Extração Clínica em Texto Livre: Uma Análise Orientada pela Complexidade Semântica

Antônio Pereira[1], Lucas Pereira[1], Miriam Allein Zago Marcolino[2], Ricardo Bertoglio Cardoso[2], Luciana Lara[2], Ana Paula Beck da Silva Etges[2,3,4], Carisi A Polanczyk[2,3,4,5], Leonardo Rocha[1]

[1]Universidade Federal de São João del-Rei (UFSJ), Brasil

[2]Instituto para Avaliação e Translação em Saúde para Doenças Crônicas e Negligência (IATS-CARE), Brasil

[3]Hospital de Clínicas de Porto Alegre (HCPA), Brasil

[4]Universidade Federal do Rio Grande do Sul (UFRGS), Brasil

[5]Centro de Inovação em Inteligência Artificial para a Saude (CI-IA Saúde)

antonio258p@gmail.com, lucasmdpereira@gmail.com, zago.miriam13@gmail.com, ricardobcardoso@gmail.com, lucianarodriguesdelara@gmail.com, ana.etges.ext@accamargo.org.br, carisi.anne@gmail.com, lcrocha@ufsj.edu.br

Resumo. *Este trabalho investiga a extração automatizada de informações clínicas a partir de notas médicas não estruturadas em português, considerando quatro fases de complexidade semântica crescente: scores e tempos clínicos, medicamentos na alta, comorbidades e complicações. Comparamos métodos determinísticos, modelos generativos com prompt engineering, versões com fine-tuning e com reasoning. Os resultados mostram que a adequação do método depende da natureza da variável: tarefas mais simples podem ser resolvidas por abordagens léxicas, enquanto as mais ambíguas exigem modelos mais robustos. Esses achados reforçam a importância de analisar, conjuntamente, desempenho e custo computacional na estruturação de dados clínicos.*

Abstract. *This work investigates the automated extraction of clinical information from unstructured medical notes in Portuguese, considering four phases of increasing semantic complexity: clinical scores and time intervals; discharge medications; comorbidities; and complications. We compare deterministic methods, generative models with prompt engineering, fine-tuned versions, and reasoning-based models. The results show that the suitability of each method depends on the nature of the clinical variable: simpler tasks can be handled by purely lexical approaches, whereas tasks with greater contextual ambiguity require more robust models. These findings highlight the importance of jointly analyzing performance and computational cost when structuring clinical data.*

1. Introdução

Prontuários hospitalares contêm um grande volume de informações clínicas relevantes para pesquisa de apoio à decisão [Anschau et al. 2022]. Notas médicas textuais registram variáveis importantes, como *scores* prognósticos, tempos de atendimento, medicamentos prescritos na alta, comorbidades pré-existentes e complicações ocorridas durante a internação. Embora essas informações sejam valiosas para estruturar bases clínicas e

viabilizar análises posteriores, como predição de reinternação, óbito e outros desfechos, elas costumam estar registradas em linguagem livre, com abreviações, variações terminológicas, negações, ambiguidades contextuais e inconsistências de escrita.

A extração automática de informações clínicas a partir de texto livre constitui uma etapa central na estruturação de dados em saúde [Reading Turchioe et al. 2022, Zanotto et al. 2021]. Embora modelos generativos recentes tenham ampliado a capacidade de processar narrativas clínicas não estruturadas [Singhal et al. 2023], sua adoção suscita uma questão prática relevante: em que medida abordagens mais sofisticadas são, de fato, necessárias para cada tipo de variável? Em muitos cenários, o uso desses modelos implica um custo computacional elevado, sem ganhos proporcionais, especialmente em tarefas que poderiam ser tratadas por estratégias mais simples.

Este trabalho parte da premissa de que a adequação do método depende da complexidade semântica da variável clínica a ser extraída. Variáveis com representação mais regular podem ser tratadas por abordagens computacionais mais simples e rápidas, enquanto variáveis com maior variabilidade terminológica, dependência de contexto, negações ou relações temporais e causais tendem a exigir modelos mais robustos e mais custosos [Thirunavukarasu et al. 2023]. Assim, mais do que verificar se LLMs são úteis para extração clínica, investigamos em quais cenários seu uso se justifica.

