Proposta de Arcabouço Teórico para a Avaliação Sistemática de Modelos de Linguagem Quantizados

Ricardo Leonarczyk[1,2]*, Murilo Regio[1,2], Cristiano Andrade[1,2], Luan Fonseca Garcia[1,2], Dalvan Griebler[2], Ewerton de Oliveira[3], Thomas Paula[3]

[1]Núcleo Avançado de Inteligência Artificial (NAIA)

[2]Escola Politécnica - Pontifícia Universidade Católica do Rio Grande do Sul (PUCRS)

[3]Brazil R&D - HP Inc.
Porto Alegre – RS – Brazil

Abstract. *This article proposes a conceptual framework to systematize the evaluation of quantized language models, organizing evaluation into four hierarchical levels of increasing scope and rigor. It is grounded in a systematic literature review of post-training quantization, whose findings also guide the selection of appropriate metrics and benchmarks for each level. By establishing a structured, evidence-based methodology, the framework is designed to enhance the transparency, reproducibility, and comparability of research in the field.*

Resumo. *Este artigo propõe um arcabouço conceitual para sistematizar a avaliação de modelos de linguagem quantizados, organizando-a em quatro níveis hierárquicos de escopo e rigor crescentes. Ele é fundamentado em uma revisão sistemática da literatura sobre quantização pós-treinamento, cujos resultados também orientam a seleção de métricas e benchmarks apropriados para cada nível. Ao estabelecer uma metodologia estruturada e baseada em evidências, o arcabouço visa aprimorar a transparência, a reprodutibilidade e a comparabilidade da pesquisa na área.*

1. Introdução

A proliferação de grandes modelos de linguagem (LLMs) impõe um grande desafio: seu alto custo computacional e de memória muitas vezes limita seu uso a hardware de ponta. Superar essa barreira é essencial para executar IA em dispositivos de consumo, como computadores e smartphones, um domínio conhecido como inferência local ou de borda (*edge*). Uma estratégia eficaz para isso é a quantização pós-treinamento (PTQ), que reduz a precisão numérica dos parâmetros de um modelo pré-treinado sem necessidade de retreino [Gholami et al. 2022]. Nela, os pesos do modelo são convertidos de uma representação de ponto flutuante de 32 bits ($FP32$) para um formato mais compacto, como inteiros de 8 bits ($INT8$). Esse método de compressão reduz o uso de memória e acelera a computação, tornando modelos sofisticados acessíveis sem depender de *data centers* remotos.

Contudo, avaliar modelos quantizados é um desafio, especialmente ao medir o impacto da quantização na qualidade do modelo. Métodos de avaliação restritos ou mal

*Autor correspondente: ricardo.leonarczyk@edu.pucrs.br

definidos podem levar a conclusões enganosas, onde um modelo parece ter sucesso em uma métrica, mas falha criticamente em outra. Por exemplo, a quantização pode melhorar o desempenho em sumarização, mas levar o modelo a alucinar fatos, tornando-o inadequado para uso geral. Além disso, uma avaliação abrangente, especialmente para modelos grandes, exige muitos recursos e tempo. Essas limitações forçam os pesquisadores a restringir o escopo de suas avaliações, gerando a necessidade de diretrizes para priorizar experimentos e garantir resultados significativos.

Para suprir essa necessidade, propomos neste estudo um arcabouço conceitual que auxilia pesquisadores e profissionais de *machine learning* a projetar e comunicar sistematicamente o escopo de suas avaliações. O arcabouço resulta de uma revisão sistemática da literatura (RSL) em andamento. Sua relevância está em oferecer uma abordagem estruturada e baseada na literatura para lidar com a natureza muitas vezes *ad-hoc* e inconsistente da avaliação de modelos quantizados. Ao definir níveis claros de avaliação (de Básico a Abrangente), ele estabelece um vocabulário comum que aumenta a transparência, a reprodutibilidade e a comparabilidade dos resultados na área.

2. Trabalhos Relacionados

Não identificamos outros estudos que proponham um arcabouço de níveis similar para avaliar a quantização em modelos de linguagem ou que fundamentem suas recomendações em uma RSL. Contudo, alguns estudos propõem protocolos de avaliação que abordam dimensões similares. Em seu trabalho sobre estratégias de quantização, [Jin et al. 2024] sugerem um arcabouço para analisar LLMs quantizados de forma abrangente, indo além de tarefas típicas como modelagem de linguagem ou classificação. Eles consideram as dimensões de conhecimento & capacidade, alinhamento e eficiência, recomendando benchmarks para cada uma com base em uma revisão não sistemática. Outros estudos focam na robustez da avaliação. [Zhao et al. 2025] propõe um novo benchmark chamado PTQ-Bench, que avalia a robustez de estratégias de PTQ em relação a diferentes larguras de bit, arquiteturas e modalidades.

