UNIVERSIDADE FEDERAL DE OURO PRETO

INSTITUTO DE CIÊNCIAS EXATAS E BIOLÓGICAS

DEPARTAMENTO DE COMPUTAÇÃO

BEATRIZ HELENA DE MELLO ORLANDI DE DEUS

GERAÇÃO SINTÉTICA DE DIÁLOGOS PROFISSIONAIS USANDO MODELOS DE LINGUAGEM E COM A INCORPORAÇÃO DE HISTÓRICO CURRICULAR E INDICADORES DE PERFORMANCE

Ouro Preto
2026

BEATRIZ HELENA DE MELLO ORLANDI DE DEUS

GERAÇÃO SINTÉTICA DE DIÁLOGOS PROFISSIONAIS USANDO MODELOS DE LINGUAGEM E COM A INCORPORAÇÃO DE HISTÓRICO CURRICULAR E INDICADORES DE PERFORMANCE

Monografia apresentada ao Curso de Ciência da Computação da Universidade Federal de Ouro Preto como requisito parcial para obtenção do título de Bacharel em Ciência da Computação.

Orientador: Reinaldo Silva Fortes
Coorientador: Alexandre Fortes Santana

Ouro Preto
2026

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE OURO PRETO
REITORIA
INSTITUTO DE CIENCIAS EXATAS E BIOLOGICAS
DEPARTAMENTO DE COMPUTACAO

FOLHA DE APROVAÇÃO

Beatriz Helena de Mello Orlandi de Deus

Geração Sintética de Diálogos Profissionais usando Modelos de Linguagem e com a Incorporação de Histórico Curricular e Indicadores de Performance

Monografia apresentada ao Curso de Ciência da Computação da Universidade Federal de Ouro Preto como requisito parcial para obtenção do título de Bacharel em Ciência da Computação

Aprovada em 2 de Março de 2026

Membros da banca

Reinaldo Silva Fortes (Orientador) - Doutor - Universidade Federal de Ouro Preto
Alexandre Fortes Santana (Coorientador) - Bacharel - Efí Bank
Valéria de Carvalho Santos (Examinadora) - Doutora - Universidade Federal de Ouro Preto
Daniel José Chaves Ferreira (Examinador) - Bacharel - PPGCC / UFOP

Reinaldo Silva Fortes, orientador do trabalho, aprovou a versão final e autorizou seu depósito na Biblioteca Digital de Trabalhos de Conclusão de Curso da UFOP em 02/03/2026

Documento assinado eletronicamente por **Reinaldo Silva Fortes**, **PROFESSOR DE MAGISTERIO SUPERIOR**, em 03/03/2026, às 09:40, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site http://sei.ufop.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0 , informando o código verificador **1062478** e o código CRC **1CF13CF2**.

R. Diogo de Vasconcelos, 122, - Bairro Pilar Ouro Preto/MG, CEP 35402-163
Telefone: 3135591692 - www.ufop.br

À minha mãe, dona do mais belo sorriso, de personalidade intensa e coração grande, que me ensinou que amar é sempre excessivo. Ao meu pai que em seu modo firme de existir, me mostrou que cuidado também pode ter mãos calejadas e que nelas há o abraço mais seguro que já conheci. Aos meus irmãos, Daniel e Danilo, que me mostraram, em cada brincadeira de criança, que a vida exige coragem para aceitar os tombos. À minha irmã Bárbara, minha Flor do Dia, minha primeira referência de mundo. Foi com você que aprendi quase tudo, inclusive a arte de irritar com precisão e, ainda assim, permanecer amada como se o amor fosse sempre maior que qualquer provocação. Ao meu noivo Samuel, amigo de infância e amor de toda a vida, por caminhar ao meu lado com ternura e apoio incondicional, acreditando em mim até quando eu mesma duvidei. E a Deus, que permanece, mesmo quando tudo em mim oscila. A cada um de vocês, entrego estas páginas como quem devolve o que recebeu, porque aprendi com todos, que sonhar é uma construção feita de muitos corações.

Agradecimentos

Aos professores, que com tanto zelo e dedicação me ensinaram até chegar a esta etapa final, deixo minha eterna gratidão. Em especial, ao meu orientador Reinaldo, pelo apoio, paciência e confiança em meu trabalho. À UFOP e ao IFMG, que me mostraram que o conhecimento abre infinitos caminhos e possibilidades, e aos professores do IFMG, que me apresentaram à programação e me incentivaram a ingressar na graduação. Às minhas escolas de ensino fundamental, Laura Queiroz e Raul Soares, por despertarem em mim, ainda tão cedo, a vontade de aprender. E à minha avó Sônia, que, ao lado da minha mãe, me ensinou a ler e a escrever nas tardes que guardo com tanto carinho na memória.

Ao Efí Bank, por me proporcionar um espaço de crescimento, aprendizado e realização. Em especial ao Alexandre e à Danielle, que me inspiram a cada dia e acreditaram no meu potencial. À tríade dos J's — Júlia, Josué e João — que, além de colegas de estágio, tornaram-se grandes amigos, dividindo comigo, desafios e conquistas. À Júlia (Agrê), Leonardo (Nubreu) e Mateus (Preso), companhias insubstituíveis em tantos trabalhos e estudos que atravessaram madrugadas, semestre após semestre, me lembrando que o processo se torna mais leve quando não estamos sozinhos.

Se fosse possível escrever, um a um, os nomes de todos que contribuíram para que eu chegasse até aqui, estes agradecimentos se estenderiam até a última página desta monografia. A todos vocês, registro de coração: nada disso seria possível sem o apoio, a amizade e o amor que recebi ao longo desta jornada chamada vida.

"Os limites da minha linguagem significam os limites do meu mundo."
Wittgenstein (1922). Tractatus Logico-Philosophicus.

Resumo

Avaliar atributos profissionais, como desempenho, produtividade e habilidade, a partir de interações textuais é um desafio central para organizações que operam em ambientes digitais. Contudo, a escassez de dados reais de comunicação corporativa, aliada a restrições de privacidade, limita o avanço de pesquisas nessa direção. Este trabalho propõe uma abordagem experimental para a construção e avaliação de um *corpus* sintético de conversas empresariais, gerado por modelos de linguagem de grande escala e condicionado a variáveis latentes ocupacionais extraídas de históricos de carreira reais. O método desenvolvido integra simulação estatística, formação de times por similaridade de trajetória profissional e engenharia estruturada de *prompt*, produzindo diálogos que refletem comportamentos associados a diferentes perfis, sem revelar explicitamente os atributos no texto. A avaliação combina validação estrutural, rubricas de qualidade e um protocolo de inferência inversa, no qual os modelos avaliadores tentam recuperar os atributos latentes apenas a partir da leitura das conversas. Os resultados demonstram que o cenário estruturado supera consistentemente a linha de base aleatória, confirmando que modelos de linguagem conseguem codificar e recuperar parcialmente sinais comportamentais em diálogos sintéticos. A análise revela, ainda, um viés sistemático em direção a avaliações positivas por parte dos inferidores e uma tensão entre a naturalidade do texto e a controlabilidade dos atributos, abrindo caminho para o aprimoramento de métodos de geração e avaliação de dados sintéticos corporativos.

Palavras-chave: Dados sintéticos. Diálogos corporativos. Inferência inversa.

Abstract

Assessing professional attributes such as performance, productivity, and ability from textual interactions is a central challenge for organizations operating in digital environments. However, the scarcity of real corporate communication data, combined with privacy constraints, limits research progress in this direction. This work proposes an experimental approach for building and evaluating a synthetic corpus of corporate conversations, generated by large language models and conditioned on occupational latent variables derived from real career histories. The developed method integrates statistical simulation, team formation based on professional trajectory similarity, and structured prompt engineering, producing dialogues that reflect behaviors associated with different profiles without explicitly revealing the attributes in the text. The evaluation combines structural validation, quality rubrics, and an inverse inference protocol in which evaluator models attempt to recover the latent attributes solely from reading the conversations. Results show that the structured scenario consistently outperforms the random baseline, confirming that language models can partially encode and recover behavioral signals in synthetic dialogues. The analysis further reveals a systematic positivity bias among inferrers and a tension between text naturalness and attribute controllability, opening avenues for improving synthetic corporate data generation and evaluation methods.

Keywords: Synthetic data. Corporate dialogues. Inverse inference.

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas e Siglas

API *Application Programming Interface*, Interface de Programação de Aplicações

AR(1) *Autoregressive process of order 1*, Processo autorregressivo de ordem 1

BERT *Bidirectional Encoder Representations from Transformers*, Modelo de Representações de Codificador Bidirecional baseado em Transformers

BLEU *Bilingual Evaluation Understudy*, Métrica automática de avaliação de qualidade de texto gerado

CSR *Compressed Sparse Row*, Representação esparsa de matriz

ESCO *European Skills, Competences, Qualifications and Occupations*

FWER *Family-Wise Error Rate*, Taxa de Erro Familiar

GAN *Generative Adversarial Network*, Rede Generativa Adversária

GPG *Guided Profile Generation*, Geração Condicionada a Perfis

IDF *Inverse Document Frequency*, Frequência Inversa de Documento

JSON *JavaScript Object Notation*, Formato de serialização estruturada para troca de dados

LLM *Large Language Model*, Modelo de Linguagem de Grande Escala

MAE *Mean Absolute Error*, Erro Absoluto Médio

MCC *Matthews Correlation Coefficient*, Coeficiente de Correlação de Matthews

MCSE *Monte Carlo Standard Error*, Erro Padrão de Monte Carlo

NLU *Natural Language Understanding*, Compreensão de Linguagem Natural

Parquet Formato colunar para armazenamento eficiente e leitura em lote

PLN Processamento de Linguagem Natural

RH Recursos Humanos

SDG *Synthetic Data Generation*, Geração de Dados Sintéticos

SGD *Schema-Guided Dialogue*, Diálogo Guiado por Esquema

TF-IDF *Term Frequency-Inverse Document Frequency*, Métrica que pondera a importância de termos em documentos

VAE *Variational Autoencoder*, *Autoencoder* Variacional

Weighted Jaccard Variação ponderada do índice de Jaccard

Sumário

1 Introdução

Nos últimos anos, os avanços em Processamentos de Linguagem Natural (PLNs) e *Natural Language Understandings*, Compreensão de Linguagem Natural (NLUs) foram intensificados pelo surgimento dos *Large Language Models*, Modelos de Linguagem de Grande Escala (LLMs), modelos capazes de compreender e gerar linguagem natural com elevada coerência e adaptabilidade contextual (BROWN et al., 2020). Esses modelos tornaram-se essenciais para aplicações em diversas áreas, incluindo atendimento ao cliente, sistemas de recomendação e assistentes virtuais. Contudo, seu desempenho é fortemente influenciado pela qualidade, diversidade e domínio dos dados utilizados para treinamento (GOYAL; MAHMOUD, 2024).

No contexto organizacional, a comunicação corporativa apresenta particularidades como formalidade, uso de jargões específicos e dinâmicas hierárquicas, o que torna desafiador o desenvolvimento de modelos especializados sem dados contextualizados (CAMILLERI, 2021). Entretanto, a indisponibilidade de bases públicas com diálogos reais limita significativamente o progresso de pesquisas nessa direção, em virtude de restrições legais, éticas e logísticas (YANG; ISLAM, 2020).

A *Synthetic Data Generation*, Geração de Dados Sintéticos (SDG), notadamente com uso de LLMs, desponta como alternativa viável para suprir essa lacuna. A criação de corpora sintéticos permite simular interações autênticas sem violar a privacidade de dados sensíveis (JORDON et al., 2022). No entanto, a geração de diálogos com realismo linguístico, coerência pragmática e diversidade funcional requer métodos rigorosos de modelagem e validação (LIU et al., 2023).

Diante desse cenário, esta monografia propõe uma metodologia para geração e avaliação de uma base sintética de conversas empresariais, utilizando LLMs para produzir interações entre personas corporativas com perfis profissionais distintos. A abordagem inclui a simulação controlada de variáveis latentes ocupacionais (desempenho, produtividade e habilidade), a formação de times por similaridade de trajetória profissional e a avaliação do *corpus* por meio de inferência inversa, na qual modelos avaliadores tentam recuperar os atributos latentes a partir do texto gerado (ABDULLIN et al., 2023).

No restante deste capítulo serão apresentadas justificativas para o trabalho, na Seção 1.1, os objetivos do trabalho, na Seção 1.2 e, por fim, a organização deste documento, na Seção 1.3.

1.1 Justificativa

A ausência de conjuntos de dados públicos representativos das interações internas em corporações constitui uma limitação expressiva para o avanço da pesquisa em PLN aplicada a

contextos empresariais. A natureza sensível das informações, aliada à dificuldade de obtenção de consentimento e à exigência de anonimização, restringe o uso de dados reais em larga escala (KAUR et al., 2021; JARKE; BREITER, 2020).

A adoção de SDG, especialmente com LLMs, viabiliza a criação de dados sintéticos fidedignos, desde que acompanhada de estratégias de validação e controle de qualidade. Estudos recentes destacam que, com modelagem adequada, é possível gerar textos sintéticos com alto grau de realismo, respeitando parâmetros linguísticos e pragmáticos do domínio-alvo (FERRARA, 2023; SOUDANI; HASIBI; KANOULAS, 2024).

Este trabalho contribui para esse campo ao propor uma metodologia replicável e parametrizada de geração sintética, com base em perfis profissionais extraídos de fontes estruturadas. Os perfis são derivados da base *Karrierewege* (SENGER et al., 2024), que contém trajetórias profissionais reais, permitindo a criação de personas consistentes com a realidade corporativa. A formação de times por similaridade de trajetória textual, utilizando técnicas de *Term Frequency-Inverse Document Frequency*, Métrica que pondera a importância de termos em documentos (TF-IDF) e similaridade de Jaccard ponderada, aproxima a estrutura de interações em canais corporativos (KUMMERFELD et al., 2019).

A base gerada possibilita o treinamento e a avaliação de aplicações específicas, como sistemas de análise de sentimento em mensagens internas, detecção de riscos de evasão de talentos, progressão de carreira e assistentes conversacionais corporativos (ZHENG et al., 2023b).

Por fim, a presente monografia também busca colaborar com a pesquisa em estruturação de dados sintéticos a partir de atributos funcionais e relacionais dos membros de times, considerando habilidades técnicas, competências interpessoais e comportamentos colaborativos conforme identificado na literatura sobre times organizacionais (KOZLOWSKI; ILGEN, 2006).

1.2 Objetivos

Este trabalho tem como objetivo geral a construção e avaliação de um *corpus* sintético de conversas empresariais, gerado por LLMs, que represente de forma coerente e diversificada as interações entre colaboradores em distintos perfis e contextos organizacionais, com variáveis latentes controláveis e recuperáveis por inferência inversa.

Para alcançar este objetivo geral, definem-se os seguintes objetivos específicos:

1. **Metodologia de simulação de variáveis latentes:** uma abordagem parametrizada para simulação controlada de atributos ocupacionais (desempenho, produtividade e habilidade) por experiência profissional, com validação estatística por replicação Monte Carlo e estimação de erro *Monte Carlo Standard Error*, Erro Padrão de Monte Carlo (MCSE) por cenário.

2. **Taxonomia de personas corporativas:** um conjunto validado de perfis profissionais derivados da base *Karrierewege* (SENGER et al., 2024), com atributos funcionais e comunicacionais quantificáveis, agregados por pessoa para condicionamento da geração de diálogos.

3. **Mecanismo de formação de times por similaridade textual:** agrupamentos de personas baseados em similaridade de trajetória profissional, utilizando representação TF-IDF e similaridade de Jaccard ponderada, com bandas de afinidade calibradas por percentis (MANNING; RAGHAVAN; SCHÜTZE, 2008).

4. **Protocolo reprodutível de geração condicionada:** um *pipeline* de engenharia de *prompt* estruturado em etapas (planejamento, geração e validação), com dossiê de projeto fixo por time, roteiro de execução replicável e conversão de rótulos latentes em descritores comportamentais sem vazamento explícito.

5. **Framework de avaliação multinível do *corpus*:** um protocolo de avaliação em três níveis: validação estrutural, avaliação de qualidade por rubricas com LLM e inferência inversa dos atributos latentes, com métricas de classificação (acurácia, F1 macro, *Matthews Correlation Coefficient*, Coeficiente de Correlação de Matthews (MCC) e comparação entre modelos avaliadores de diferentes provedores.