Para isso, organizamos o problema em quatro fases de complexidade crescente, com base em notas médicas reais em português do Hospital de Clínicas de Porto Alegre. Avaliamos desde tarefas mais diretas, como a extração de *scores* clínicos e a classificação de categorias de medicamentos na alta, até tarefas semanticamente mais desafiadoras, como a identificação de comorbidades pré-existentes e complicações durante a internação. Ao longo dessas fases, comparamos estratégias determinísticas de similaridade textual, modelos generativos de pequeno porte com *prompt engineering*, versões adaptadas por *fine-tuning* e modelos com capacidade explícita de *reasoning*.

Os resultados mostram que não existe uma solução ótima para todas as variáveis: em tarefas de baixa complexidade semântica, métodos simples e transparentes alcançam desempenho competitivo a custo muito inferior, enquanto o aumento da variabilidade de representação e da inferência contextual torna os modelos especializados mais importantes. Isso sugere que a estruturação automática de dados clínicos pode se beneficiar de estratégias escalonadas, acionando diferentes métodos conforme a natureza da variável.

2. Trabalhos Relacionados

A extração automática de informações de textos clínicos tem sido explorada como etapa central para estruturar bases de dados em saúde, apoiar a pesquisa clínica e a predição de desfechos [Anschau et al. 2022, Zanotto et al. 2021]. Esses estudos mostram que variáveis relevantes podem ser recuperadas de prontuários eletrônicos por meio de diferentes estratégias computacionais, incluindo métodos supervisionados e modelos baseados em *transformers*. Recentemente, LLMs têm sido investigados nesse contexto devido à sua capacidade de lidar com linguagem não estruturada e tarefas que exigem maior inferência semântica [Agrawal et al. 2022, Singhal et al. 2023, Thirunavukarasu et al. 2023]. Em paralelo, esforços como o SemClinBr ampliaram a infraestrutura disponível para o processamento de notas clínicas em português, favorecendo o desenvolvimento de abordagens neurais nesse domínio [Oliveira et al. 2022].

Apesar dos avanços, a literatura tem se concentrado em avaliar a efetividade de modelos específicos em tarefas clínicas isoladas. Ainda permanece pouco discutido em que medida a complexidade semântica da variável extraída deve orientar a escolha entre métodos mais simples, estratégias de *prompt engineering*, modelos ajustados e abordagens mais robustas. Mais do que apresentar achados inéditos, este trabalho **corrobora** evidências consolidadas: sua contribuição é o **arcabouço de análise orientado pela complexidade semântica**, que organiza o problema em fases e investiga em quais cenários as abordagens sofisticadas se justificam, conjugando desempenho e custo computacional.

3. Abordagem Proposta

Partimos da premissa de que a extração de informações clínicas de notas médicas não constitui um problema homogêneo. Diferentes variáveis demandam níveis distintos de interpretação linguística, variando desde menções lexicalmente estáveis até situações que exigem desambiguação contextual, tratamento de negações e inferência temporal ou causal. Estruturamos o problema como uma sequência de tarefas com complexidade semântica crescente: a **Fase 1** contempla a extração de *scores* e tempos clínicos, isto é, variáveis numéricas ou temporais frequentemente expressas por acrônimos, abreviações e padrões relativamente recorrentes. A **Fase 2** trata da identificação de categorias de medicamentos prescritos na alta hospitalar, uma tarefa de reconhecimento de termos cuja principal dificuldade reside em isolar corretamente a seção relevante do prontuário. A **Fase 3** aborda a identificação de comorbidades pré-existentes, exigindo o reconhecimento de sinônimos, o tratamento de negações e a distinção entre condições crônicas e eventos agudos. A **Fase 4** considera a identificação de complicações ocorridas na internação, cenário em que a extração depende de relações temporais e causais mais complexas.

