adicionalmente, a integridade do fluxo de gradientes foi avaliada por meio de um smoke-test de overfitting em lote único.

RESULTADOS E DISCUSSÃO

A execução dos testes confirmou uma contagem total de 109.529.856 parâmetros, dos quais 84.953.856 correspondem à parte não-embedding da arquitetura, em concordância com os cálculos teóricos. A suíte completa do projeto registrou aprovação dos 16 testes executados. No smoke-test de overfitting em lote único, a função de perda reduziu de $\ln(32000) \approx 10{,}37$ para valores inferiores a 0,5 em aproximadamente 50 passos, confirmando o correto funcionamento da cadeia embedding → RoPE → atenção → SwiGLU → RMSNorm → camada de saída.

Os componentes de conhecimento também foram auditados. O grafo veterinário foi processado com sucesso, totalizando 454 doenças e 9.546 relações sem duplicidades ou inconsistências estruturais. O motor de raciocínio clínico produziu 3.000 exemplos de treinamento validados automaticamente, sem falhas de integridade. A auditoria independente do dataset atribuiu nota geral de 8,5/10, alcançando 10/10 em qualidade científica e rastreabilidade das fontes, evidenciando a robustez da base de conhecimento utilizada para treinamento e recuperação de informações (Figura 3).

Na avaliação qualitativa em cenário clínico, envolvendo um gato geriátrico com anorexia, perda de peso e polidipsia, a SophiaLLM organizou o diagnóstico diferencial priorizando as hipóteses por probabilidade clínica, destacando a doença renal crônica (DRC) como principal suspeita, seguida por hipertireoidismo e diabetes mellitus, e indicou de forma objetiva o perfil bioquímico associado à urinálise como exame inicial de maior poder discriminante para o caso. Embora os modelos generalistas (ChatGPT, Gemini e Claude) tenham sido capazes de identificar hipóteses diagnósticas plausíveis, apenas a SophiaLLM apresentou raciocínio diferencial estruturado, com ranqueamento explícito das hipóteses por probabilidade clínica, posicionamento adequado do hipertireoidismo no contexto do caso, indicação de exames discriminantes e conexão direta entre sinais clínicos, exames complementares e hipóteses diagnósticas (Figuras 4 e 5). Embora a avaliação tenha sido qualitativa e baseada em um único caso representativo, os achados fornecem evidências preliminares de que a SophiaLLM não apenas recupera conhecimento clínico, mas também organiza esse conhecimento em uma sequência lógica de investigação diagnóstica.

A SophiaLLM foi concebida como uma arquitetura escalável, organizada em três versões sucessivas (Figura 6), preservando o mesmo tokenizador SentencePiece de 32 mil tokens e a mesma interface RAG baseada em tokens especiais. A versão inicial (V1), validada neste trabalho, possui aproximadamente 109,5 milhões de parâmetros totais, dos quais cerca de 85 milhões correspondem aos componentes não relacionados aos embeddings, com dimensão oculta (d_model) de 768, 12 camadas Transformer e 12 cabeças de atenção. A evolução planejada inclui a SophiaLLM V2, com aproximadamente 226 milhões de parâmetros não-embedding (d_model = 1024, 18 camadas e 16 cabeças), incorporando treinamento supervisionado (Supervised Fine-Tuning – SFT) além do mecanismo RAG. A SophiaLLM V3 amplia a escala para aproximadamente 471 milhões de parâmetros não-embedding (d_model = 1280, 24 camadas e 20 cabeças), adicionando capacidades multimodais integradas ao mesmo framework de recuperação de conhecimento.