1.3 Organização do Trabalho

Esta monografia está estruturada em cinco capítulos, apresentados a seguir:

Capítulo 2: Revisão Bibliográfica - Discute os principais conceitos sobre geração de dados sintéticos, diálogos orientados a tarefas, trajetórias profissionais e avaliação de texto gerado por LLMs, bem como trabalhos correlatos.

Capítulo 3: Metodologia - Detalha o processo de simulação de variáveis latentes, formação de times por similaridade de trajetória, geração de diálogos com LLMs via engenharia de *prompt* e o protocolo de avaliação multinível.

Capítulo 4: Resultados e Discussão - Apresenta os resultados da inferência inversa, incluindo métricas de classificação, comparação entre cenários e provedores, análise de viés, testes estatísticos e qualidade do *corpus*.

Capítulo 5.1: Conclusão - Sintetiza as contribuições, discute limitações e propõe direções para trabalhos futuros.

2 Revisão Bibliográfica

Este capítulo examina, de forma crítica, a literatura e os recursos empíricos indispensáveis à construção de uma *base sintética de conversas corporativas*. A discussão está organizada em seis eixos: (i) metodologias de Geração de Dados Sintéticos (SDG), (ii) enriquecimento de dados e fusão de fontes heterogêneas, (iii) representação semântica e recuperação de informação, (iv) estrutura organizacional e de times, (v) bases de dados correlacionadas ao domínio de Recursos Humanos (RH) e comunicação empresarial e (vi) geração de diálogos com Grandes Modelos de Linguagem (LLMs). Essa estrutura evidencia avanços, lacunas e fundamenta a metodologia proposta.

A presente revisão bibliográfica aprofunda cada um desses eixos de forma integrada, estruturando o debate a partir de avanços metodológicos, desafios e oportunidades identificados na literatura especializada. Inicialmente na Seção 2.1, explora-se o avanço das metodologias de **geração de dados sintéticos**, desde abordagens estatísticas tradicionais e arquiteturas profundas, como *Generative Adversarial Network*, Rede Generativa Adversária (GAN) e *Variational Autoencoder*, *Autoencoder* Variacional (VAE), até o emprego dos LLMs, ressaltando desafios de avaliação em termos de *realismo*, *diversidade* e mitigação de vieses e "alucinações" (GOYAL; MAHMOUD, 2024; JI; LEE et al., 2023). Em seguida na Seção 2.3 , discute-se o **enriquecimento de perfis** por meio da **fusão de fontes heterogêneas** 2.2, evidenciando como repositórios como o *Karrierewege* (SENGER et al., 2024) podem ancorar as *personas* em históricos autênticos, ampliando a fidelidade factual dos dados gerados (BLEIH; BELAID, 2016). Complementarmente, aborda-se a **representação semântica** e os mecanismos de **recuperação de informação** - desde *TF-IDF* até *embeddings* contextuais de modelos como o *Bidirectional Encoder Representations from Transformers*, Modelo de Representações de Codificador Bidirecional baseado em Transformers (BERT) e a aplicação de **busca vetorial** para correspondência semântica precisa entre perfis e diálogos (MANNING; RAGHAVAN; SCHÜTZE, 2008; DEVLIN et al., 2019).

Na segunda parte, examina-se a **modelagem das estruturas organizacionais e de times** na Seção 2.4, fundamentada em distinções conceituais entre *grupo*, *equipe* e *time*, e na aplicação de técnicas de **clusterização** sobre o *dataset* **IBM HR Analytics Employee Attrition e Performance** para delineamento de *perfis arquetípicos* Mello, Arima e Neves (2020), Yang e Islam (2020). Prossegue-se na seção 2.5 com a **análise crítica de bases correlacionadas**, do *Corpus* de E-mails Enron a *datasets* de assistentes virtuais (*Schema-Guided Dialogue*, Diálogo Guiado por Esquema (SGD), *MSDialog*), mapeando suas contribuições e lacunas em termos de cobertura de interações corporativas multi-departamentais (KLIMT; YANG, 2004; RASTOGI et al., 2020a). Por fim, a seção 2.6 dedica-se à **geração de diálogos com LLMs**, enfatizando práticas avançadas de **engenharia de *prompt*** condicionada a perfis e métricas de coerência e aderência ao contexto, conforme proposto em abordagens como *Guided Profile Generation*

Zhang et al. (2024), Soudani, Hasibi e Kanoulas (2024).

2.1 Metodologias de Geração de Dados Sintéticos

Jordon et al. (2022) definem *synthetic data* como dados gerados por um modelo ou algoritmo desenvolvido com o propósito explícito de suportar tarefas de ciência de dados. Em termos operacionais, dados sintéticos consistem em registros produzidos artificialmente para replicar propriedades estatísticas e estruturais de conjuntos reais - com a finalidade de viabilizar investigação, desenvolvimento de modelos e validação sem recorrer diretamente a amostras sensíveis oriundas do mundo real.

A SDG evoluiu de modelos estatísticos para arquiteturas profundas como GAN e VAE. Entretanto, as abordagens baseadas em LLMs despontam como o estado da arte para dados textuais, por permitirem um controle granular sobre estilo, tom e coerência de domínio (GOYAL; MAHMOUD, 2024). A qualidade dos dados gerados resulta do equilíbrio entre *realismo*, *diversidade* e *representatividade*. *Frameworks* modernos recomendam métricas automáticas (ex.: *BLEU*, *ROUGE*, *BARTScore* e *Distinct-n*) e avaliações humanas para aferir naturalidade e utilidade (SOUDANI; HASIBI; KANOULAS, 2024). Estudos recentes também alertam para vieses induzidos pelo treinamento e o risco de "alucinação" - geração de informações factualmente incorretas -, destacando a necessidade de métricas padronizadas e estratégias de mitigação (JI; LEE et al., 2023).

2.2 Enriquecimento de Dados e Fusão de Fontes Heterogêneas

Uma limitação inerente à SDG puramente generativa é a dependência exclusiva do conhecimento paramétrico do LLM, o que pode comprometer a fidelidade factual dos dados gerados. Para mitigar este risco, a literatura recente explora técnicas de **fusão de dados**, que combinam informações de múltiplas fontes para criar um repositório sintético mais rico e ancorado na realidade (BLEIH; BELAID, 2016).

Esta abordagem é particularmente relevante para a criação de perfis ou "*personas*". Em vez de solicitar a um LLM que invente um histórico profissional, pode-se enriquecer um perfil com dados de carreira autênticos. O trabalho seminal de Park et al. (2023) sobre **agentes generativos** demonstra a eficácia desta abordagem. A nossa metodologia adota um princípio análogo, utilizando uma base de dados de carreiras sintéticos, como o ***Karrierewege*** (SENGER et al., 2024), para enriquecer perfis derivados de uma base de dados corporativa. Esta fusão entre dados quantitativos (performance e demografia) e dados textuais não estruturados (histórico de carreira) visa gerar *personas* com um nível de realismo e profundidade inatingível por métodos puramente generativos.

2.3 Representação Semântica e Recuperação de Informação

A fusão eficaz de dados textuais depende da capacidade de medir a similaridade de significado entre documentos. A Ciência da Computação aborda este desafio através da **Recuperação de Informação** (MANNING; RAGHAVAN; SCHÜTZE, 2008). Métodos tradicionais baseados em contagem de palavras são limitados por não capturarem a semântica.

A revolução dos transformadores, notavelmente com o modelo BERT (DEVLIN et al., 2019), introduziu os *embeddings* contextuais. Um *embedding* é uma representação vetorial densa de um texto num espaço de alta dimensionalidade, onde a distância entre vetores reflete a similaridade semântica. A técnica de **Busca Vetorial**, fundamentada na similaridade de cosseno, é particularmente robusta para identificar currículos e perfis com maior aderência semântica, superando as limitações da busca por palavras-chave.

2.4 Estrutura de Cargos e Times

Mello, Arima e Neves (2020) distingue *grupos* de *times*, ressaltando que estes últimos envolvem interdependência de tarefas, metas compartilhadas, papéis coordenados e fronteiras sociais claras; em contextos organizacionais, os papéis estabelecem expectativas sobre responsabilidades, autoridade e padrões de desempenho, orientando a coordenação e a comunicação entre os membros. Compreender "time" e "papel" vai além da delimitação de posições hierárquicas, exigindo o reconhecimento de uma arquitetura de interdependências que torna as atribuições legíveis e previsíveis - base conceitual robusta para modelar perfis profissionais e as relações que os articulam (KOZLOWSKI; ILGEN, 2006; KAHN; KATZ; JACOBS, 1964).

Sob perspectiva semântica, títulos ocupacionais funcionam como âncoras para conjuntos de tarefas, conhecimentos e habilidades; assim, taxonomias ocupacionais padronizadas oferecem vocabulário controlado e relações semânticas úteis para identificar ocupações interligadas ou sinônimos sistemáticos. A *European Skills, Competences, Qualifications and Occupations* (ESCO) organiza ocupações com termos preferenciais e não preferenciais, além de vincular competências, facilitando a expansão consistente de termos e a redução de ambiguidade terminológica (European Commission, 2025). Por sua vez, o modelo de conteúdo do O*NET estabelece domínios descritivos - como conhecimentos, habilidades, atividades de trabalho e contexto organizacional - que permitem inferir proximidade funcional entre cargos com títulos distintos (O*NET Resource Center, 2025; HANDEL, 2016). Tais taxonomias fornecem uma base conceitual sólida para o mapeamento de responsabilidades e famílias ocupacionais, mesmo se não incorporadas diretamente à metodologia empírica.

Para ilustrar a topologia densa e a formação de comunidades em redes corporativas reais, a Figura 2.1 apresenta o grafo da rede de *e-mails* da Enron (vide Seção 2.5.3), um *corpus* público amplamente utilizado em pesquisas de comunicação organizacional e ciência de redes.

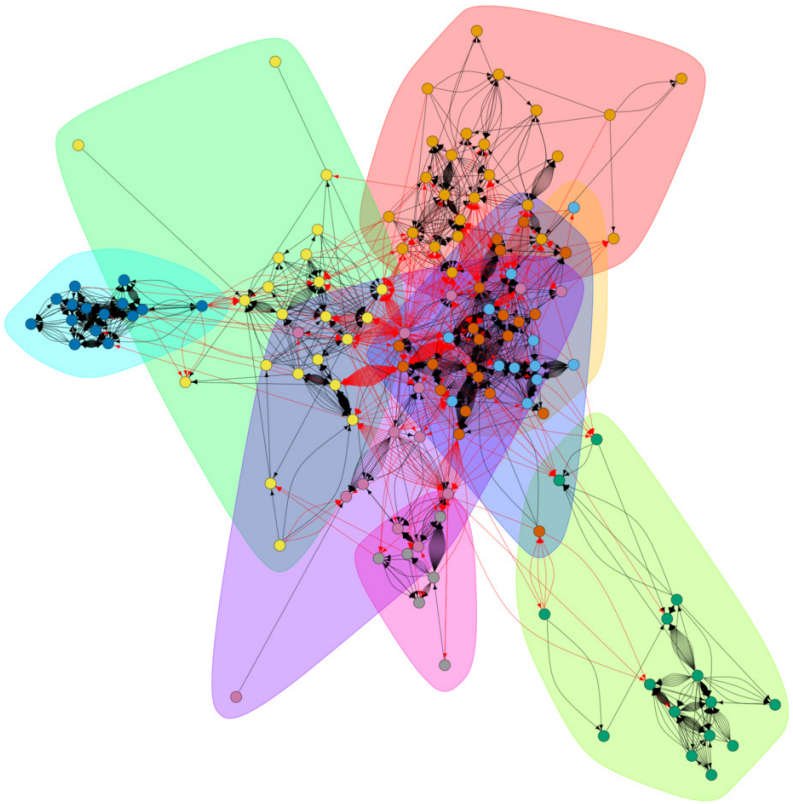

Figura 2.1 – Rede de *e-mails* da Enron com comunidades (cores). *Autor*: Peter Prevos. *Fonte*: Wikimedia Commons. Licença: CC BY-SA 4.0.

2.5 Bases de Dados Correlacionadas

A construção de um *corpus* sintético robusto requer uma análise prévia dos recursos de dados existentes para identificar os atributos disponíveis e as lacunas a serem preenchidas. A seguir apresenta-se, organizada em quatro categorias complementares, a natureza dos dados, seu escopo de aplicação e as limitações em termos de diálogo corporativo multi-departamental, destacando como essas fontes podem ser integradas ou suplementadas para suprir as exigências de realismo, variedade de contexto e profundidade conversacional.

2.5.1 *Datasets* de Perfis Profissionais e de RH

Estes datasets oferecem ricas informações sobre atributos de funcionários, mas carecem de dados de diálogo.

- ***IBM HR Analytics Employee Attrition e Performance*** (S; DESHPANDE; SCIENTISTS, 2023): Um *dataset* canónico com 1.470 registos sintéticos, detalhando demografia, cargo, nível hierárquico, satisfação, performance e rotatividade.

- ***OpenResume*** (YAMASHITA; DOM; PUROHIT, 2024): Um repositório em larga escala com 75.000 currículos reais e anonimizados. Os dados são estruturados em trajetória profissional, competências e educação.
- ***Karrierewege*** (SENGER et al., 2024; SENGER et al., 2025): Um *dataset* público em larga escala, com com mais de 500 mil trajetórias de carreira, vinculado à taxonomia **ESCO** para normalização ocupacional. Além das sequências temporais de ocupações, a coleção expõe *job titles* normalizados, **descrições ocupacionais** e listas de **skills** associadas via ESCO.

2.5.2 *Datasets* Conversacionais

Plataformas de gestão de tarefas, como o **Jira**, são fontes comuns de dados de conversação técnica. Diversos conjuntos públicos foram extraídos de repositórios de código aberto. Por exemplo, Zhang et al. (2023) apresentam um *dataset* com 16 repositórios públicos, contendo cerca de 2,7 milhões de *issues*, 9 milhões de comentários e 32 milhões de alterações, devidamente anonimizados. Em outro estudo, Alenezi, Banitaan e Maletic (2013) compila dados de comunidades como *Apache*, *Spring*, *JBoss* e *CodeHaus*, totalizando mais de 700 mil relatórios e 2 milhões de comentários. Esses recursos, embora acessíveis publicamente, são restritos ao domínio de desenvolvimento de software, não contemplando a diversidade temática de interações corporativas.

2.5.3 *Corpora* de *E-mail* Corporativo

O ***Corpus* de *E-mails* Enron** (KLIMT; YANG, 2004) contém mais de 600 mil mensagens trocadas entre funcionários da Enron, disponibilizado publicamente após investigações regulatórias. Apesar de amplamente utilizado para estudos de redes sociais internas e análise de sentimentos, seu conteúdo mistura comunicações profissionais, pessoais e logísticas, apresentando pouca representatividade de discussões estruturadas sobre projetos.

2.5.4 *Datasets* de Assistentes e Tarefas

Corpora como o **Schema-Guided Dialogue (SGD)** (RASTOGI et al., 2020a) e o **MSDialog** (FENG et al., 2020) contêm dezenas de milhares de conversas, mas focadas em interações orientadas a tarefas com assistentes virtuais. Embora relevantes para modelagem de sistemas de diálogo, não representam a complexidade da comunicação colaborativa entre pares humanos no contexto corporativo.

2.5.5 Síntese dos *datasets* correlacionados

A análise da literatura e dos *datasets* disponíveis revela uma lacuna: não existe, publicamente, uma base de dados que integre perfis profissionais com histórico de carreira e métricas

de performance com o conteúdo textual de suas interações conversacionais em um ambiente corporativo multi-departamental. Os recursos existentes são especializados: ou em perfis de RH e carreira profissional (*IBM HR Analytics Employee Attrition e Performance* (YANG; ISLAM, 2020)*OpenResume* e *Karrierewege*(Hugging Face, 2025)), ou em diálogos orientados a tarefas com assistentes (*SGD* e *MSDialog*), ou em conversas de um domínio técnico único (Jira) como o apresentado por Alenezi, Banitaan e Maletic (2013).