3.1. Modelos avaliados

Na **Fase 1**, comparamos três abordagens. A primeira é um método determinístico baseado em similaridade textual dual, denominado **DSA** (*Dual Similarity Algorithm*), que combina medidas de Levenshtein [Levenshtein 1966] e Jaccard para localizar menções às variáveis-alvo em textos não padronizados e extrair os valores associados a partir do contexto local. A segunda é o modelo generativo **Llama 3.1 8B-Instruct** [Meta AI 2024], empregado com *prompt engineering* e sem adaptação supervisionada, com o objetivo de verificar até que ponto um LLM de pequeno porte é capaz de extrair variáveis clínicas a partir do texto livre. A terceira corresponde à adaptação do modelo base **Llama 3.1 8B** [Meta AI 2024] por *fine-tuning* supervisionado (SFT) via **LoRA** [Hu et al. 2021, Dettmers et al. 2023], empregada para especializá-lo ao domínio clínico e lidar com variáveis que apresentam maior diversidade terminológica e múltiplas formas de expressão.

Na **Fase 2**, utilizamos uma pipeline baseada no modelo **Llama 3.1 8B-Instruct** [Meta AI 2024] em duas etapas. Primeiro, o modelo extrai da nota clínica completa apenas o trecho relativo às orientações de alta. Esse trecho é usado para classificar categorias farmacológicas predefinidas. Embora a tarefa final envolva reconhecimento de termos relativamente bem delimitados, o uso do modelo é necessário para separar a prescrição de alta de outras menções medicamentosas distribuídas ao longo do prontuário.

Nas **Fases 3** e **4**, comparamos quatro estratégias. A primeira é novamente o **DSA**, adaptado para classificação binária com tratamento explícito de negações. A

segunda é o modelo **Llama 3.1 8B-Instruct** [Meta AI 2024], utilizado em regime *zero-shot*, com *prompts* contendo instruções sobre presença, ausência e sinônimos. A terceira é o modelo **Llama 3.1 70B-Instruct** [Meta AI 2024], incluído para avaliar o impacto do aumento de escala em tarefas mais ambíguas. A quarta é o modelo **Qwen3 32B-AWQ** [Yang et al. 2025], empregado com capacidade explícita de *reasoning*, de modo a verificar se os traços intermediários de *reasoning* contribuem em casos nos quais a decisão depende de contexto, temporalidade ou causalidade.

A seleção equilibra custo e capacidade semântica, fundamentada na literatura: o **DSA** é um *baseline* transparente e sem GPU, competitivo em variáveis de expressão regular [Levenshtein 1966]; a família **Llama 3.1** [Meta AI 2024] isola especialização (uso via *prompt* [Agrawal et al. 2022] e *fine-tuning* eficiente com LoRA/QLoRA [Hu et al. 2021, Dettmers et al. 2023]) e escala (8B para 70B, viabilizada por quantização); e o **Qwen3 32B-AWQ** [Yang et al. 2025] testa se o *reasoning* explícito beneficia tarefas de maior inferência contextual [Thirunavukarasu et al. 2023], ao custo de inferência elevado.