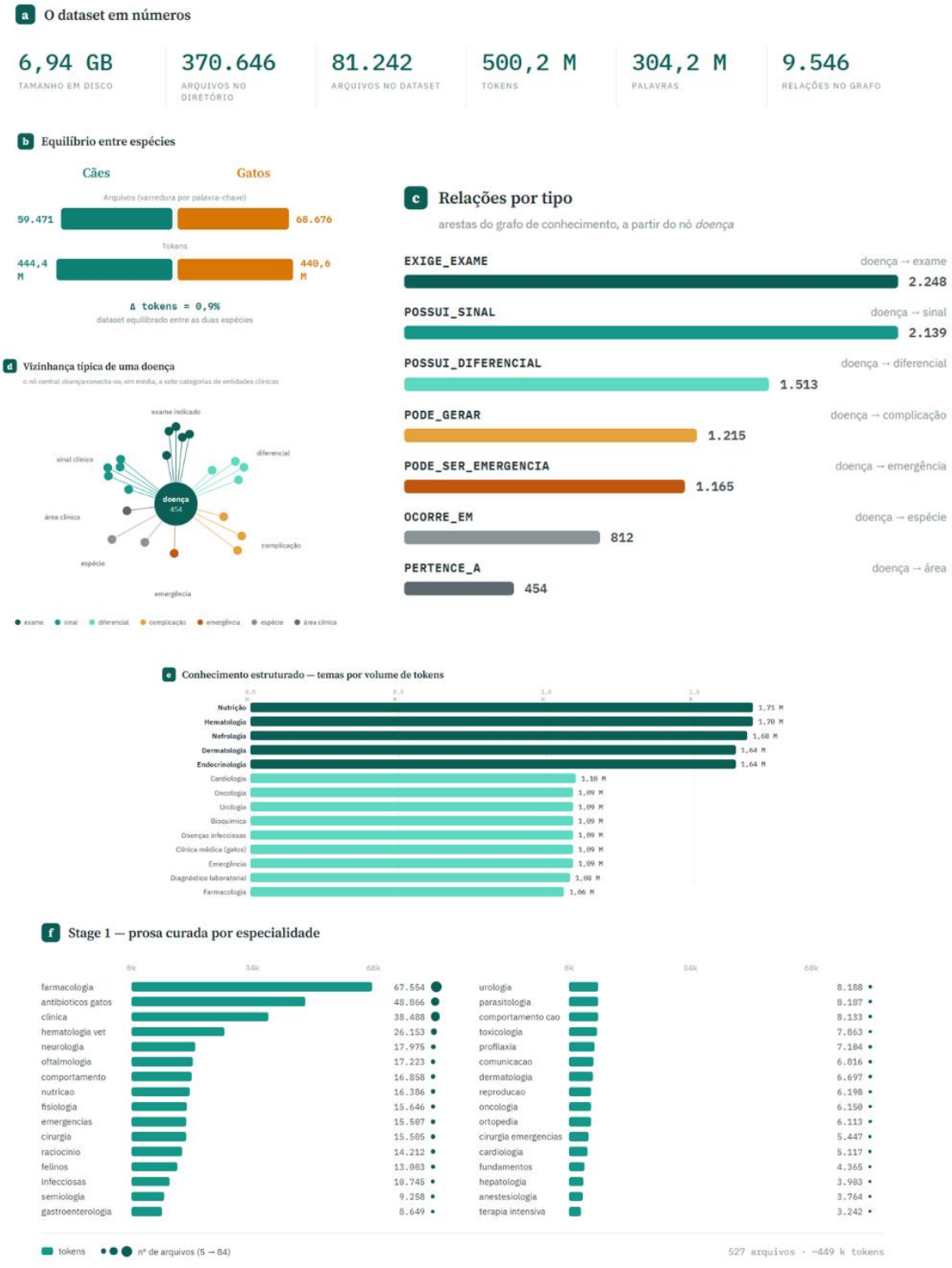

Figura 3 - Caracterização do corpus veterinário. (a) Métricas globais do dataset. (b) Comparação entre cães e gatos por número de arquivos e por tokens; a diferença de tokens é de apenas 0,9%, indicando um corpus balanceado. (c) Contagem de arestas do grafo de conhecimento por tipo de relação, partindo do nó doença. (d) Esquema da vizinhança média de um nó doença (n = 454), conectado a exames, sinais, diferenciais, complicações, emergências, espécies e áreas clínicas — código de cores compartilhado com o painel (c). (e) Volume de tokens por tema do conhecimento estruturado. (f) Prosa curada na Etapa 1, por especialidade: comprimento da barra = tokens; diâmetro do ponto = nº de arquivos. Cor de cães em verde-petróleo; gatos em laranja.