2.6 Geração de Diálogos com Grandes Modelos de Linguagem

Segundo Jurafsky e Martin (2023) e Gao, Galley e Li (2019), **geração de diálogos** consiste em produzir, de forma automática, *sequências multi-turno* de enunciados condicionadas ao histórico conversacional, aos objetivos comunicativos e, quando pertinente, a conhecimento externo e perfis de participantes; essa formulação abrange *chatbots* de domínio aberto, sistemas orientados a tarefas e cenários de busca de informação, com requisitos de **coerência global**, **consistência pragmática** e **gestão de contexto**.

Em alinhamento com essa definição, a literatura recente sistematiza componentes da geração *multi-turn* - criação de dados semente, produção de enunciados e *filtros* de qualidade e sintetiza critérios de avaliação como *adequação semântica*, **consistência de *persona***, **factualidade** e **diversidade lexical/estrutural**; nessas revisões, LLMs destacam-se como mecanismo dominante para a síntese de diálogos, desde que apoiados por *prompts* estruturados e mecanismos de ancoragem ao contexto (SOUDANI; HASIBI; KANOULAS, 2024).

No paradigma *orientados a tarefas*, o objetivo é completar metas concretas (p. ex., reservar um recurso, consultar um serviço, atualizar um registro) sob *restrições* explícitas (*slots*, valores e políticas). A literatura organiza o problema em componentes clássicos: compreensão de linguagem, *dialogue state tracking*, política e realização de linguagem e também em abordagens *end-to-end*. Como bases de referência, destacam-se **MultiWOZ** (multi-domínio, anotações de estado/atos, métricas como *Inform*, *Success*, *Bilingual Evaluation Understudy*, Métrica automática de avaliação de qualidade de texto gerado (BLEU)) e SDG (esquemas dinâmicos de *Application Programming Interfaces*, Interfaces de Programação de Aplicações (APIs) e *zero-shot* para novos serviços), frequentemente avaliados com *toolkits* como o **ConvLab**, que padronizam treinamentos e diagnósticos. (BUDZIANOWSKI et al., 2018; RASTOGI et al., 2020b; ZHU et al., 2020; BALARAMAN; SHEIKHALISHAHI; MAGNINI, 2021)

No presente trabalho, adota-se um **paradigma condicional** em que múltiplos LLMs de ponta (GPT-5.1, Gemini 2.5 Pro e Claude Haiku 4.5) geram *turnos* de diálogo a partir de **dossiês de *persona*** (performance, trajetória e competências) e de **cenários comunicacionais** próprios do ambiente corporativo; a **engenharia de *prompt*** explicita papéis, objetivos e restrições de forma a maximizar **coerência interturnos** e **aderência ao papel** (OPENAI, 2026; GOOGLE, 2026; ANTHROPIC, 2026).

Para operacionalizar o condicionamento aos perfis, emprega-se *Guided Profile Generation*, Geração Condicionada a Perfis (GPG), procedimento em que o modelo sintetiza um *perfil linguístico* intermediário, antes de produzir os enunciados; essa etapa melhora a incorporação de contexto pessoal/profissional e a consistência do comportamento gerado Zhang et al. (2024).

2.7 Testes Estatísticos para Comparação de Classificadores

A avaliação rigorosa de modelos de classificação exige não apenas o cálculo de métricas de desempenho, mas também a aplicação de testes estatísticos que permitam determinar se as diferenças observadas entre classificadores são estatisticamente significativas ou se decorrem de variação amostral. Esta seção apresenta os fundamentos dos testes empregados no presente trabalho.

2.7.1 Teste de McNemar

O teste de *McNemar*, proposto originalmente por McNemar (1947), é um teste não paramétrico para dados pareados em tabelas de contingência 2×2. No contexto de comparação de classificadores, o teste avalia se dois modelos cometem erros em observações distintas, focando exclusivamente nos pares discordantes, isto é, nas observações em que um classificador acerta e o outro erra. Seja b o número de observações em que o classificador A acerta e o classificador B erra, e c o número de observações em que B acerta e A erra. A estatística de teste é dada por:

$$\chi^2 = \frac{(b-c)^2}{b+c}, \tag{2.1}$$

que segue, sob a hipótese nula de desempenho equivalente, uma distribuição χ^2 com um grau de liberdade.

Dietterich (1998) realizou uma análise comparativa de cinco testes estatísticos aproximados para comparação de algoritmos de aprendizado supervisionado, avaliando a probabilidade de erro tipo I (rejeição incorreta da hipótese nula) e o poder estatístico de cada teste. O estudo demonstrou que o teste de *McNemar* apresenta taxa de erro tipo I adequada e poder estatístico superior ao teste t pareado de 5×2 *cross-validation* e ao teste t de validação cruzada com k *folds*, sendo recomendado como procedimento padrão para comparação pareada de classificadores em um único conjunto de dados.

2.7.2 Correção de Holm-Bonferroni para comparações múltiplas

Quando múltiplas comparações são realizadas simultaneamente, a probabilidade de cometer ao menos um erro tipo I (rejeição falsa) aumenta com o número de testes, fenômeno conhecido como inflação da *Family-Wise Error Rate*, Taxa de Erro Familiar (FWER). A correção de Bonferroni clássica controla a FWER dividindo o nível de significância α pelo número de

comparações k, porém é reconhecidamente conservadora, reduzindo o poder estatístico (HOLM, 1979).

Holm (1979) propôs um procedimento sequencialmente rejeitor que oferece controle da FWER com maior poder estatístico do que a correção de *Bonferroni*. O método ordena os p-valores das k comparações em ordem crescente ($p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(k)}$) e compara cada $p_{(i)}$ com o limiar $\alpha/(k-i+1)$. A rejeição prossegue sequencialmente até que um p-valor não atinja o limiar correspondente, momento em que todas as hipóteses restantes são retidas. Esse procedimento é uniformemente mais poderoso que a correção de Bonferroni e não requer independência entre os testes.

Demšar (2006) sistematizou a aplicação de testes não paramétricos para comparação de classificadores em aprendizado de máquina, recomendando o uso do teste de Friedman com *post-hoc* de *Nemenyi* para comparações de múltiplos classificadores em múltiplos conjuntos de dados, e o teste de *Wilcoxon* com correção de *Holm-Bonferroni* para comparações pareadas. O trabalho consolidou a prática de aplicar correções para comparações múltiplas em estudos experimentais de aprendizado de máquina, evitando conclusões espúrias decorrentes da multiplicidade de testes.

2.7.3 *Bootstrap* para intervalos de confiança

O *bootstrap*, introduzido por Efron e Tibshirani (1993), é uma técnica de reamostragem que permite estimar a distribuição amostral de uma estatística sem depender de suposições paramétricas sobre a distribuição dos dados. O procedimento consiste em gerar B amostras de mesmo tamanho que a amostra original, obtidas por amostragem com reposição, e calcular a estatística de interesse em cada reamostra. O intervalo de confiança percentílico de $(1-\alpha)\times 100\%$ é obtido pelos quantis $\alpha/2$ e $1-\alpha/2$ da distribuição *bootstrap*.

No contexto de avaliação de classificadores, o *bootstrap* é particularmente útil para construir intervalos de confiança para métricas como acurácia e *macro F1*, permitindo avaliar a incerteza das estimativas e determinar se o desempenho de um classificador é estatisticamente distinguível de um *baseline* de referência, como a predição aleatória (EFRON; TIBSHIRANI, 1993; DIETTERICH, 1998).

2.8 Síntese da Revisão Bibliográfica

A presente revisão bibliográfica evidencia um conjunto de achados que fundamentam a metodologia proposta. Em primeiro lugar, confirma-se que LLMs constituem a abordagem de vanguarda para SDG textual, oferecendo fluência e controle condicional superiores aos modelos estatísticos e às arquiteturas profundas tradicionais, ainda que imponham riscos de *alucinação* e vieses que demandam estratégias de mitigação e avaliação específicas (GOYAL; MAHMOUD, 2024; JI; LEE et al., 2023). Em segundo lugar, a **fusão de fontes heterogêneas**, como na integração de *datasets* distintos, mostra-se essencial para ancorar perfis e enriquecer

atributos narrativos das *personas* (BLEIH; BELAID, 2016). Terceiro, técnicas de **representação semântica** baseadas em *embeddings* contextuais e busca vetorial são o procedimento mais robusto para o *matching* entre perfis estruturados e trechos textuais, superando abordagens por contagem de termos e ampliando a precisão do processo de recuperação (DEVLIN et al., 2019).

As evidências apontam que a ***Guided Profile Generation*, Geração Condicionada a Perfis** e a **engenharia de *prompt*** estruturada aumentam a consistência de *persona* e a coerência interturnos, desde que complementadas por mecanismos de ancoragem factual e por avaliações automáticas e humanas de qualidade Zhang et al. (2024), Soudani, Hasibi e Kanoulas (2024).

Além disso, as evidências bibliográficas justificam: (i) priorizar modelos de geração textual conditionais e ancorados em fontes externas; (ii) empregar recuperação semântica para integração de atributos; (iii) utilizar métricas automáticas consolidadas, como *BLEU*, *ROUGE*, *BARTScore*, *Distinct-n* e medidas de factualidade e análises quantitativas de coerência e diversidade para avaliar a qualidade dos diálogos sintéticos; e (iv) aplicar testes estatísticos não paramétricos, como o teste de *McNemar* com correção de *Holm-Bonferroni* e intervalos de confiança por *bootstrap*, para fundamentar comparações entre classificadores com rigor estatístico (DIETTERICH, 1998; DEMŠAR, 2006; EFRON; TIBSHIRANI, 1993). Estas conclusões fundamentam as escolhas metodológicas explicitadas no Capítulo 3.

3 Desenvolvimento

Este capítulo apresenta a metodologia desenvolvida para construção e avaliação do *corpus* sintético de diálogos corporativos. A estrutura está organizada em cinco seções: visão geral do *pipeline* (Seção 3.1), preparação e consolidação dos dados de carreira (Seção 3.2), definição e simulação de cenários e variáveis latentes (Seção 3.3), formação de times e agrupamento por trajetória (Seção 3.4), formação de times e agrupamento por trajetória (Seção 3.4), e o protocolo de geração e avaliação com LLMs (Seção 3.5).

3.1 Visão Geral da Arquitetura

O *pipeline* desenvolvido opera em três camadas principais: estrutural, gerativa e avaliativa, conforme ilustrado na figura 3.1. A camada estrutural é responsável pela consolidação dos dados de carreira, simulação de variáveis latentes e formação de times. A camada gerativa executa a geração de diálogos condicionados aos atributos latentes. A camada avaliativa realiza a validação estrutural, atribuição de rubricas de qualidade e inferência inversa dos atributos.

O objetivo central é construir um *corpus* de diálogos sintéticos que seja controlável (variáveis latentes condicionam o comportamento dos participantes sem vazamento explícito no texto), auditável (cada execução registra cenário, semente, modelo, versão e parâmetros) e que permita avaliação inversa, ou seja, um LLM avaliador tenta inferir os atributos latentes apenas a partir do texto.

A arquitetura modular permite substituição de componentes como provedores LLM, métricas e cenários sem alteração do fluxo principal. Todos os artefatos intermediários são persistidos em formato *JavaScript Object Notation*, Formato de serialização estruturada para troca de dados (JSON) ou Formato colunar para armazenamento eficiente e leitura em lote (Parquet), garantindo reprodutibilidade e rastreabilidade. A figura 3.1 ilustra a visão geral da arquitetura do *pipeline*.

3.2 Preparação Consolidação dos Dados

3.2.1 Fontes de Dados

A base de dados é formada pela união do *dataset* Karrierewege (SENGER et al., 2024) (*split* de teste com 247.000 registros) com o Karrierewege_plus, processada via DuckDB para eficiência. O Karrierewege contém históricos profissionais reais com informações de cargos, empresas, datas e habilidades, codificados segundo a taxonomia ESCO.

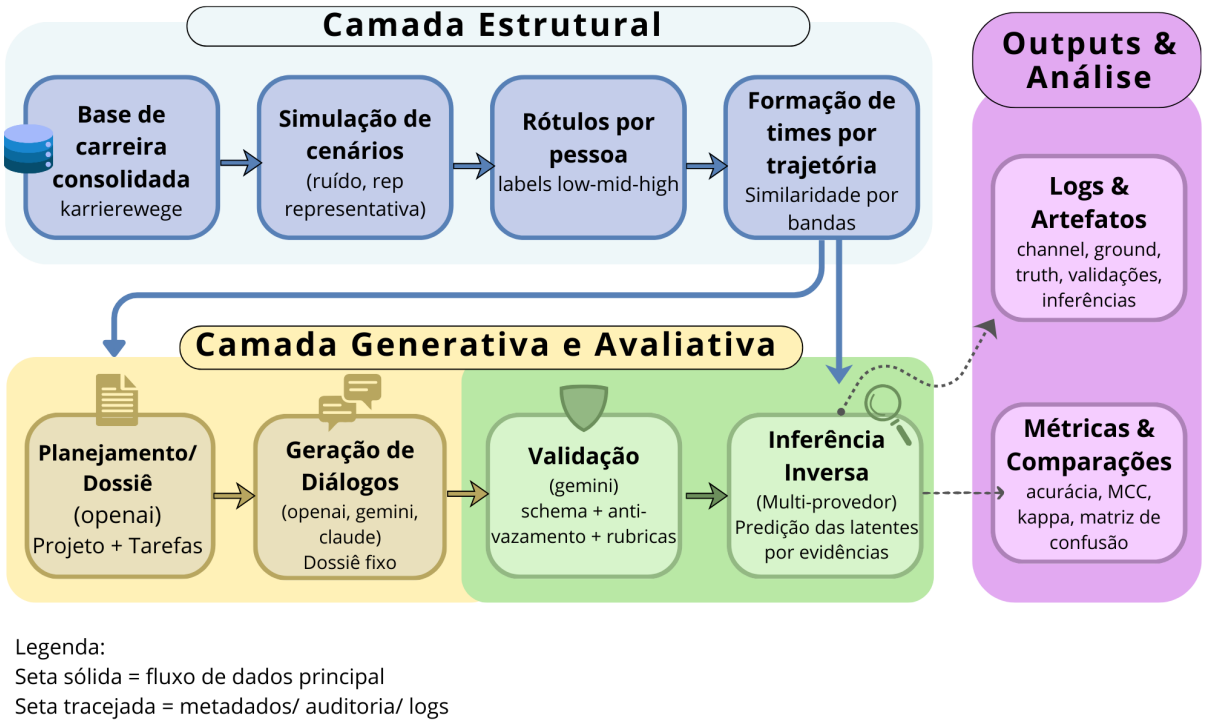

Figura 3.1 – visão geral da arquitetura para construção de corpus sintético controlável e auditável, com avaliação estrutural e inferência inversa multi-provedor.

3.2.2 Processo de Consolidação

O processo de consolidação segue cinco etapas: (i) leitura dos arquivos Parquet das fontes originais; (ii) `UNION ALL` dos *splits* do Karrierewege_plus; (iii) deduplicação por chave composta `(_id, experience_order, preferredLabel_en)`; (iv) `LEFT JOIN` com a base principal; e (v) materialização em Parquet ordenado por `(_id, experience_order)`.

A ordenação por identificador e ordem de experiência é requisito para suportar processamento incremental e componentes temporais, como o processo autorregressivo *Autoregressive process of order 1*, Processo autorregressivo de ordem 1 (AR(1)) aplicado por indivíduo na simulação de variáveis latentes.

3.2.3 Estrutura dos Perfis

Cada perfil de participante contém: identificador único, sequência de experiências profissionais (cargo, empresa, período, descrição), conjunto de habilidades (*skills*) e metadados de trajetória. Os cargos são representados por seus títulos ESCO em inglês, permitindo comparação semântica entre trajetórias.

3.3 Definição e Simulação de Cenários

Com a base consolidada, definiu-se um conjunto de cenários para simular variáveis latentes. O processo opera em dois níveis: primeiro, geram-se valores para cada experiência profissional (linha); depois, agregam-se esses valores por pessoa para obter um rótulo final.

A seguir, detalham-se os componentes de cada variável latente (Subseção 3.3.1),o procedimento de agregação por pessoa (Subseção 3.3.2) e os cenários comparados (Subseção 3.3.3).

3.3.1 Variáveis latentes e fatores de geração

Para cada experiência, foram geradas três variáveis latentes: `ability_latent` (habilidade), `performance_latent` (desempenho) e `productivity_latent` (produtividade). Essas variáveis combinam: traço individual A_i (fixo por `_id`), ajuste por trajetória (`skill_fit`), senioridade (via `experience_order`), estabilidade, efeito de time M_{team}, termo temporal E_t, ruído ϵ controlado por σ, e componente serial $U_{i,t}$ gerado por AR(1) com parâmetro ϕ.