3.2. Dados e tarefas

Os experimentos foram conduzidos com base em 2.258 notas médicas reais em português do Hospital de Clínicas de Porto Alegre (HCPA) de internações entre 2019 e 2021 para procedimentos cardiovasculares, anotadas independentemente por três especialistas clínicos, a fim de aumentar a confiabilidade do conjunto de referência e reduzir ambiguidades inerentes ao texto clínico não estruturado. Na **Fase 1**, avaliamos a extração de variáveis numéricas e temporais, incluindo *scores* prognósticos e tempos clínicos como **GRACE**, **NYHA**, **KILLIP**, **TIMI Fluxo**, **TIMI Risk**, **Delta T** e **Porta-balão** [Fox et al. 2006, Morrow et al. 2000]; nessa fase, o objetivo é recuperar corretamente o valor clínico associado a cada variável a partir da nota médica. Na **Fase 2**, investigamos a identificação de categorias de medicamentos prescritos na alta, como **antiagregantes**, **beta-bloqueadores**, **inibidores da ECA**, **estatinas** e **anticoagulantes**, após a extração da seção de orientações de alta, de modo que a tarefa consiste em classificar a presença ou ausência de categorias farmacológicas predefinidas nesse trecho. Na **Fase 3**, investigamos a identificação de comorbidades pré-existentes, formulada como uma tarefa de classificação binária por variável, incluindo **hipertensão**, **diabetes**, **tabagismo**, **obesidade**, **insuficiência renal crônica**, **AVC prévio** e **dislipidemia**. Por fim, na **Fase 4**, tratamos da identificação de complicações ocorridas durante a internação, também modelada como classificação binária, incluindo **IAM**, **arritmias**, **insuficiência renal aguda**, **sangramento**, **infecção hospitalar** e **insuficiência respiratória aguda**.

3.3. Estratégia de avaliação

Em todas as fases, os experimentos usaram validação cruzada estratificada em cinco *folds*. Na **Fase 1**, a avaliação foi binarizada em acerto ou erro na extração, pela correspondência entre o valor extraído e a anotação de referência; nas **Fases 2**, **3** e **4**, cada categoria, comorbidade ou complicação foi tratada como classificação binária independente. Reportamos *accuracy*, precisão, *recall* e F1-*Score*, com destaque para a média macro de F1, por sintetizar melhor o comportamento diante de classes desbalanceadas. Analisamos ainda o custo computacional (tempos de execução), de modo a relacionar qualidade da extração e viabilidade prática. Em cada fase, os modelos foram comparados por validação estatística com o teste de *Wilcoxon* e correção de *Bonferroni*.

4. Resultados

A Tabela 1 apresenta uma visão consolidada dos resultados obtidos nas quatro fases experimentais. Em conjunto, os resultados mostram que o desempenho dos métodos varia de forma consistente com a complexidade semântica da tarefa, reforçando a hipótese central deste trabalho: abordagens mais sofisticadas tendem a se justificar apenas quando a variabilidade de representação e a necessidade de inferência contextual aumentam.

Table 1. Síntese dos resultados por fase.

Fase	Tarefa	Melhor método	Resultado
Fase 1	Extração de *scores* e tempos clínicos	**Llama 3.1 8B + LoRA**	Melhor em 7/8 variáveis
Fase 2	Classificação de medicamentos na alta	**Llama 3.1 8B-Instruct**	Macro-F1 = 0,9231
Fase 3	Identificação de comorbidades	**Qwen3 32B-AWQ**	F1 médio = 0,8118
Fase 4	Identificação de complicações	**Llama 3.1 70B**	F1 médio = 0,6841

Na **Fase 1**, observamos que a adequação do método depende fortemente da regularidade lexical das variáveis. Em variáveis com formas de expressão mais estáveis, como **GRACE** e **NYHA**, o **DSA** já apresentou desempenho elevado, alcançando F1 de 0,987 e 0,967, respectivamente. No entanto, em variáveis com maior diversidade terminológica, como **KILLIP** e **Delta T**, abordagens mais sofisticadas tornaram-se claramente necessárias. Nessa fase, o melhor desempenho global foi obtido pelo **Llama 3.1 8B** adaptado por *fine-tuning* via **LoRA**, que apresentou o maior F1 em 7 das 8 variáveis analisadas, com o **DSA** apresentando valores bem próximos. As maiores diferenças ocorreram nas variáveis mais heterogêneas, como **KILLIP** (F1=0,962) e **Delta T** (F1=0,933), indicando que a adaptação supervisionada foi decisiva quando a rigidez lexical deixou de ser suficiente. Assim, para tarefas de menor complexidade semântica, métodos determinísticos podem ser muito eficazes e mais eficientes. Já em variáveis com maior variabilidade de representação, o *fine-tuning* passa a ser uma estratégia mais adequada.