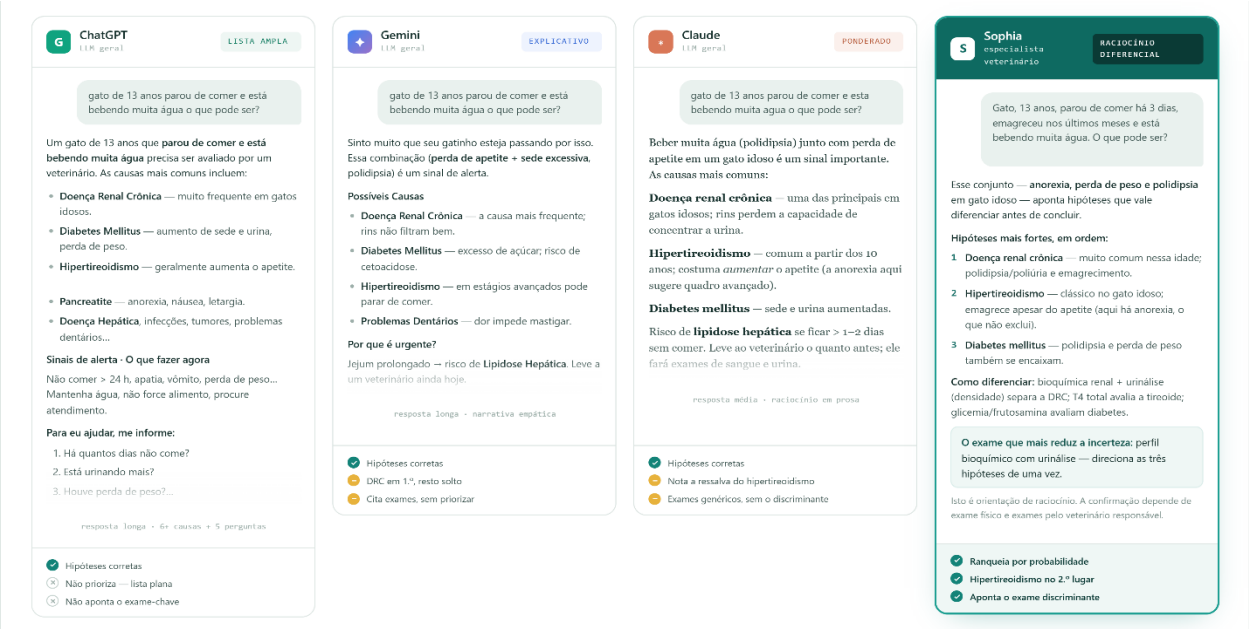

Figura 4 - Comparação qualitativa entre diferentes modelos de linguagem (ChatGPT, Gemini, Claude e SophiaLLM) na análise de um caso clínico veterinário envolvendo um gato geriátrico com anorexia, perda de peso e polidipsia. Enquanto os modelos generalistas apresentam listas de hipóteses diagnósticas e recomendações clínicas em diferentes níveis de detalhamento, a SophiaLLM utiliza um mecanismo de raciocínio diferencial especializado em medicina veterinária, priorizando hipóteses por probabilidade clínica, justificando cada inferência e indicando exames complementares capazes de reduzir a incerteza diagnóstica. Observa-se que a abordagem da SophiaLLM enfatiza a interpretação contextual dos sinais clínicos e a tomada de decisão orientada por evidências, características essenciais para sistemas de apoio ao raciocínio veterinário.