As latentes foram transformadas em escalas interpretáveis: scores em 0-100 via sigmoide (desempenho e habilidade), produtividade positiva via *softplus*, e classes em 3 níveis (LOW/MID/HIGH) por discretização com limiares τ_1, τ_2.

A interpretação dos três níveis apoia-se na literatura de gestão de pessoas e economia do trabalho:

- **Desempenho**: adotou-se a perspectiva multidimensional de Rotundo e Sackett (2002), que decompõe o desempenho global em contribuições de tarefa, cidadania organizacional e comportamentos contraproducentes.
  - *LOW*: entregas incompletas, retrabalho frequente e dependência de terceiros para cumprir atribuições básicas.
  - *MID*: cumprimento regular das atribuições com qualidade adequada, sem desvios significativos.
  - *HIGH*: contribuições que excedem as expectativas, com iniciativa, impacto positivo no time e comportamentos de cidadania organizacional.
- **Produtividade**: seguiu-se a definição de Syverson (2011), que a trata como a eficiência na conversão de insumos em produtos.
  - *LOW*: baixo volume de entregas ou ciclos longos de execução em relação aos pares.
  - *MID*: ritmo de entrega compatível com a média do grupo, sem gargalos evidentes.
  - *HIGH*: *throughput* elevado com manutenção de qualidade, encaminhamentos rápidos e capacidade de absorver demandas adicionais.

- **Habilidade**: utilizou-se a tipologia de Winterton, Deist e Stringfellow (2006), que distingue conhecimentos, habilidades técnicas e competências comportamentais.
  - *LOW*: domínio limitado do repertório necessário ao cargo, com dúvidas frequentes sobre conceitos fundamentais.
  - *MID*: proficiência funcional com lacunas pontuais, capaz de executar tarefas rotineiras com autonomia.
  - *HIGH*: domínio amplo do repertório técnico e comportamental, com capacidade de síntese, transferência entre contextos e orientação de pares.

3.3.2 Agregação por pessoa

Para cada pessoa, calculou-se a **média das variáveis latentes** ao longo de todas as suas experiências profissionais. A tabela `per_id` resultante contém: médias das latentes e dos scores, probabilidades empíricas de classe (proporção de experiências em cada nível), e o rótulo final por dimensão (obtido por *argmax* das probabilidades). Esse rótulo agregado é o *ground truth* utilizado para condicionar a geração de diálogos e avaliar a inferência inversa.

3.3.3 Cenários utilizados

Foram definidos seis cenários com $R = 25$ replicações cada, variando presença de latentes, nível de ruído (σ) e aleatoriedade. Entretanto, para a geração de diálogos, utilizaram-se apenas dois cenários:

- `S_full_high_noise`**:** cenário completo com variáveis latentes estruturalmente correlacionadas à trajetória e alto ruído ($\sigma = 1{,}4$, $\phi = 0{,}6$).
- `S_random_with_latents`**:** *baseline* com latentes geradas aleatoriamente, sem correlação estrutural com a trajetória.

A escolha de apenas dois cenários justifica-se pela análise prévia das replicações: a comparação entre cenários com e sem ruído (e.g., `S_full_low_noise` vs. `S_full_high_noise`) não produziu alterações nas categorias dos rótulos agregados por pessoa. Isso ocorre porque a agregação por média suaviza as flutuações introduzidas pelo ruído, embora os valores por experiência variem, a média ao longo da trajetória converge para valores similares, resultando nos mesmos rótulos discretizados (LOW/MID/HIGH). Assim, optou-se por manter apenas o cenário de alto ruído (mais desafiador) e o aleatório (para validação por contraste). A Tabela 3.1 resume os seis cenários definidos, indicando com $\star$ os dois efetivamente utilizados na geração de diálogos.

A estabilidade das estimativas foi verificada por meio do erro padrão de Monte Carlo (MCSE), definido pela Equação 3.1.

Tabela 3.1 – Cenários de simulação.

Cenário	**Latentes**	σ	ϕ	**Aleatória**
`S_full_high_noise`*	sim	1.4	0.6	sim
`S_random_with_latents`*	sim	1.0	0.6	sim
`S_full_low_noise`	sim	0.6	0.6	não
`S_no_latents_low_noise`	não	0.6	0.0	não
`S_no_latents_high_noise`	não	1.4	0.0	não
`S_random_no_latents`	não	1.0	0.0	não

$$\mathrm{MCSE}(\bar{x}_m) = \frac{s_m}{\sqrt{R}}, \tag{3.1}$$

onde $\bar{x}_m$ é a média da métrica m ao longo das R replicações, s_m é o desvio padrão amostral dessa métrica entre replicações e R é o número total de replicações ($R = 25$).

O MCSE quantifica a incerteza da estimativa de Monte Carlo: valores pequenos indicam que a média observada é estável e que novas replicações não alterariam substancialmente o resultado. As métricas-chave atingiram estabilidade (MCSE absoluto pequeno), confirmando que $R = 25$ replicações são suficientes para garantir estimativas confiáveis dos cenários simulados.

3.3.4 Seleção de repetição representativa

Para gerar diálogos condicionados por um conjunto coerente de perfis, selecionou-se uma repetição representativa por cenário via *medoid*: padroniza-se um vetor de métricas por repetição (z-score), calcula-se a distância ao centro do cenário, e seleciona-se a repetição de menor distância. Esse procedimento preserva a consistência interna do *dataset* e facilita a auditoria.

3.4 Formação de Times por Trajetória Textual

A formação de times visa aproximar a estrutura de interações corporativas em que coortes compartilham contexto ocupacional e histórico de carreira antes de aplicar LLMs na geração de mensagens. Em canais corporativos, é comum haver múltiplas conversas paralelas no mesmo espaço, organizadas em tópicos (*threads*) e retomadas; portanto, definir um "time" por similaridade de trajetória fornece um contexto plausível para a geração posterior (KUMMERFELD et al., 2019).

A seguir, detalham-se a representação textual da trajetória (Subseção 3.4.1), a calibragem por percentis (Subseção 3.4.2) e o algoritmo de agrupamento (Subseção 3.4.3).

3.4.1 Representação da trajetória

Para reduzir o impacto de termos genéricos, a trajetória foi representada como um documento textual com os últimos K cargos (ex.: $K = 10$). Em seguida, aplicou-se TF-IDF - técnica que pondera termos pela frequência no documento e pela raridade no corpus, reduzindo o peso de palavras muito comuns com:

- **n-gramas (1,2)** para capturar expressões compostas (ex.: "gerente de produto");
- **max_df** para remover termos que aparecem em uma fração grande dos indivíduos (típicos de cargos genéricos);
- **min_df** para remover ruído raríssimo;
- matriz esparsa no formato *Compressed Sparse Row*, Representações CSR (CSRs) para eficiência.

A similaridade entre duas pessoas foi definida por Variação ponderada do índice de Jaccard (Weighted Jaccard), usando pesos IDF:

$$\mathrm{WJ}(i,j) = \frac{\sum_{t \in T_i \cap T_j} w_t}{\sum_{t \in T_i \cup T_j} w_t}, \tag{3.2}$$

onde: T_i é o conjunto de termos (unigramas/bigramas) presentes na trajetória de i e w_t é o peso IDF do termo.

Essa escolha mantém a interpretação de sobreposições (adequado para trajetórias) e, ao mesmo tempo, reduz a influência de *tokens* muito frequentes.

3.4.2 Calibragem por percentis e bandas de afinidade

Os limiares de similaridade Weighted Jaccard (Equação 3.2) foram calibrados com base em percentis de similaridade amostrada em pares que compartilham termos informativos. Definiram-se três bandas:

- **Alta afinidade (HIGH):** similaridades $\geq P_{95}$ - trajetórias muito similares;
- **Afinidade média (MID):** similaridades entre P_{60} e P_{75} - sobreposição moderada;
- **Baixa afinidade (LOW):** similaridades entre P_{35} e P_{45} - trajetórias distintas, porém não aleatórias.

3.4.3 Algoritmo de agrupamento e restrições

Para cada banda, os grupos são montados de forma gulosa ao redor de uma semente:

1. gera-se um conjunto candidato por índice invertido (termos de maior *Inverse Document Frequency*, Frequência Inversa de Documento (IDF));
2. selecionam-se membros cuja similaridade com a semente esteja na banda;
3. aplica-se checagem intra-grupo e poda até respeitar coesão mínima;
4. impõem-se limites de tamanho: $\mathrm{MIN} = 5$ e $\mathrm{MAX} = 15$.

O processo produz: estatísticas e termos mais frequentes por grupo, mapeamento grupo-membro e relatórios de descarte (IDs que não entraram em nenhum grupo por banda), permitindo auditoria da cobertura.

3.5 Geração de Diálogos

A geração de diálogos foi implementada como um fluxo de **engenharia de prompt** com saídas estruturadas em JSON. A controlabilidade decorre de três decisões de projeto: (i) decompor o problema em etapas (planejar $\rightarrow$ redigir $\rightarrow$ validar); (ii) condicionar explicitamente a geração por meio de perfis, limites e regras; e (iii) aplicar auditoria automática com regeneração quando necessário. Em conjunto, essas escolhas priorizam **reprodutibilidade** e **padronização** do conteúdo central entre os cenários, reduzindo a variação espúria.

3.5.1 Modelos de linguagem empregados e papéis no pipeline

Foram utilizados modelos de múltiplos provedores para reduzir a dependência de um único estilo gerativo e permitir comparação entre avaliadores. A Tabela 3.2 resume os modelos e seus papéis no *pipeline*.

Tabela 3.2 – Modelos LLM utilizados e seus papéis no *pipeline*.

Provedor	**Modelo**	**Papéis**
OpenAI	GPT-5.1	Planejador (P1), Gerador (P2), Inferidor (P4, *self*)
Google	Gemini 2.5 Pro	Gerador (P2), Avaliador (P3), Inferidor (P4, todos)
Anthropic	Claude Haiku 4.5	Gerador (P2), Inferidor (P4, *self*)

O Gemini atua como inferidor universal (avalia diálogos de todos os geradores), enquanto GPT e Claude fazem *self-evaluation* (inferem apenas sobre diálogos que eles próprios geraram). Essa configuração permite comparar vieses entre provedores e investigar o efeito de autoavaliação. Os *templates* e esquemas de saída foram mantidos idênticos entre provedores, registrando o identificador do modelo em cada execução (OPENAI, 2026; GOOGLE, 2026; ANTHROPIC, 2026).

O *pipeline* separa **modelo gerador** (produção do diálogo) e **modelo avaliador** (controle de qualidade e inferência inversa), evitando que o próprio modelo avalie sistematicamente textos que ele próprio produziu. Sendo aplicada avaliação cruzada entre provedores.

3.5.2 Custos computacionais e consumo de tokens

A execução completa do *pipeline* envolveu custos significativos de API, decorrentes das múltiplas etapas de geração, validação e inferência inversa com três provedores. A Tabela 3.3 resume o consumo de tokens e os custos por provedor.

Tabela 3.3 – Custos e consumo de tokens por provedor LLM.

Provedor	**Tokens totais**	**Requisições**	**Custo (USD)**
OpenAI (GPT-5.1)	6.639.017	690	51,11
Google (Gemini 2.5 Pro)	83.764.475	-	205,38
Anthropic (Claude Haiku 4.5)	11.672.883	-	26,69
Total consolidado	**102.076.375**	-	**283,18**

O Gemini 2.5 Pro concentra a maior parte do consumo de tokens (82% do total) e do custo (72,5%), o que se justifica pelo seu papel como inferidor universal (avaliando diálogos de todos os geradores) e como avaliador de qualidade (Prompt 3). O Claude Haiku 4.5 apresenta o menor custo unitário, enquanto o GPT registrou 690 requisições com custo intermediário. O custo total de USD 283,18 para a geração e avaliação de todo o *corpus* representa uma limitação relevante para a reprodutibilidade e escalabilidade do *pipeline*, conforme discutido na Seção 4.11.

3.5.3 Unidade de geração: canal com múltiplos tópicos

A unidade de geração é uma simulação de **canal corporativo** com múltiplos tópicos (*threads*). Em *chats* corporativos reais, a colaboração tende a ocorrer por meio de conversas paralelas, retomadas e atualizações curtas, em vez de um único diálogo linear. Por esse motivo, cada time é representado como um canal contendo tópicos independentes, mas coerentes entre si.

A parametrização operacional do canal foi definida por faixas, para manter variabilidade realista sem perder controle do volume:

- número de tópicos por canal: $T \in [3, 6]$;
- número de mensagens por tópico: tipicamente $N \in [15, 25]$;
- presença de poucos tópicos longos ($N \in [35, 60]$) para capturar cauda longa (sínteses, atas e decisões).

3.5.4 Dossiê do projeto fixo e roteiro de execução replicável

Para permitir comparação entre cenários, o conteúdo central do trabalho do time foi fixado em um **dossiê do projeto**, gerado uma única vez por time a partir da trajetória (Karrierewege + plus). A função do dossiê é definir *o que* o time está fazendo e *por que* isso é plausível para aquele conjunto de perfis. O dossiê do projeto descreve, de forma concisa:

- o projeto (objetivo, entregas, restrições, partes interessadas e prazo);
- temas recorrentes (3-6) coerentes com os perfis do time;
- lista de tarefas (5-12) pequenas e atribuíveis a papéis;
- artefatos citáveis (por exemplo: documento, planilha, ticket, e-mail ou incidente fictício, porém consistente).

Em seguida, é definido um **roteiro do canal**, que funciona como um plano de execução: ele explicita *como* o trabalho se desenrola em conversa. O roteiro especifica: (a) quais tópicos ocorrerão; (b) o tamanho esperado de cada tópico; e (c) quais decisões devem necessariamente aparecer no texto. Em termos práticos, o roteiro organiza:

- a sequência de tópicos (por exemplo: abertura, alinhamento, riscos, execução, revisão e entrega);
- metas de tamanho por tópico e distribuição aproximada de fala por papel;
- um conjunto de **decisões obrigatórias** que devem emergir nas conversas.

A replicação entre cenários não busca copiar o texto literal. Em vez disso, fixa-se o dossiê e o roteiro e "re-realizam-se" as conversas sob diferentes condições latentes, preservando temas, artefatos e decisões (o que permite comparações justas entre cenários).

3.5.5 Condições por cenário sem vazamento explícito

Para cada cenário, o time recebe rótulos por participante (uma linha por `_id` a partir do `per_id` representativo). Esses rótulos não são exibidos no texto; eles são convertidos em **descritores comportamentais** que orientam estilo e dinâmica conversacional. Assim, o sinal do cenário aparece na linguagem e no comportamento, e não em palavras-chave.

A conversão de rótulos para comportamento segue regras interpretáveis, como:

Produtividade alta: maior iniciativa, mais lembretes, respostas mais frequentes e encaminhamentos rápidos.

Habilidade alta: explicações mais corretas, menos dúvidas básicas, síntese melhor e melhor uso de artefatos.

Desempenho baixo: maior retrabalho, atrasos plausíveis, necessidade de correção e dependência de terceiros.

Uma regra de qualidade é aplicada como restrição dura: o diálogo **não pode mencionar** explicitamente "desempenho", "produtividade", "habilidade", "score" nem níveis (baixo/médio/alto). O objetivo é evitar vazamento e forçar que a inferência seja feita por evidências linguísticas.

3.5.6 Templates de prompt e validação

O processo foi implementado com três *templates* de *prompt*, sempre com saída estritamente em JSON, cobrindo planejamento, geração e controle de qualidade. Por questões de legibilidade, os templates completos (incluindo schemas de saída e regras) foram deslocados para o Apêndice A:

1. **Prompt 1 - Planejamento** (Apêndice A.1): define o dossiê do projeto e o roteiro do canal, fixos por time.
2. **Prompt 2 - Diálogos** (Apêndice A.2) gera o canal do cenário (tópicos, mensagens, referências a artefatos e decisões obrigatórias), executando o roteiro sem alterá-lo.
3. **Prompt 3 - Validação** (Apêndice A.3): valida a saída e, quando aplicável, corrige problemas formais.
4. **Prompt 4 - Inferência** (Apêndice A.4): a partir do texto gerado pelo *prompt 2*, retorna o rótulo provável de cada variável latente, para cada participante da conversa.