3. Um Arcabouço de Avaliação para Modelos de Linguagem Quantizados

Nosso arcabouço estrutura a avaliação de modelos quantizados em dois aspectos primários: qualidade e eficiência. A **qualidade** abrange o conhecimento e as capacidades do modelo, seu alinhamento (utilidade, honestidade e segurança) e sua robustez a perturbações e ataques. A **eficiência** engloba ganhos de desempenho, como velocidade de inferência, e melhorias no uso de recursos, como a redução do consumo de memória e energia.

A Figura 1 ilustra nosso arcabouço hierárquico para a avaliação sistemática de modelos de *machine learning* quantizados. Ele se organiza em quatro níveis aninhados, representados como regiões concêntricas de escopo e rigor crescentes. Cada nível se baseia nos requisitos do anterior, criando uma taxonomia abrangente.

Após o nível de **Escopo Preliminar** (uma análise parcial), o nível de **Cobertura Básica** estabelece o padrão mínimo para uma avaliação válida, exigindo benchmarks padronizados e medidas de validade estatística (e.g., latência média). As avaliações de qualidade focam em conhecimento e capacidade, enquanto as métricas de eficiência visam a velocidade de inferência ou o consumo de recursos.

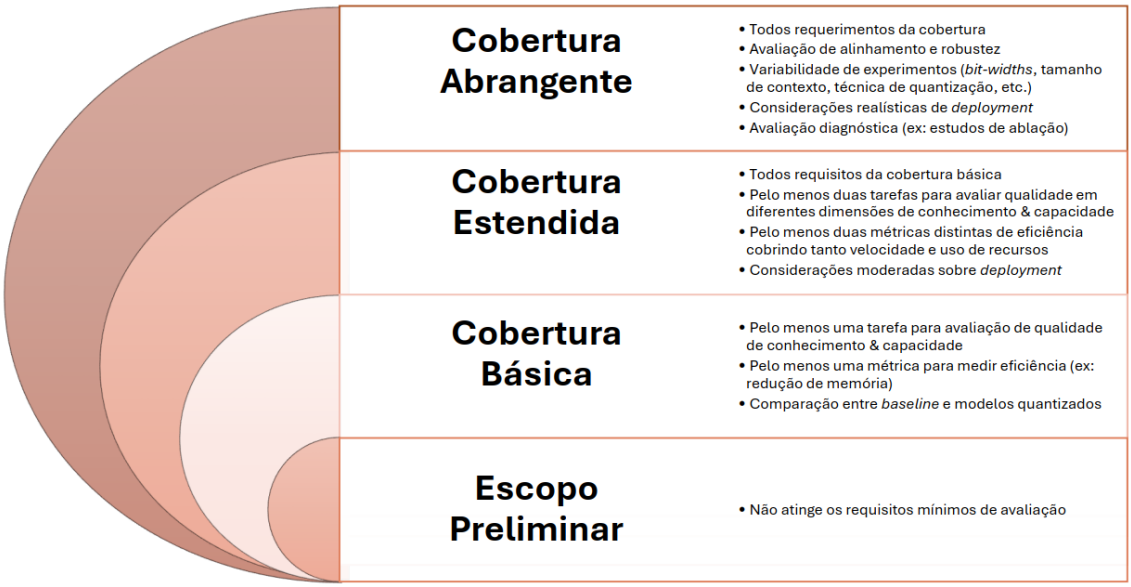

Figura 1. Camadas do Arcabouço de Avaliação

A **Cobertura Estendida** reconhece que modelos quantizados podem ter desempenho desigual entre capacidades — por exemplo, manter a performance em sumarização, mas perder conhecimento factual. Assim, este nível exige a avaliação de subdimensões distintas de qualidade e eficiência. Ele também introduz considerações práticas de implementação, como o impacto do uso de CPU na usabilidade do dispositivo durante a inferência.