Na **Fase 2**, o **Llama 3.1 8B-Instruct** alcançou um macro-F1 de 0,9231, com desempenho particularmente elevado em categorias como **antiagregantes**, **inibidores da ECA**, **estatinas** e **anticoagulantes**. O resultado indica que, uma vez corretamente isolada a seção de orientações de alta, um modelo generativo pequeno já é suficiente para tratar tarefas de reconhecimento de termos bem delimitados. Esse achado sugere que a principal dificuldade dessa fase não está na classificação farmacológica em si, mas na delimitação correta do trecho relevante do prontuário.

Na **Fase 3**, o melhor desempenho médio foi obtido pelo **Qwen3 32B-AWQ**, com F1 de 0,8118, seguido pelo **Llama 3.1 70B-Instruct** (0,8094), pelo **DSA** (0,7616) e pelo **Llama 3.1 8B-Instruct** (0,7248). Esse resultado indica que o aumento da capacidade inferencial torna-se mais relevante quando a tarefa exige uma interpretação contextual mais refinada. Ainda assim, o comportamento não foi uniforme entre as variáveis. O **DSA** permaneceu competitivo e, em alguns casos, superior em comorbidades com terminologia mais padronizada, como **hipertensão**, **tabagismo**, **obesidade** e **dislipidemia**. Por outro lado, variáveis como **infarto prévio** e **arritmia prévia** obtiveram maior benefício de modelos com maior capacidade de inferência. Esses resultados sugerem que a classificação de comorbidades ocupa uma posição intermediária: simples demais para justificar sempre o uso do modelo mais caro, mas complexa o suficiente para expor limitações de abordagens lexicais em parte das variáveis.

Na **Fase 4** o melhor desempenho médio foi obtido pelo **Llama 3.1 70B-Instruct**, com F1 de 0,6810, seguido do **Qwen3 32B-AWQ** (F1 0,4326). O **DSA** (0,3401) e o **Llama 3.1 8B-Instruct** (0,3404) apresentaram desempenho inferior, porém bastante

próximo entre si. Mesmo o melhor resultado da fase permaneceu relativamente baixo, o que reforça a dificuldade dessa tarefa. A identificação de complicações exige determinar se um evento ocorreu como consequência da internação ou do procedimento, o que demanda raciocínio temporal e causal mais rigoroso. Esses resultados indicam que o ganho obtido por modelos com *reasoning* explícito é real, mas não compensa a diferença de escala no número de parâmetros, mantendo as complicações como o cenário mais desafiador, no qual tanto a sofisticação do raciocínio quanto a capacidade bruta do modelo se mostram determinantes.

A Figura 1 ilustra o *trade-off* entre desempenho e custo computacional nas Fases 3 e 4. Na Fase 3, o **Qwen3 32B-AWQ** obteve os melhores resultados, mas a um custo substancialmente maior. Na Fase 4, o **Llama 3.1 70B-Instruct** apresentou o melhor desempenho em menor tempo do que o **Qwen3 32B-AWQ**. O **DSA** permaneceu rápido e competitivo em cenários semânticos mais simples.

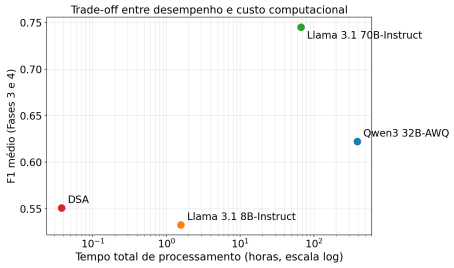

Figure 1. *Trade-off* entre desempenho e custo computacional nas Fases 3 e 4. O eixo x apresenta o tempo total de processamento em escala logarítmica, enquanto o eixo y mostra o F1 médio obtido por cada método.