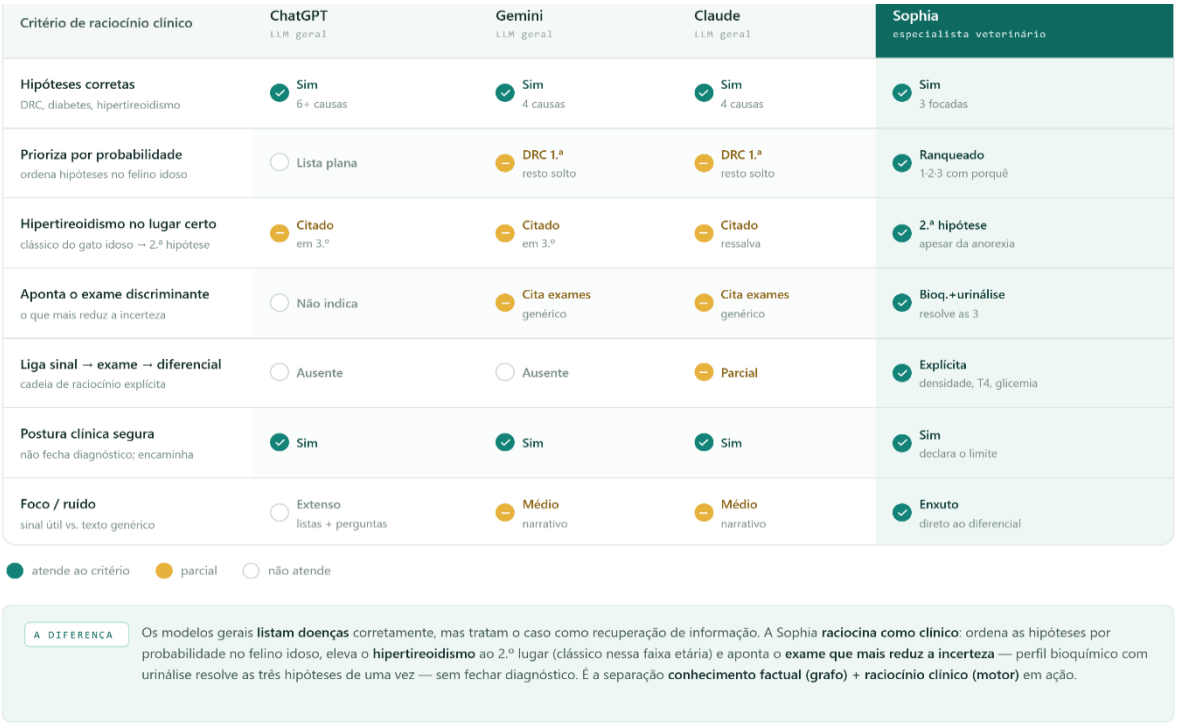

Figura 5 - Avaliação comparativa do raciocínio clínico entre modelos de linguagem generalistas (ChatGPT, Gemini e Claude) e a SophiaLLM em um caso veterinário envolvendo um gato idoso com anorexia, perda de peso e polidipsia. Embora todos os modelos tenham identificado hipóteses diagnósticas plausíveis, apenas a SophiaLLM apresentou raciocínio diferencial estruturado, com ranqueamento explícito das hipóteses por probabilidade clínica, posicionamento adequado do hipertireoidismo no contexto do caso, indicação de exames discriminantes e conexão direta entre sinais clínicos, exames complementares e hipóteses diagnósticas. Os resultados evidenciam a capacidade da SophiaLLM de integrar conhecimento factual e raciocínio clínico especializado para apoio à tomada de decisão em medicina veterinária.

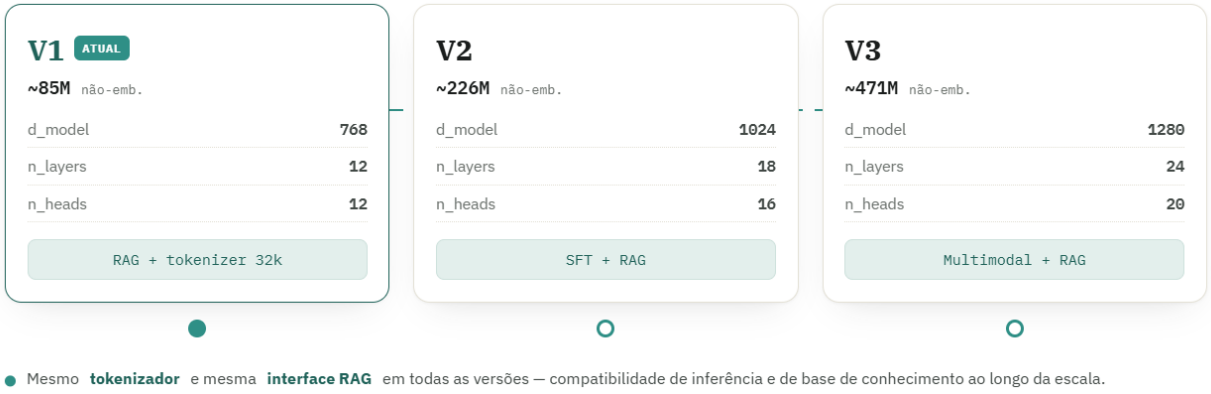

Figura 6 - Evolução planejada da SophiaLLM. Os valores apresentados correspondem aos parâmetros não-embedding do núcleo Transformer, excluindo a matriz de embeddings compartilhada. A versão V1 validada neste trabalho possui aproximadamente 85 milhões de parâmetros não-embedding e 109,5 milhões de parâmetros totais.

A principal estratégia adotada na SophiaLLM foi a separação entre conhecimento factual e conhecimento paramétrico. Em vez de concentrar todo o conhecimento clínico nos pesos do modelo, a arquitetura utiliza uma combinação de corpus especializado, recuperação de informação (RAG), grafo de conhecimento e motor de raciocínio clínico. Revisões recentes apontam que sistemas RAG apresentam vantagens importantes ao permitir acesso a fontes externas atualizadas, melhorando consistência factual e reduzindo erros decorrentes de conhecimento desatualizado (NEHA et al., 2025). Essa característica é particularmente relevante em medicina veterinária, área na qual protocolos diagnósticos e terapêuticos evoluem continuamente. Estudos recentes têm mostrado que a integração entre grafos de conhecimento e modelos de linguagem pode melhorar a rastreabilidade das inferências, favorecer raciocínio multi-etapas e aumentar a interpretabilidade dos sistemas de apoio à decisão; abordagens como MedRAG e DR.KNOWS demonstraram que a incorporação explícita de relações clínicas pode auxiliar na geração de hipóteses diagnósticas mais consistentes e justificáveis (ZHAO et al., 2025).