3.5.7 Controle de variância, reprodutibilidade e registros

Cada execução foi registrada para permitir auditoria e reprodutibilidade. Além de registrar o conteúdo gerado, os *logs* permitem rastrear *como* o texto foi produzido e quais regras foram aplicadas. Em cada execução, armazenou-se:

- identificadores de execução (cenário, grupo, semente) e marca temporal;
- modelo, versão e parâmetros (por exemplo: temperatura e top-p);
- entradas de geração (dossiê do projeto, roteiro do canal, perfis e descritores comportamentais);
- saída em JSON e relatório de validação.

3.6 Avaliação do *Corpus* de Diálogos

A avaliação foi definida em três níveis, combinando checagens automáticas e avaliação por LLM. A motivação é separar problemas de **estrutura** (fáceis de detectar) de problemas de **qualidade comunicacional** (mais subjetivos), além de mensurar o sinal latente via inferência inversa com comparação contra a verdade de referência.

3.6.1 Nível 1: Validação estrutural e sanidade

O primeiro nível concentra verificações determinísticas, para garantir que o conjunto gerado é utilizável como *dataset* e que respeita restrições de não-vazamento. Essas verificações operam sobre o JSON e sobre padrões textuais básicos. As checagens incluem:

- integridade do schema (JSON) e presença de chaves obrigatórias;
- contagem de tópicos e mensagens dentro dos limites definidos;
- distribuição de participação por interlocutor (por exemplo: entropia e Gini de mensagens);
- testes de vazamento: ausência de rótulos internos e ausência de notas numéricas associáveis a atributos latentes.

3.6.2 Nível 2: Avaliação de qualidade por rubricas com LLM

No segundo nível, cada diálogo é avaliado por rubricas (coerência, aderência ao objetivo, realismo corporativo, utilidade e consistência de papéis), gerando notas e justificativas curtas. A avaliação por LLM foi estruturada como preenchimento de rubricas com evidências do próprio texto, seguindo a lógica de avaliação guiada por critérios e justificativas. Para discussão de métodos LLM-baseados de avaliação e seus desafios, ver Gao et al. (2024). Para reduzir variância e instabilidade do avaliador, aplicou-se:

- múltiplas amostras por diálogo e agregação por mediana;
- checagem cruzada com prompts alternativos (mesmas rubricas, redação diferente);
- comparação entre avaliadores de provedores distintos para detectar divergências sistemáticas.

3.6.3 Nível 3: Avaliação inversa (inferência dos atributos a partir do texto)

A avaliação inversa mede se os atributos latentes do cenário foram refletidos no comportamento linguístico *sem vazamento explícito*. Aqui, o LLM atua como um "inferidor" que estima, a partir do diálogo, métricas e níveis de desempenho/produtividade/habilidade e permite calcular a acurácia do inferidor em relação à verdade de referência. O procedimento foi:

1. **Verdade de referência:** para cada participante, registrar os níveis por cenário (baixo/-médio/alto) e, quando disponíveis, os valores contínuos (por exemplo: score em 0-100) derivados do `per_id` representativo.

2. **Inferência inversa:** após a geração do diálogo, aplica-se um quarto template (Apêndice A.4) que recebe apenas (a) o canal gerado e (b) os perfis de trajetória (sem descritores comportamentais e sem rótulos do cenário). O inferidor deve:

   - Estimar, para cada participante, os níveis (LOW/MID/HIGH) de desempenho, produtividade e habilidade;
   - Estimar um valor contínuo (0-100) por atributo, acompanhado de confiança;
   - Apontar evidências textuais (referenciando `msg_id` e trechos curtos) que sustentem a inferência.

3. **Acurácia por modelo inferidor:** o prompt é executado com múltiplos modelos (GPT, Gemini e Claude). Para cada modelo, calculam-se métricas de classificação (acurácia macro, F1 macro, matrizes de confusão) por atributo e agregadas no conjunto. Quando houver valores contínuos estimados, avaliam-se correlação (Spearman) e erro absoluto médio.

Dessa forma, além de medir a qualidade do texto (Seção 3.6.2), o protocolo mede o quanto o *corpus* preserva sinal sobre as latentes a ponto de ser recuperável por inferência, e compara diretamente a capacidade de diferentes LLMs em recuperar esse sinal (acurácia do inferidor).

Como resultado do desenvolvimento, obteve-se: (i) uma base consolidada de históricos profissionais (*Karrierewege* + *Karrierewege_plus*) em Parquet; (ii) dois cenários de simulação executados (`S_full_high_noise` e `S_random_with_latents`) com $R = 25$ replicações cada, incluindo tabelas por experiência e por pessoa; (iii) relatórios de validação com MCSE por cenário e seleção de repetição representativa por *medoid*; (iv) 105 times formados por similaridade de trajetória via TF-IDF e Weighted Jaccard, distribuídos em três bandas de afinidade (HIGH, MID, LOW); (v) diálogos gerados por três provedores LLM (GPT-5.1, Gemini 2.5 Pro e Claude Haiku 4.5), com validação estrutural (Prompt 3 A.3) e avaliação por rubricas; (vi) inferências inversas (Prompt 4 A.4) executadas por três inferidores, com comparação contra *ground truth*; e (vii) um módulo de análise (`analise.py`) com métricas de classificação, testes estatísticos e visualizações.

3.7 Síntese do Desenvolvimento

O desenvolvimento descrito neste capítulo resultou em um *pipeline* completo e auditável para geração e avaliação de diálogos corporativos sintéticos condicionados a variáveis latentes.

Os principais artefatos produzidos incluem: (i) uma base consolidada de históricos profissionais em Parquet; (ii) cenários de simulação com replicações de Monte Carlo validadas por MCSE; (iii) 105 times formados por similaridade de trajetória via TF-IDF e Weighted Jaccard; (iv) diálogos gerados por três provedores de LLM com validação estrutural e avaliação por rubricas; e (v) inferências inversas com comparação contra *ground truth*. Os resultados obtidos com este *pipeline* são apresentados no Capítulo 4.

4 Resultados e Discussão

Este capítulo apresenta os resultados obtidos com a execução completa do *pipeline* descrito no Capítulo 3. A análise abrange a descrição do *corpus* gerado (Seção 4.1), as métricas globais de inferência inversa (Seção 4.2), a comparação entre cenários (Seção 4.3), a comparação entre provedores (Seção 4.4), a análise de viés de predição (Seção 4.5), os testes estatísticos (Seção 4.6), a qualidade do *corpus* (Seção 4.7), a análise detalhada da validação P3 (Seção 4.8), a análise por banda de afinidade (Seção 4.9) e a discussão integrada dos achados (Seção 4.11).

4.1 Descrição do *Corpus*

O *corpus* analisado compreende **16.524 observações** de inferência inversa, distribuídas uniformemente em três dimensões latentes (habilidade, produtividade e desempenho), com 5.508 observações por dimensão. Os diálogos foram gerados por três provedores LLM - GPT-5.1 (OPENAI, 2026), Gemini 2.5 Pro (GOOGLE, 2026) e Claude Haiku 4.5 (ANTHROPIC, 2026) e avaliados por três inferidores (os mesmos provedores), totalizando 105 times compostos por 541 membros únicos, sob dois cenários experimentais.

A distribuição de classes no *ground truth* apresenta desbalanceamento moderado, com predominância da classe MID (≈43-46%) sobre LOW (≈25-28%) e HIGH (≈29-31%), refletindo a discretização por limiares τ_1 e τ_2 aplicada sobre as variáveis latentes contínuas. A Tabela 4.1 apresenta a distribuição por dimensão.

Tabela 4.1 – Distribuição de classes no *ground truth* por dimensão.

Dimensão	**LOW (%)**	**MID (%)**	**HIGH (%)**
Habilidade	27,65	43,57	28,78
Produtividade	25,20	46,21	28,59
Desempenho	26,11	43,21	30,68

4.2 Métricas Globais de Inferência Inversa

As métricas selecionadas seguem recomendações da literatura para classificação ordinal multiclasse (SOKOLOVA; LAPALME, 2009; CHICCO; JURMAN, 2020). Foram adotadas: *accuracy* (proporção de acertos exatos), *balanced accuracy* (média das taxas de *recall* por classe), *macro F1* (média harmônica não ponderada de *precision* e *recall* por classe), MCCs, kappa quadrático de Cohen (concordância além do acaso com ponderação ordinal (COHEN, 1968)), *Mean Absolute Errors*, Erros Absolutos Médios (MAEs) ordinal (erro absoluto médio na escala

LOW=0, MID=1, HIGH=2) e *within-1 accuracy* (proporção de predições a no máximo uma classe de distância).

4.2.1 Resultados agregados por dimensão

A Tabela 4.2 apresenta as métricas globais com todos os provedores agregados.

Tabela 4.2 – Métricas globais por dimensão (todos os provedores agregados).

Dimensão	**n**	**Acc.**	**Bal. Acc.**	**Macro F1**	**MCC**	**κ quad.**	**MAE ord.**	**W-1 Acc.**
Habilidade	5508	0,464	0,484	0,432	0,235	0,343	0,635	0,901
Produtividade	5508	0,460	0,470	0,432	0,204	0,312	0,629	0,910
Desempenho	5508	0,484	0,494	0,455	0,254	0,381	0,599	0,917

A *accuracy* global situa-se entre 46,0% e 48,4%, significativamente acima do *baseline* de chance (33,3% para 3 classes equiprováveis), porém modesta em termos absolutos. O MCC entre 0,204 e 0,254 indica correlação fraca a moderada entre predições e *ground truth*. O kappa quadrático (0,312-0,381) sugere concordância *fair* a *moderate* segundo a escala de Landis e Koch (1977).

A dimensão **desempenho** apresenta os melhores resultados em todas as métricas, sugerindo que os LLMs codificam e recuperam sinais de desempenho com maior fidelidade do que habilidade ou produtividade. Isso é consistente com a hipótese de que comportamentos associados ao desempenho (retrabalho, atrasos, necessidade de correção) são mais salientes linguisticamente do que indicadores de habilidade ou produtividade (ROTUNDO; SACKETT, 2002).

A *within-1 accuracy* elevada (90-92%) indica que, embora os modelos errem frequentemente a classe exata, raramente cometem erros de duas classes de distância (e.g., predizer HIGH quando o *ground truth* é LOW).

4.2.2 Métricas por classe e o viés HIGH

A análise por classe revela um padrão crítico, apresentado na Tabela 4.3.

O padrão demonstra que os inferidores apresentam **viés massivo para a classe HIGH**, com *recall* de 87-95% para HIGH contra apenas 17-19% para LOW e 19-24% para MID. A *precision* de LOW é alta (65-81%), indicando que quando o modelo prediz LOW, geralmente acerta - mas raramente o faz. Este fenômeno é consistente com o *positivity bias* documentado na literatura de LLMs (ZHENG et al., 2023a; WANG et al., 2023), onde modelos de linguagem tendem a atribuir avaliações mais favoráveis.

A Figura 4.1 apresenta a matriz de confusão global para habilidade, a Figura 4.2 para produtividade e a Figura 4.3 para desempenho, evidenciando a concentração de predições na classe HIGH.

Tabela 4.3 – Métricas por classe e dimensão (*precision*, *recall* e F1).

Dimensão	**Classe**	**Precision**	**Recall**	**F1**	**Suporte**
Habilidade	LOW	0,803	0,215	0,339	1523
	MID	0,496	0,325	0,392	2400
	HIGH	0,410	0,913	0,566	1585
Produtividade	LOW	0,664	0,216	0,326	1388
	MID	0,519	0,367	0,430	2545
	HIGH	0,400	0,827	0,539	1575
Desempenho	LOW	0,802	0,248	0,379	1438
	MID	0,500	0,324	0,393	2380
	HIGH	0,438	0,911	0,591	1690

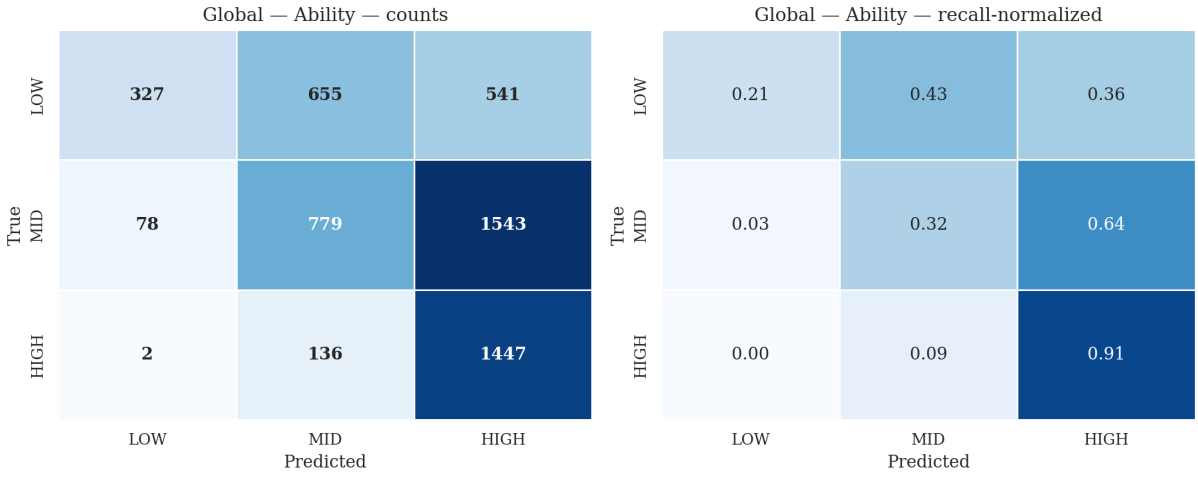

Figura 4.1 – Matriz de confusão global - habilidade (todos os provedores agregados).

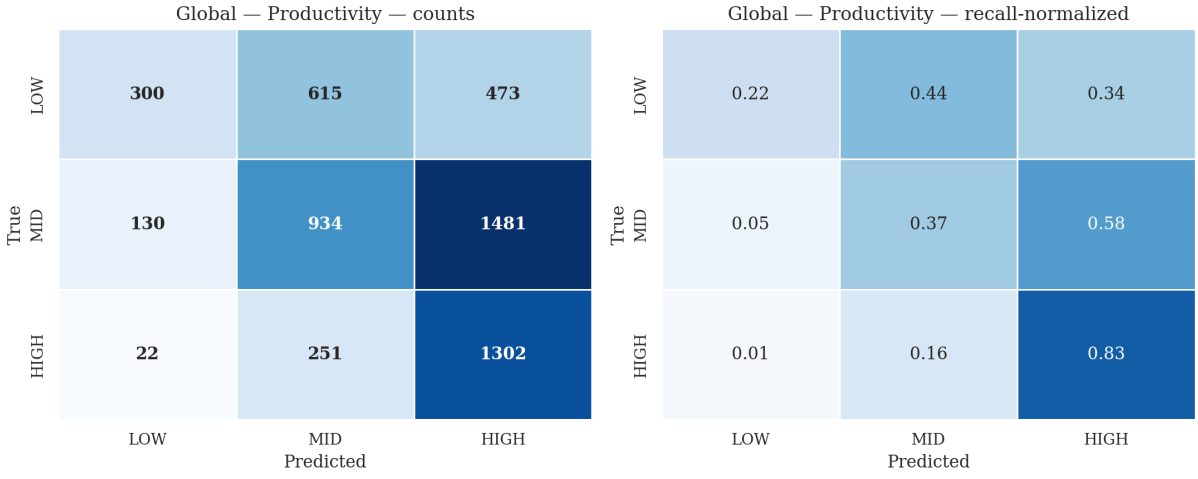

Figura 4.2 – Matriz de confusão global - produtividade (todos os provedores agregados).