5. Conclusão e Trabalhos Futuros

Este trabalho investigou a extração automatizada de informações clínicas a partir de notas médicas não estruturadas em tarefas com diferentes níveis de complexidade semântica. Em conjunto, os resultados mostram que não existe uma estratégia única adequada para todas as variáveis clínicas. Em tarefas mais simples, com expressão lexical mais estável ou categorias bem delimitadas, abordagens simples já se mostram suficientes. Já em cenários com maior variabilidade terminológica, ambiguidade contextual e necessidade de inferência temporal e causal, modelos mais robustos, especialmente os baseados em *fine-tuning* e *reasoning*, tornam-se mais vantajosos. Assim, a escolha do método deve considerar não apenas o desempenho obtido, mas também a natureza da variável e o custo computacional envolvido. Como trabalho futuro, pretendemos investigar arquiteturas em cascata que combinem métodos leves e modelos generativos especializados, bem como validar a abordagem em outros contextos clínicos.

Agradecimentos

Este trabalho foi apoiado por CI-IA Saúde, FAPEMIG, FAPESP, UNIMED Belo Horizonte, CNPq, Capes, AWS e IATS Care. Alguns experimentos deste trabalho utilizaram os recursos da infraestrutura PCAD, no INF/UFRGS.

Declaração de Uso de IA Generativa.

Ferramentas de IA foram utilizadas apenas como apoio à revisão gramatical; os autores revisaram todo o texto e assumem responsabilidade pela integridade das informações.

References

Agrawal, M., Hegselmann, S., Lang, H., Kim, Y., and Sontag, D. (2022). Large language models are few-shot clinical information extractors. In *Proceedings of the 2022 conference on empirical methods in natural language processing*, pages 1998–2022.

Anschau, F., Aredes, N. D. A., and others. (2022). Cohort study protocol of the brazilian collaborative research network on covid-19: strengthening who global data. *BMJ Open*, 12(11).

Dettmers, T., Pagnoni, A., Holtzman, A., and Zettlemoyer, L. (2023). QLoRA: Efficient finetuning of quantized language models. *arXiv preprint arXiv:2305.14314*.

Fox, K. A. A., Dabbous, O. H., Goldberg, R. J., Pieper, K. S., Eagle, K. A., Van de Werf, F., Avezum, Á., Goodman, S. G., Flather, M. D., Anderson, F. A., and Granger, C. B. (2006). Prediction of risk of death and myocardial infarction in the six months after presentation with acute coronary syndrome: prospective multinational observational study (GRACE). *BMJ*, 332(7553):1091–1094.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. (2021). LoRA: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8):707–710.

Meta AI (2024). The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Morrow, D. A., Antman, E. M., Charlesworth, A., et al. (2000). A new risk score for patients with acute coronary syndromes treated with percutaneous coronary intervention. *Journal of the American College of Cardiology*, 43(1):89–96.

Oliveira, L. E. S. e., Peters, A. C., Da Silva, A. M. P., Gebeluca, C. P., Gumiel, Y. B., Cintho, L. M. M., Carvalho, D. R., Al Hasan, S., and Moro, C. M. C. (2022). Semclinbr-a multi-institutional and multi-specialty semantically annotated corpus for portuguese clinical nlp tasks. *Journal of Biomedical Semantics*, 13(1):13.

Reading Turchioe, M., Volodarskiy, A., Pathak, J., Wright, D. N., Tcheng, J. E., and Slotwiner, D. (2022). Systematic review of current natural language processing methods and applications in cardiology. *Heart*, 108(12):909–916.

Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., et al. (2023). Large language models encode clinical knowledge. *Nature*, 620:172–180.

Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, K., Gutierrez, L., Tan, T. F., and Ting, D. S. W. (2023). Large language models in medicine. *Nature medicine*, 29(8):1930–1940.

Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., and et al. (2025). Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Zanotto, B. S., Beck da Silva Etges, A. P., Dal Bosco, A., et al. (2021). Stroke outcome measurements from electronic medical records: cross-sectional study on the effectiveness of neural and nonneural classifiers. *JMIR Medical Informatics*, 9(11):e29120.