A **Cobertura Abrangente** adiciona avaliações de alinhamento e robustez, além de variabilidade experimental sistemática. Esta última pode incluir análises em diferentes precisões de quantização, estudos de sensibilidade de parâmetros de quantização (variações de escala e ponto zero) e testes em múltiplas arquiteturas. Este nível enfatiza uma metodologia de diagnóstico com métodos orientados por hipóteses, como estudos de ablação, para explicar os padrões de desempenho e avançar de resultados descritivos para um entendimento dos mecanismos subjacentes.

Isso mantém a precisão taxonômica enquanto reconhece contribuições de pesquisa incrementais.

4. Resultados da revisão sobre quantização pós-treino em modelos de linguagem

O arcabouço de avaliação proposto na Seção 3 resulta de nossa RSL em andamento sobre quantização pós-treinamento para modelos de linguagem somente textuais ou multimodais. A RSL pode servir como referência para selecionar tarefas, métricas e benchmarks para cada nível do arcabouço. A metodologia da RSL partiu da análise de artigos seminais para definir as palavras-chave relevantes. A busca foi feita em sete bibliotecas digitais (incluindo Scopus e arXiv), resultando em 487 artigos únicos. A seleção teve três etapas: (1) aplicação de critérios de exclusão (e.g., publicações anteriores a 2019 ou com menos de quatro páginas) , (2) análise de escopo com base nos resumos e (3) uma análise de conteúdo detalhada a partir da leitura completa dos textos. Após as etapas, 97

artigos foram obtidos. A análise da literatura revela que, para sumarização e tradução, as métricas de n-grama como ROUGE e BLEU são predominantes. Em QA, a acurácia (base em número de respostas corretas) é a métrica mais utilizada. Um dos estudos utiliza a entropia relativa para comparar as diferenças de respostas entre os modelos base e quantizados. Alguns aspectos do alinhamento dos modelos são avaliados em conversas de múltiplos turnos, empregando benchmarks como o MT-Bench e o FollowBench em conjunto com um modelo avaliador ou juiz (LLM-as-a-judge). Para a robustez, benchmarks como o AdvGLUE são utilizados, com a acurácia sendo computada com base na queda de desempenho entre as versões adversarial e não-adversarial do benchmark.

5. Caso de Uso do Arcabouço de Avaliação

Um possível caso de uso é a avaliação de um modelo quantizado para sumarização (TinyLlama-1.1B[1]) para execução em um computador pessoal. Usando o arcabouço, um pesquisador visa garantir tanto a qualidade da sumarização quanto o baixo impacto na usabilidade do computador. Guiado pelo nível de Cobertura Estendida, ele usa a RSL para selecionar a métrica ROUGE e o benchmark CNN/DailyMail[2] na avaliação da sumarização, e mede a latência, o consumo de CPU e de memória para a eficiência. Como seu foco é apenas sumarizar textos, ele não testa uma segunda tarefa na dimensão de qualidade (assim deixando de completar a camada de avaliação estendida). Ao final, ele reporta ter realizado uma avaliação de Cobertura Básica com análises adicionais em eficiência e deployment.

6. Conclusão

Este estudo apresentou um arcabouço para classificar a abrangência da avaliação de modelos de linguagem quantizados. Também fornecemos alguns resultados de uma RSL em andamento que podem orientar a escolha de métricas e benchmarks ao usar o arcabouço. Como trabalhos futuros, planejamos publicar a RSL e estender o arcabouço para que seus níveis considerem tipos específicos de métricas.

Agradecimentos

Pesquisa financiada por HP Brasil Indústria e Comércio de Equipamentos Eletrônicos Ltda. utilizando incentivos fiscais de ressarcimento de IPI de acordo com a Lei (Lei nº 8.248 de 1991)

Referências

Gholami, A., Kim, S., Dong, Z., Yao, Z., Mahoney, M. W., and Keutzer, K. (2022). A survey of quantization methods for efficient neural network inference. In *Low-Power Computer Vision*, pages 291–326. Chapman and Hall/CRC.

Jin, R., Du, J., Huang, W., Liu, W., Luan, J., Wang, B., and Xiong, D. (2024). A comprehensive evaluation of quantization strategies for large language models. *ArXiv*, abs/2402.16775.

Zhao, J., Wang, M., Zhang, M., Shang, Y., Liu, X., Wang, Y., Zhang, M., and Nie, L. (2025). Benchmarking post-training quantization in llms: Comprehensive taxonomy, unified evaluation, and comparative analysis. *ArXiv*, abs/2502.13178.

[1]https://huggingface.co/TinyLlama

[2]https://github.com/abisee/cnn-dailymail