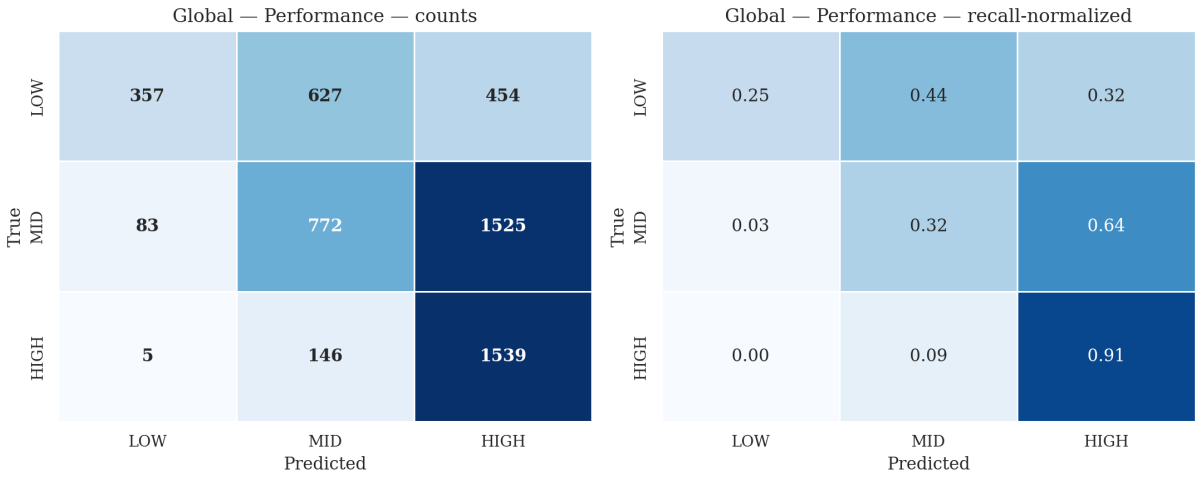

Figura 4.3 – Matriz de confusão global - desempenho (todos os provedores agregados).

A magnitude do viés fica evidente ao comparar as distribuições de predições e *ground truth*: os inferidores predizem HIGH em ≈70-74% dos casos, quando a prevalência real é de ≈29-31%. A Figura 4.4 ilustra essa discrepância.

4.3 Comparação entre Cenários

Dois cenários foram avaliados: **S_full_high_noise**, com variáveis latentes estruturalmente correlacionadas com a trajetória profissional e alto ruído ($\sigma = 1{,}4$); e **S_random_with_latents**, com variáveis latentes geradas aleatoriamente (*baseline* sem correlação estrutural). A Tabela 4.4 apresenta a comparação.

Tabela 4.4 – Métricas por cenário e dimensão.

Dimensão	**Métrica**	**S_full_high_noise**	**S_random_w_latents**	**Δ**
Habilidade	Accuracy	0,507	0,420	+8,6pp
	Macro F1	0,469	0,375	+9,5pp
	MCC	0,300	0,166	+13,4pp
Produtividade	Accuracy	0,526	0,395	+13,0pp
	Macro F1	0,492	0,350	+14,2pp
	MCC	0,296	0,100	+19,5pp
Desempenho	Accuracy	0,554	0,415	+14,0pp
	Macro F1	0,511	0,382	+12,9pp
	MCC	0,346	0,157	+18,8pp

O cenário estruturado supera consistentemente o *baseline* aleatório em todas as dimensões e métricas, com ganhos de 7,9 a 17,7 pontos percentuais. Isso confirma a hipótese central do trabalho: quando as variáveis latentes são estruturalmente correlacionadas com a trajetória profis-

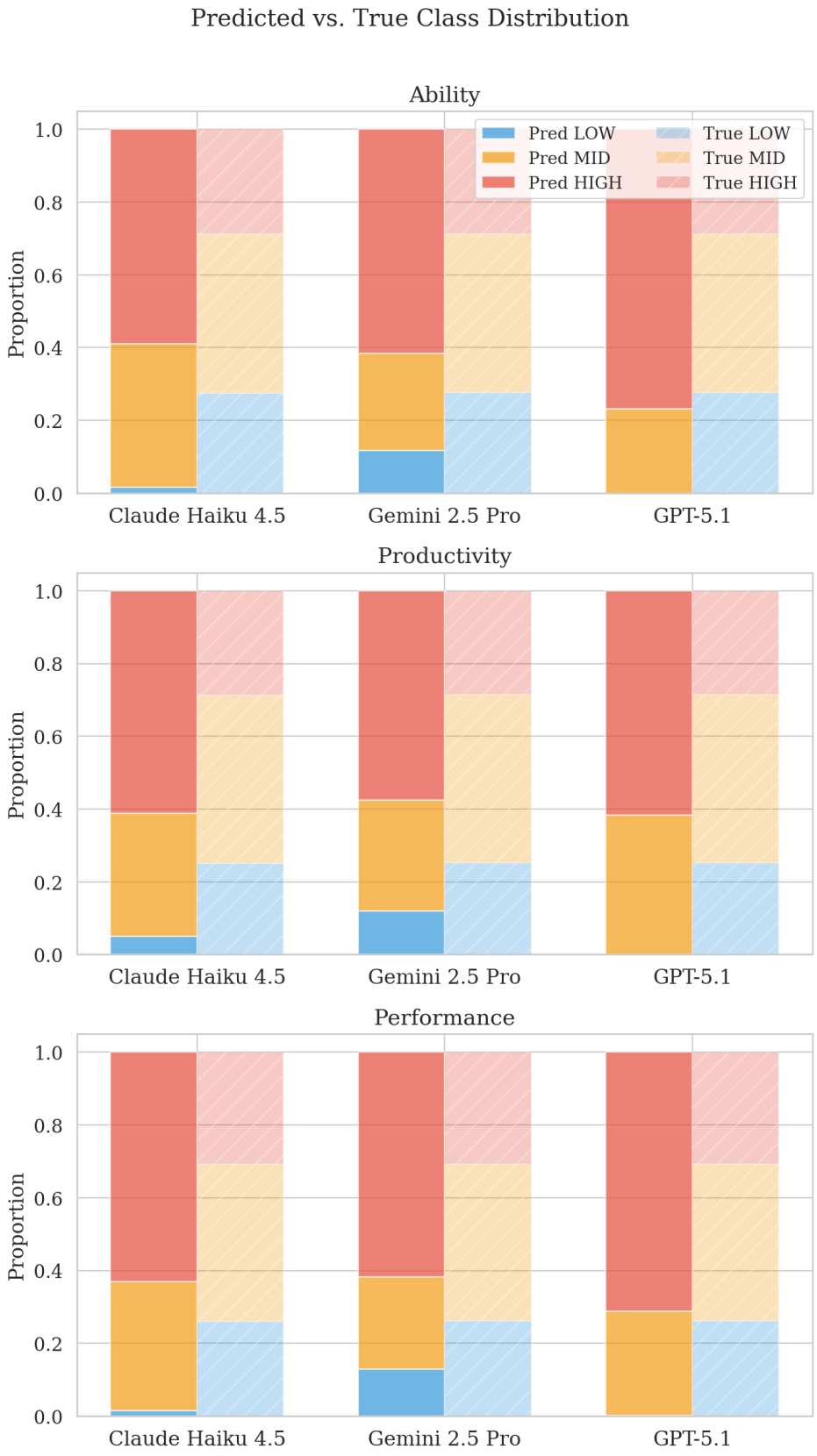

Figura 4.4 – Distribuição de classes: *ground truth* vs. predições por dimensão.

sional, os LLMs geradores conseguem codificar sinais comportamentais que são parcialmente recuperáveis por LLMs inferidores.

O MCC do cenário aleatório para produtividade (0,100) é próximo de zero, indicando predição essencialmente aleatória, exatamente o esperado quando não há correlação estrutural entre latentes e trajetória. Isso valida o *design* experimental e confirma que os ganhos observados no cenário estruturado não são artefatos metodológicos. A Figura 4.5 ilustra a comparação.

4.4 Comparação entre Provedores

4.4.1 Provedores como geradores

A Tabela 4.5 apresenta as métricas quando os diálogos são segmentados pelo provedor gerador.

Tabela 4.5 – Métricas por provedor gerador e dimensão.

Dimensão	Métrica	Claude Haiku 4.5	Gemini 2.5 Pro	GPT-5.1
Habilidade	Accuracy	0,457	**0,623**	0,391
	Macro F1	0,398	**0,628**	0,332
	MCC	0,209	**0,473**	0,130
Produtividade	Accuracy	0,449	**0,568**	0,418
	Macro F1	0,412	**0,573**	0,356
	MCC	0,186	**0,378**	0,128
Desempenho	Accuracy	0,468	**0,636**	0,425
	Macro F1	0,417	**0,641**	0,359
	MCC	0,218	**0,499**	0,158

O Gemini 2.5 Pro destaca-se como o melhor gerador em todas as dimensões, com margens substanciais. Seus diálogos produzem *accuracy* de 57-64% e MCC de 0,38-0,50, indicando que o sinal latente é codificado de forma mais recuperável. O GPT-5.1 apresenta os piores resultados como gerador, com MCC entre 0,13 e 0,16. A superioridade do Gemini como gerador pode estar relacionada à sua capacidade de seguir instruções complexas de condicionamento comportamental, conforme discutido na literatura sobre *instruction-following* (ZHOU et al., 2023). A Figura 4.6 ilustra a comparação.

4.4.2 Provedores como inferidores

A Tabela 4.6 apresenta as métricas segmentadas pelo provedor inferidor.

O Gemini 2.5 Pro também é o melhor inferidor, com vantagem expressiva em *macro F1* e MCC. O GPT-5.1 apresenta desempenho consistentemente inferior, com MCC de 0,036 em habilidade, praticamente indistinguível de predição aleatória. A diferença entre Gemini e os demais inferidores é mais pronunciada em *macro F1* do que em *accuracy*, indicando que o

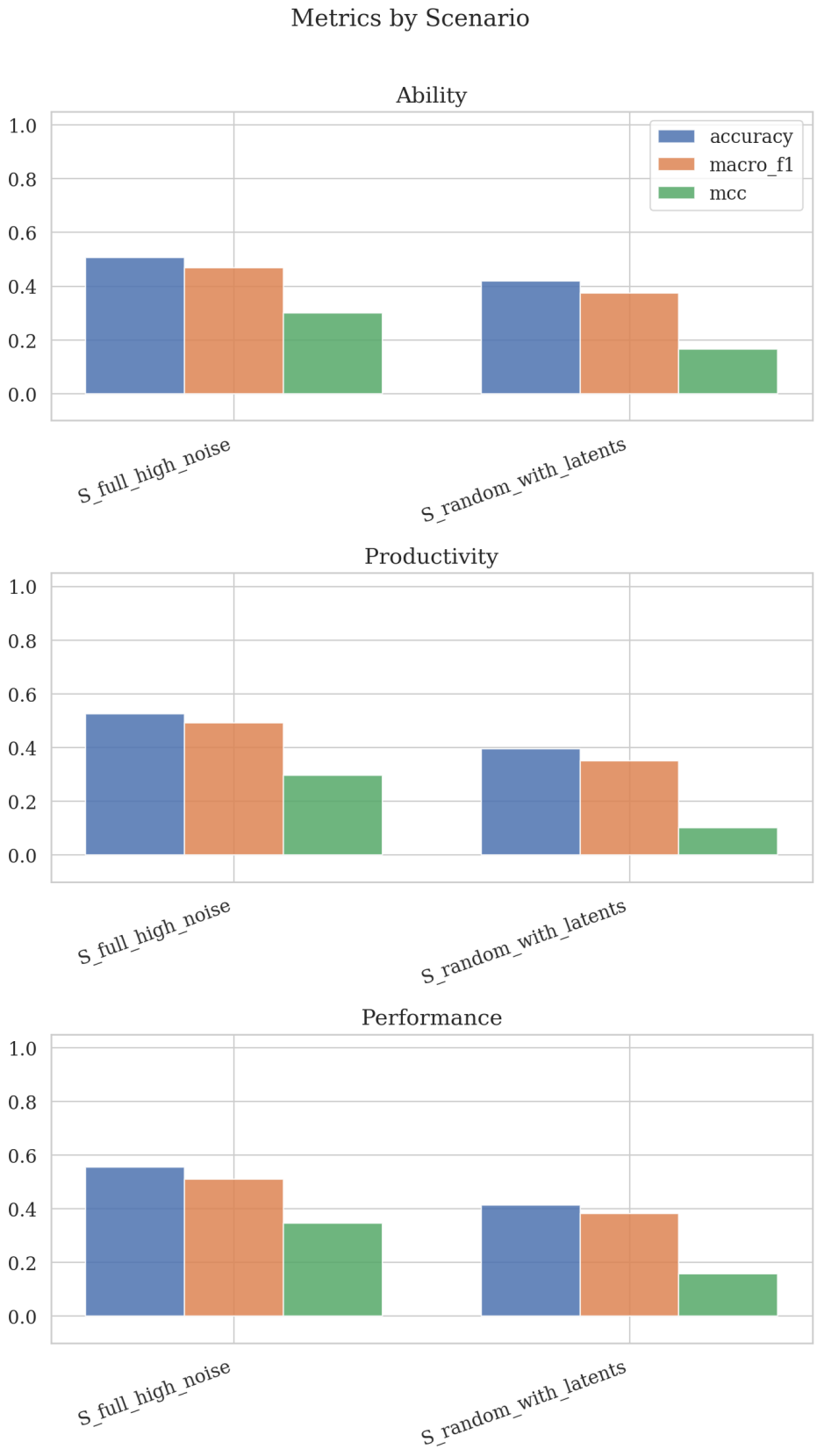

Figura 4.5 – Comparação de métricas entre cenários por dimensão.

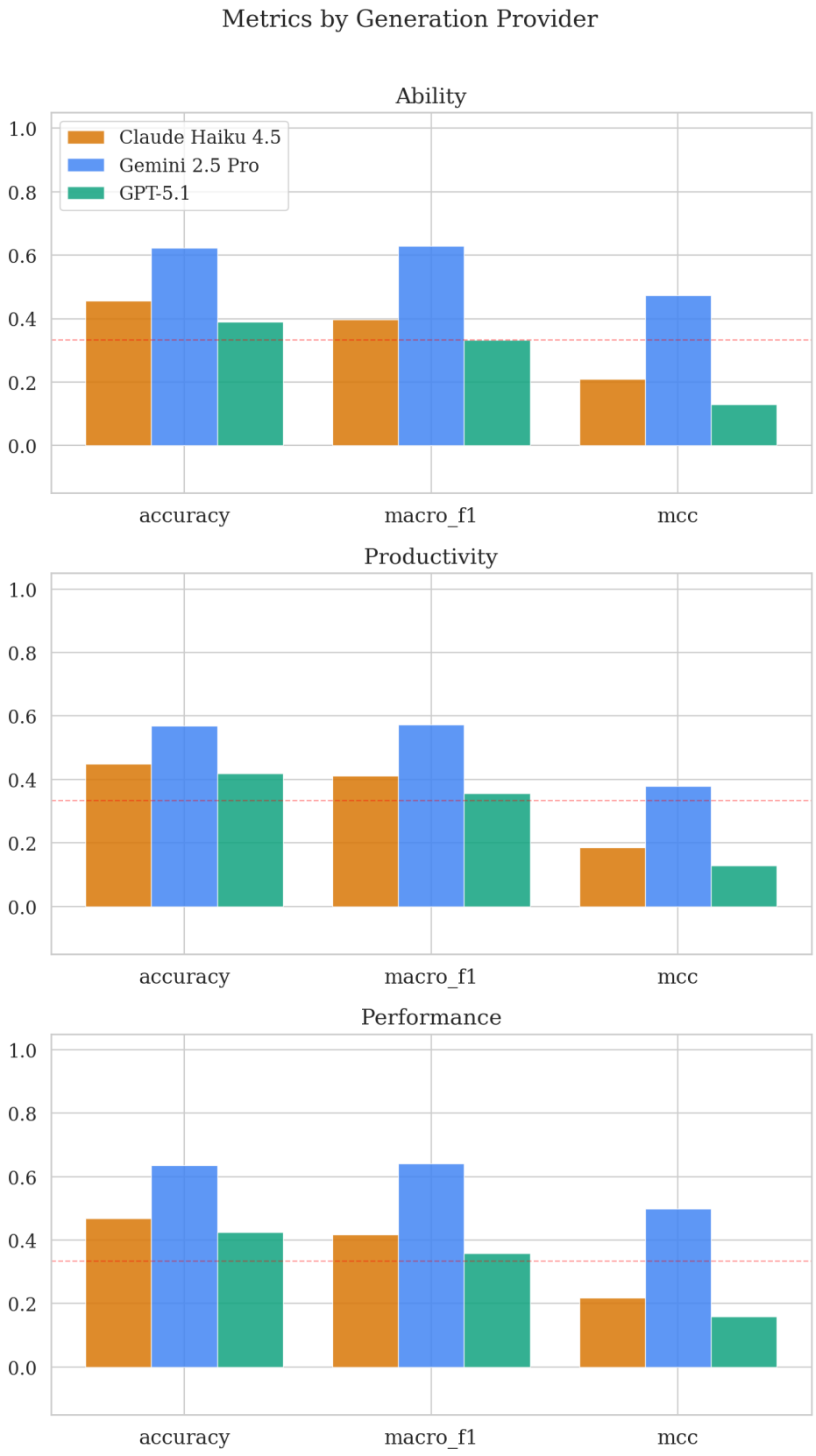

Figura 4.6 – Métricas de inferência inversa por provedor gerador.