A escolha por uma arquitetura treinada desde sua concepção inicial também possui implicações importantes. Enquanto a maioria dos projetos atuais utiliza estratégias de fine-tuning sobre modelos fundacionais, a construção integral da SophiaLLM permitiu controle total sobre o vocabulário, a tokenização e a interface de recuperação de conhecimento. Embora a SophiaLLM V1 possua aproximadamente 109,5 milhões de parâmetros, número significativamente menor que os bilhões de parâmetros presentes em modelos generalistas modernos, sua base de conhecimento contém aproximadamente 500 milhões de tokens especializados, sugerindo que modelos menores podem obter desempenho competitivo em domínios restritos quando combinados com recuperação de conhecimento de alta qualidade e bases estruturadas cuidadosamente curadas.

Apesar dos resultados promissores, algumas limitações permanecem. A avaliação clínica apresentada neste trabalho é preliminar e baseada em um número reduzido de casos representativos. Estudos futuros deverão incluir benchmarks quantitativos, conjuntos independentes de validação e avaliação por médicos veterinários especialistas. Além disso, embora o corpus apresente ampla cobertura temática, áreas como oncologia, anestesiologia, ortopedia, reprodução e terapia intensiva ainda demandam aprofundamento da base curada para atingir o mesmo nível de maturidade observado em clínica médica, farmacologia e hematologia.

CONCLUSÃO

A SophiaLLM demonstra a viabilidade técnica de uma inteligência artificial especializada em medicina veterinária de pequenos animais em português brasileiro, integrando modelo de linguagem, recuperação de conhecimento, grafo veterinário e motor de raciocínio clínico. A validação da arquitetura, a construção de uma base de conhecimento ampla e a capacidade de organizar hipóteses diagnósticas de forma estruturada indicam que modelos menores, aliados a bases estruturadas, podem apoiar a educação continuada e a tomada de decisão baseada em evidências, oferecendo uma ferramenta auditável e alinhada à prática veterinária.

AGRADECIMENTOS

Os autores agradecem o apoio institucional recebido durante o desenvolvimento deste trabalho.

REFERÊNCIAS

BOLTON, E. et al. BioMedLM: a 2.7B parameter language model trained on biomedical text. Stanford: Stanford CRFM, 2022.

BRASIL. Senado Federal. Brasil tem terceira maior população pet do mundo. Brasília, 2024. Disponível em: https://www12.senado.leg.br/noticias/infomaterias/2024/12/brasil-tem-terceira-maior-populacao-pet-do-mundo-veja-os-projetos-do-senado-sobre-o-assunto. Acesso em: 5 jun. 2026.

HUR, B.; BALDWIN, T.; VERSPOOR, K.; HARDEFELDT, L.; GILKERSON, J. Domain adaptation and instance selection for disease syndrome classification over veterinary clinical notes. In: SIGBIOMED WORKSHOP ON BIOMEDICAL LANGUAGE PROCESSING (BioNLP), 19., 2020. Anais [...]. 2020. p. 156-166.

LUO, R. et al. BioGPT: generative pre-trained transformer for biomedical text generation and mining. Briefings in Bioinformatics, 2022.

NEHA, F.; BHATI, D.; SHUKLA, D. K. Retrieval-augmented generation (RAG) in healthcare: a comprehensive review. AI, v. 6, n. 9, p. 226, 2025.

RADFORD, A.; WU, J.; CHILD, R.; LUAN, D.; AMODEI, D.; SUTSKEVER, I. *Language models are unsupervised multitask learners*. [S. l.]: OpenAI, 2019.SINGHAL, K. et al. Large language models encode clinical knowledge. Nature, v. 620, n. 7972, p. 172-180, 2023.

TOUVRON, H. et al. LLaMA: open and efficient foundation language models. arXiv:2302.13971, 2023.

ZHAO, X. et al. MedRAG: enhancing retrieval-augmented generation with knowledge graph-elicited reasoning for healthcare copilot. In: ACM WEB CONFERENCE, 2025. Anais [...]. 2025. p. 4442-4457.