Tabela 4.6 – Métricas por provedor inferidor e dimensão.

Dimensão	Métrica	Claude Haiku 4.5	Gemini 2.5 Pro	GPT-5.1
Habilidade	Accuracy	0,460	**0,506**	0,339
	Macro F1	0,379	**0,494**	0,240
	MCC	0,200	**0,305**	0,036
Produtividade	Accuracy	0,438	**0,486**	0,407
	Macro F1	0,387	**0,474**	0,309
	MCC	0,165	**0,251**	0,097
Desempenho	Accuracy	0,445	**0,528**	0,393
	Macro F1	0,358	**0,520**	0,287
	MCC	0,168	**0,330**	0,097

Gemini distribui melhor suas predições entre as três classes. A Figura 4.7 apresenta a comparação visual.

4.4.3 Cruzamento gerador × inferidor

A Tabela 4.7 apresenta a *accuracy* e o *macro F1* para cada par (gerador, inferidor), revelando interações importantes.

Tabela 4.7 – *Accuracy* por par gerador → inferidor e dimensão.

Dimensão	Gerador → Inferidor	Accuracy	Macro F1
Habilidade	Claude → Claude (*self*)	0,460	0,379
	Claude → Gemini (*cross*)	0,454	0,411
	Gemini → Gemini (*self*)	**0,623**	**0,628**
	OpenAI → Gemini (*cross*)	0,442	0,411
	OpenAI → OpenAI (*self*)	0,339	0,240
Desempenho	Claude → Claude (*self*)	0,445	0,358
	Claude → Gemini (*cross*)	0,492	0,465
	Gemini → Gemini (*self*)	**0,636**	**0,641**
	OpenAI → Gemini (*cross*)	0,456	0,419
	OpenAI → OpenAI (*self*)	0,393	0,287

O par Gemini → Gemini (*self-evaluation*) domina em todas as dimensões, atingindo 63,6% de *accuracy* em desempenho. As Figuras 4.8 e 4.9 apresentam os *heatmaps* de *accuracy*, *macro F1* e MCC por par gerador × inferidor.

4.4.4 *Self-evaluation* vs. *cross-evaluation*

Para investigar o viés de autoavaliação, compararam-se as métricas quando o inferidor avalia diálogos que ele próprio gerou (*self*) versus diálogos de outros geradores (*cross*). A Tabela 4.8 apresenta os resultados.

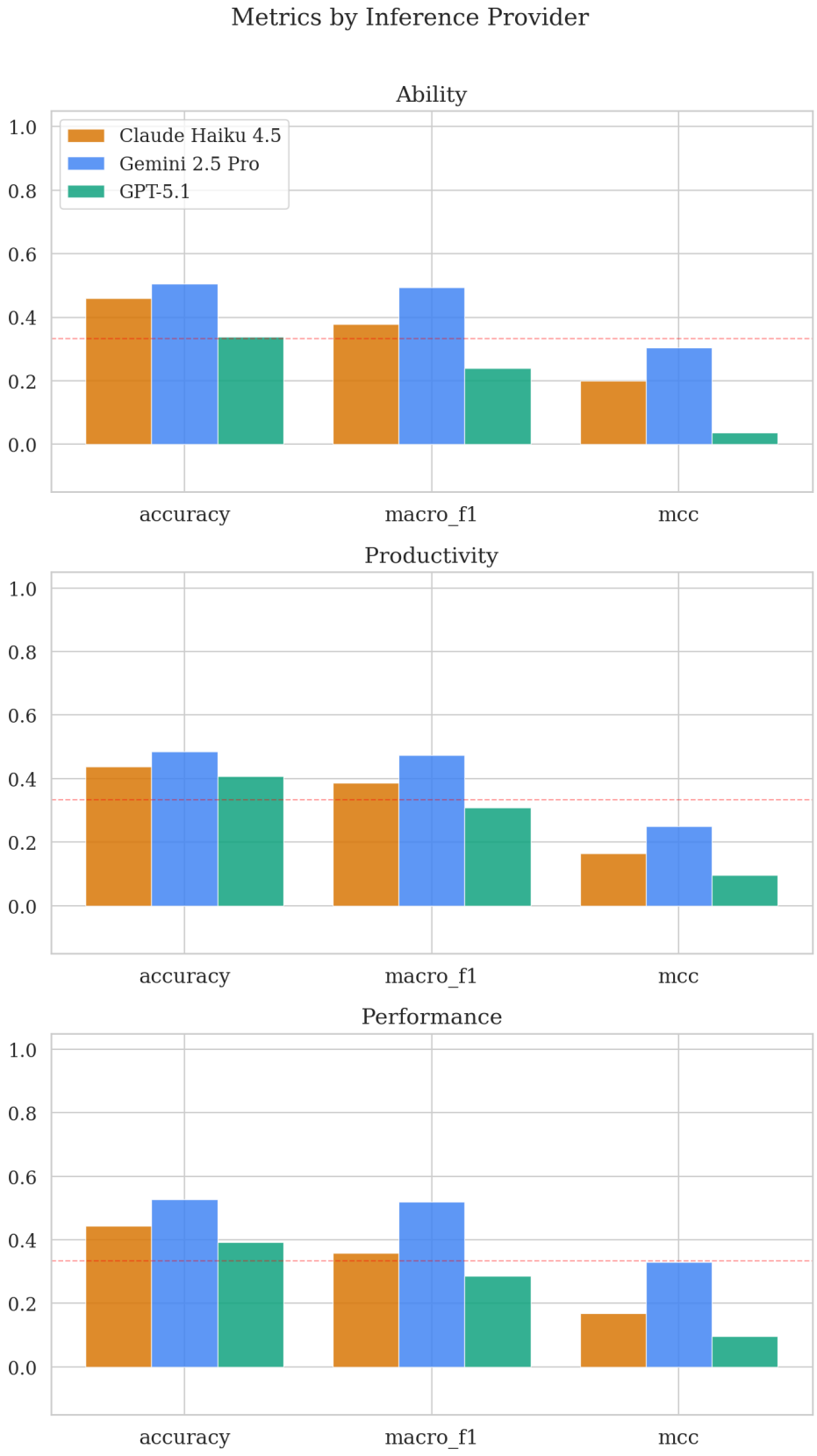

Figura 4.7 – Métricas de inferência inversa por provedor inferidor.

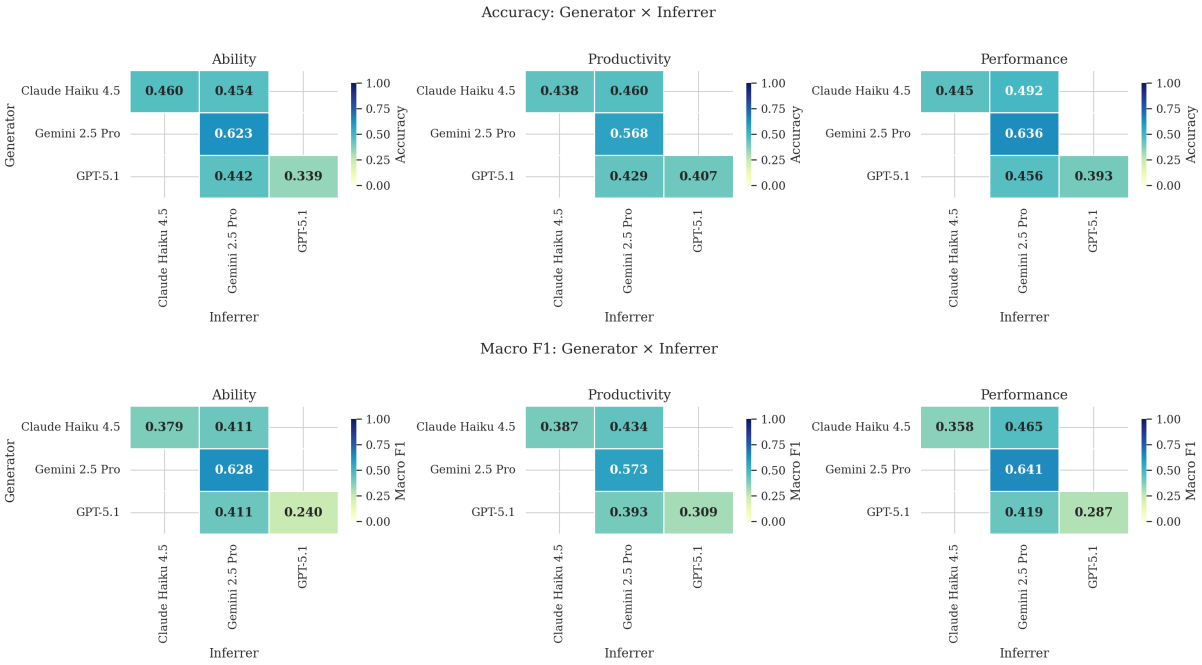

Figura 4.8 – *Heatmaps* de *accuracy* (acima) e *macro F1* (abaixo) por par gerador × inferidor.

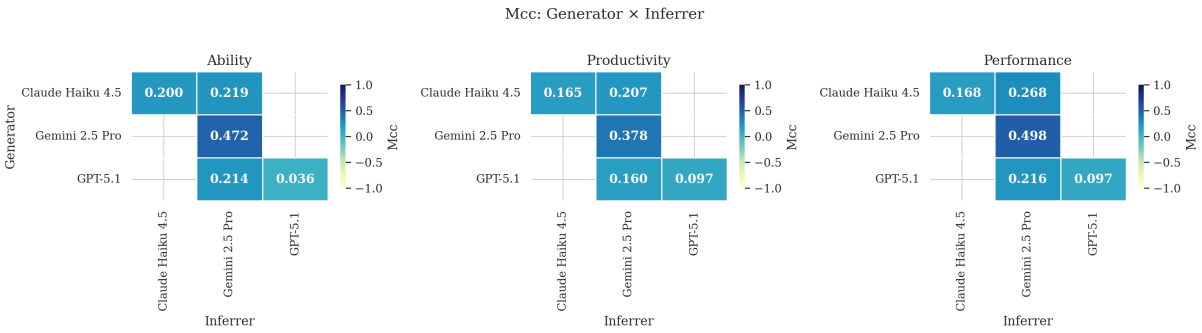

Figura 4.9 – *Heatmap* de MCC por par gerador × inferidor.

Tabela 4.8 – Comparação *self-evaluation* vs. *cross-evaluation*.

Dimensão	**Accuracy**		**Macro F1**		**MCC**	
	Self	*Cross*	*Self*	*Cross*	*Self*	*Cross*
Habilidade	0,474	0,448	0,446	0,412	0,248	0,216
Produtividade	0,471	0,445	0,444	0,414	0,218	0,184
Desempenho	0,491	0,474	0,463	0,442	0,262	0,243

A *self-evaluation* supera consistentemente a *cross-evaluation* em 2,1-3,4 pontos percentuais de *macro F1*. Essa vantagem, embora menor que em experimentos anteriores, ainda pode ter duas interpretações não mutuamente exclusivas: (i) **viés de autoavaliação**, onde o modelo reconhece padrões estilísticos que ele próprio utiliza para codificar os atributos latentes (ZHENG et al., 2023a); e (ii) **consistência interna**, onde o modelo que gera e o que infere compartilham o mesmo "vocabulário comportamental" para representar LOW/MID/HIGH. A Figura 4.10 ilustra a comparação.

4.5 Viés de Predição

A análise de viés revela que todos os inferidores sub-predizem MID e sobre-predizem HIGH. Contrariamente à expectativa de um "viés MID" (tendência a predizer a classe central como opção segura), todos os inferidores apresentam viés negativo para MID e viés positivo para HIGH. O Gemini é o mais extremo, predizendo HIGH em ≈62-65% dos casos.

Esse padrão sugere que os LLMs, ao analisar diálogos corporativos, tendem a interpretar participação ativa em conversas como indicativo de alto desempenho, um viés de *halo effect* linguístico, onde a presença e articulação verbal são confundidas com competência elevada. A Figura 4.11 apresenta o viés de predição por inferidor e dimensão.

4.6 Testes Estatísticos

4.6.1 McNemar pareado com correção de Holm-Bonferroni

O teste de McNemar avalia se dois classificadores cometem erros em observações diferentes, indicando diferença estatisticamente significativa entre eles. A correção de Holm-Bonferroni controla a taxa de erro familiar (*FWER*) ao realizar múltiplas comparações. A Tabela 4.9 apresenta os resultados.

Tabela 4.9 – Teste de McNemar pareado entre inferidores (com correção de Holm-Bonferroni).

Dimensão	**Par**	χ^2	**p-valor**	**p-adj**	**Sig.**
Habilidade	Claude vs. Gemini	0,145	0,703	0,703	ns
Habilidade	Gemini vs. OpenAI	61,190	<0,001	<0,001	***
Produtividade	Claude vs. Gemini	3,218	0,073	0,219	ns
Produtividade	Gemini vs. OpenAI	3,182	0,074	0,219	ns
Desempenho	Claude vs. Gemini	14,307	<0,001	<0,001	***
Desempenho	Gemini vs. OpenAI	24,291	<0,001	<0,001	***

Após correção de Holm, apenas as comparações em habilidade mantêm significância estatística ($p < 0{,}05$), indicando que, para essa dimensão, os inferidores diferem significativamente em seus padrões de acerto/erro.

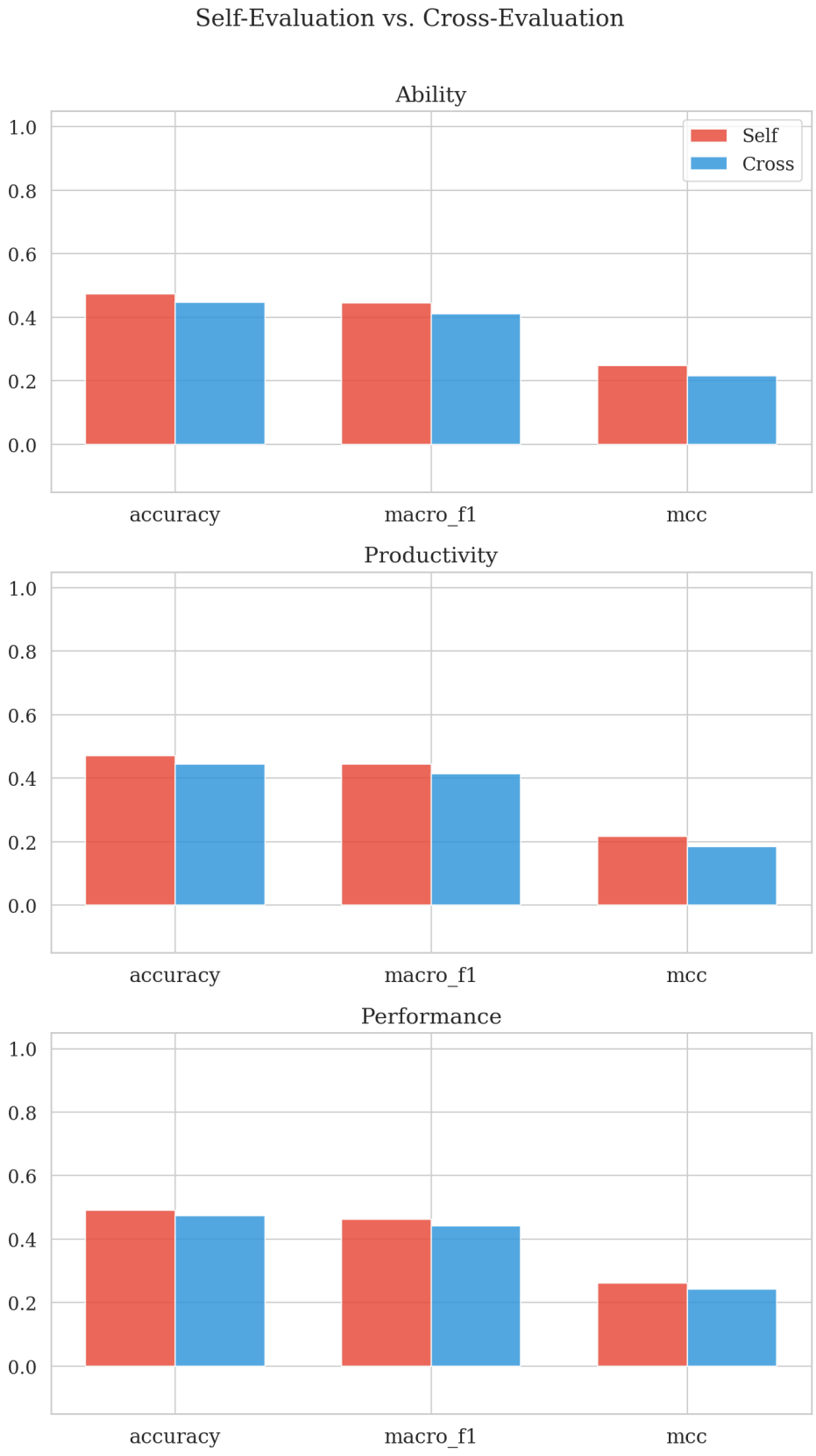

Figura 4.10 – Comparação de métricas entre *self-evaluation* e *cross-evaluation*.

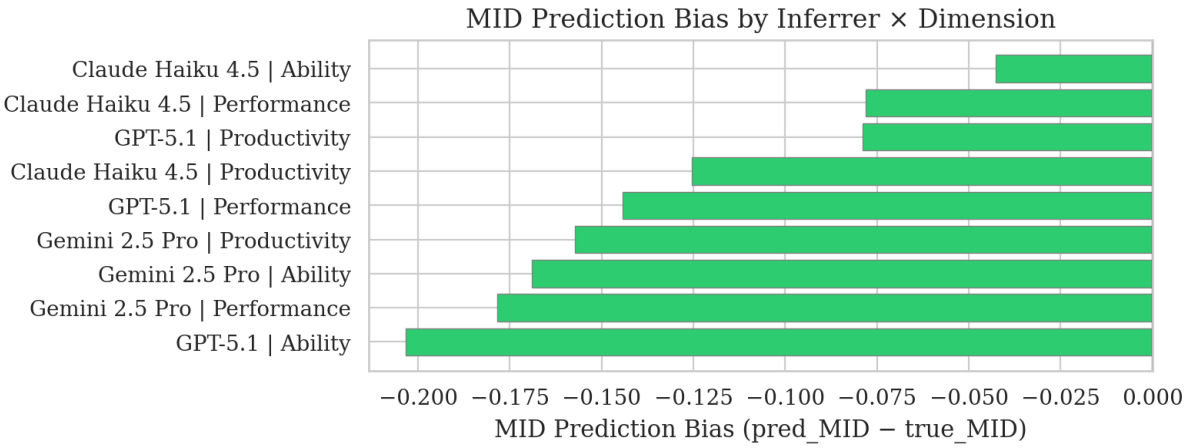

Figura 4.11 – Viés de predição MID por provedor inferidor e dimensão.

4.6.2 Intervalos de confiança por *bootstrap*

Os intervalos de confiança por *bootstrap* (2.000 reamostras) confirmam a robustez das estimativas. A Tabela 4.10 apresenta os resultados e a Figura 4.12 os *forest plots* correspondentes.

Tabela 4.10 – Intervalos de confiança de 95% por *bootstrap* para *accuracy* e *macro F1*.

Inferidor	**Dimensão**	**Acc.**	**IC 95% Acc.**	**Macro F1**	**IC 95% F1**
Claude Haiku 4.5	Habilidade	0,460	[0,431; 0,489]	0,379	[0,352; 0,405]
Claude Haiku 4.5	Desempenho	0,445	[0,415; 0,474]	0,358	[0,331; 0,382]
Claude Haiku 4.5	Produtividade	0,438	[0,410; 0,468]	0,387	[0,358; 0,418]
Gemini 2.5 Pro	Habilidade	0,506	[0,489; 0,524]	0,494	[0,477; 0,512]
Gemini 2.5 Pro	Desempenho	0,528	[0,512; 0,545]	0,520	[0,502; 0,537]
Gemini 2.5 Pro	Produtividade	0,486	[0,469; 0,503]	0,474	[0,456; 0,491]
GPT-5.1	Habilidade	0,339	[0,311; 0,367]	0,240	[0,222; 0,259]
GPT-5.1	Desempenho	0,393	[0,364; 0,422]	0,287	[0,267; 0,308]
GPT-5.1	Produtividade	0,407	[0,378; 0,437]	0,309	[0,289; 0,330]

Os intervalos de confiança do GPT-5.1 em habilidade ([0,311; 0,367]) incluem o *baseline* de chance (0,333), confirmando que seu desempenho como inferidor nessa dimensão é estatisticamente indistinguível de predição aleatória. Todos os demais intervalos excluem 0,333, indicando desempenho acima do acaso.

4.7 Qualidade do *Corpus*

4.7.1 Distribuição das rubricas

As rubricas de qualidade (1-5) atribuídas pelo Gemini como avaliador (Prompt 3) revelam diferenças substanciais entre geradores. A Tabela 4.11 apresenta as estatísticas descritivas.

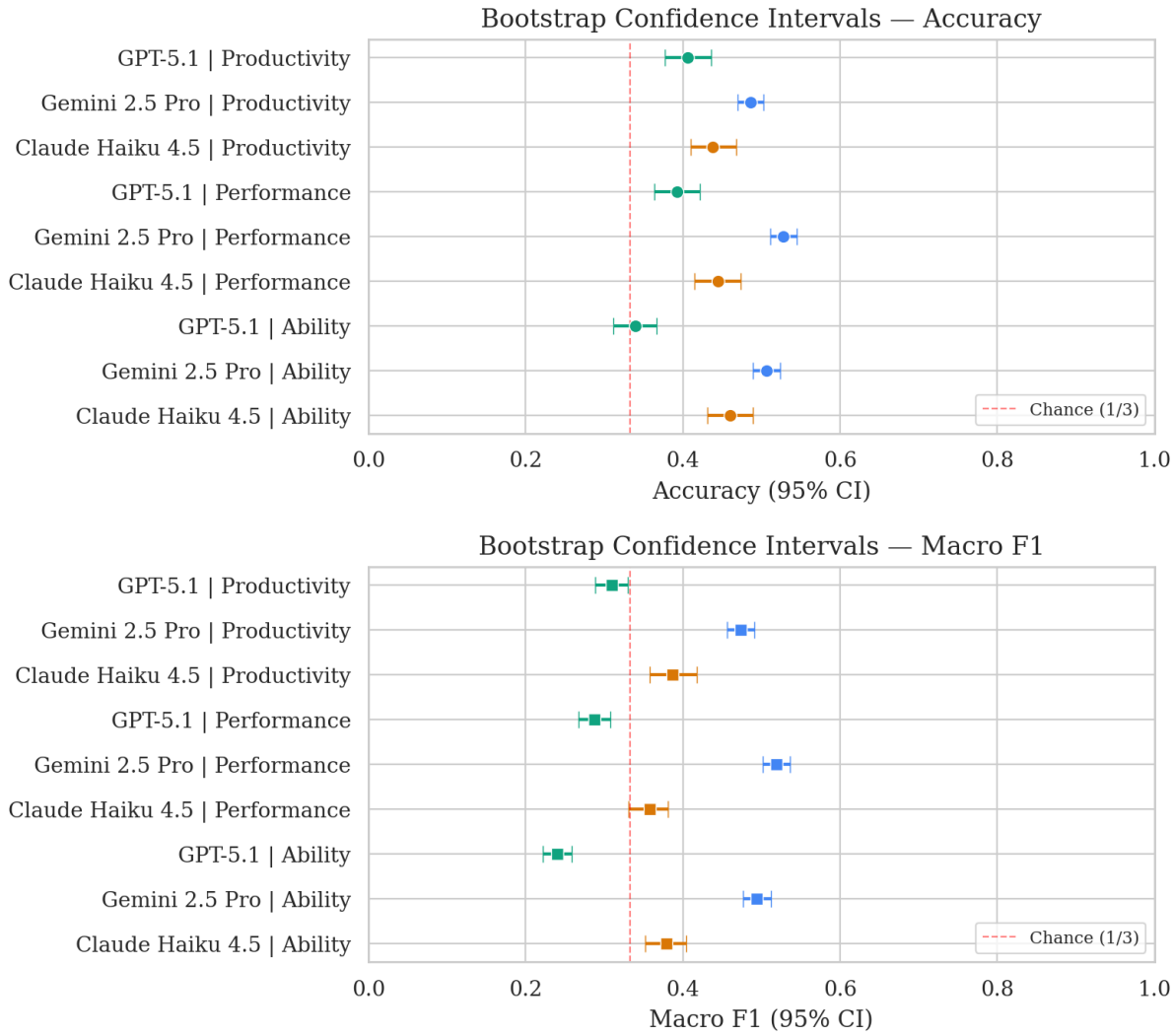

Figura 4.12 – *Forest plots* com intervalos de confiança 95% para *accuracy* (esquerda) e *macro F1* (direita) por inferidor e dimensão.

Tabela 4.11 – Estatísticas descritivas das rubricas de qualidade por provedor gerador.

Gerador	**Coerência** ($\mu \pm \sigma$)	**Realismo** ($\mu \pm \sigma$)	**Div. de Voz** ($\mu \pm \sigma$)
Claude Haiku 4.5	3,54 ± 1,15	3,98 ± 0,99	3,75 ± 1,27
Gemini 2.5 Pro	4,53 ± 0,79	4,39 ± 0,76	4,56 ± 0,68
GPT-5.1	4,57 ± 0,80	4,74 ± 0,66	4,82 ± 0,65

O GPT-5.1 recebe as melhores rubricas de qualidade (coerência 4,57; realismo 4,74; diversidade de voz 4,82), seguido de perto pelo Gemini. O Claude Haiku apresenta *scores* significativamente inferiores, especialmente em coerência (3,54). A Figura 4.13 apresenta a distribuição das rubricas.

4.7.2 O paradoxo qualidade vs. recuperabilidade

Um resultado contraintuitivo emerge ao correlacionar a qualidade dos diálogos com a acurácia de inferência. A Tabela 4.12 apresenta as correlações.

Tabela 4.12 – Correlação entre rubricas de qualidade e acurácia de inferência por time.

Rubrica	**Pearson** r	**Spearman** ρ	p **(Spearman)**
Coerência	+0,135	+0,063	0,117 (ns)
Realismo	−0,076	−0,166	<0,001 ***
Diversidade de Voz	−0,229	−0,317	<0,001 ***

A diversidade de voz apresenta correlação negativa significativa com a acurácia de inferência ($\rho = -0{,}271$, $p < 0{,}001$). Isso significa que diálogos com vozes mais distintas entre participantes são mais difíceis de inferir, pois quando cada participante tem um estilo linguístico muito diferenciado, os padrões comportamentais associados às variáveis latentes ficam confundidos com idiossincrasias estilísticas.

O realismo também correlaciona negativamente ($\rho = -0{,}121$, $p = 0{,}002$), sugerindo que diálogos mais realistas são mais difíceis de analisar para inferência de atributos latentes. Esse paradoxo tem implicações importantes: o GPT-5.1 gera os diálogos de maior qualidade percebida, mas seus diálogos são os mais difíceis de inferir. O Gemini, com qualidade ligeiramente inferior, produz diálogos onde o sinal latente é mais recuperável. Isso sugere um ***trade-off*** **entre naturalidade e controlabilidade**: diálogos mais naturais e diversificados diluem os sinais comportamentais que codificam as variáveis latentes, conforme documentado na literatura de geração controlada de texto (DATHATHRI et al., 2020; YANG; KLEIN, 2021). A Figura 4.14 ilustra essa relação.

4.8 Análise Detalhada da Validação P3

A validação P3 (Prompt 3) avalia a qualidade dos diálogos gerados segundo critérios estruturais e semânticos, identificando problemas (*issues*) e atribuindo rubricas de qualidade. Esta seção apresenta uma análise aprofundada dos resultados de validação.

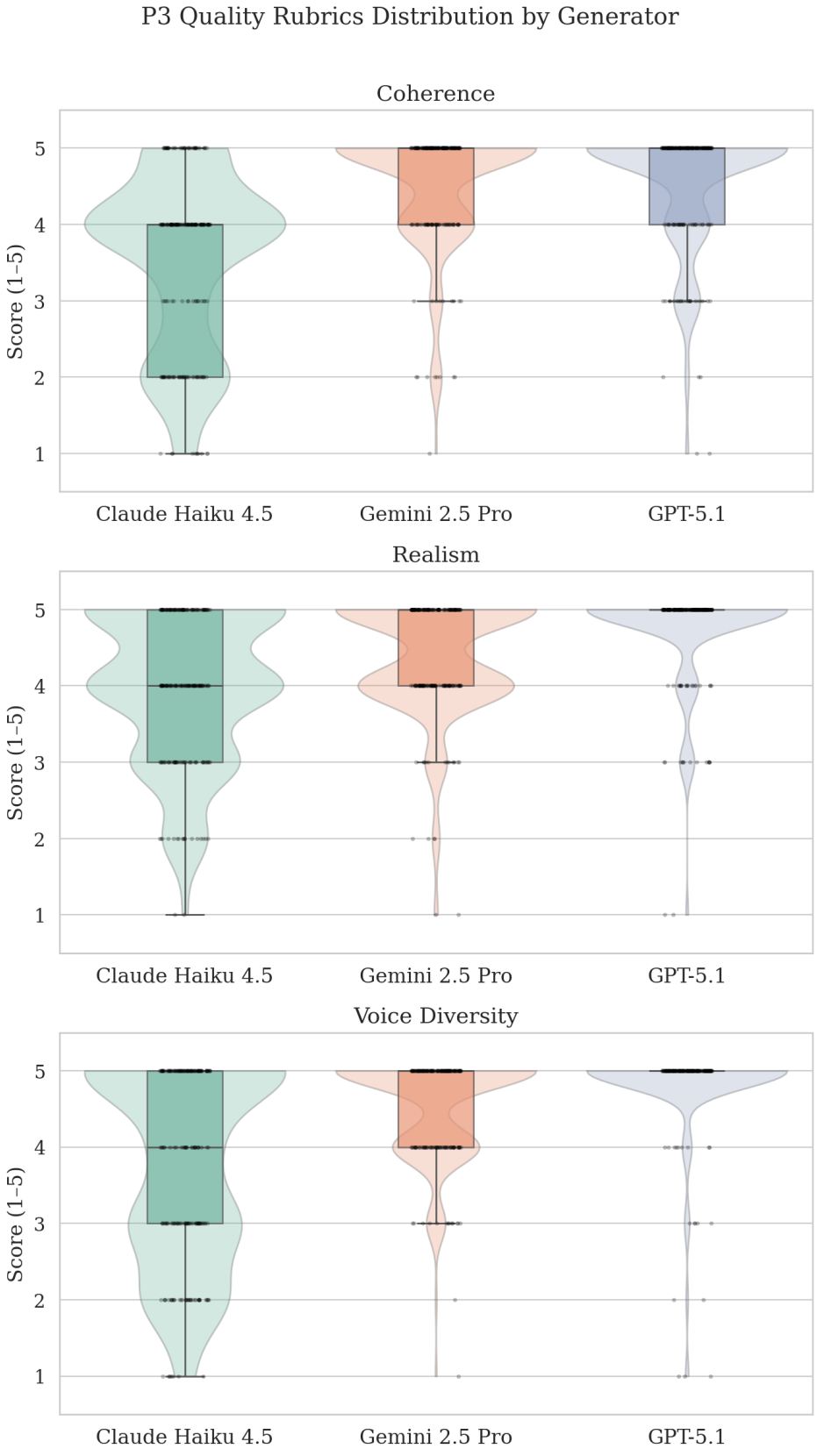

Figura 4.13 – Distribuição das rubricas de qualidade por provedor gerador.

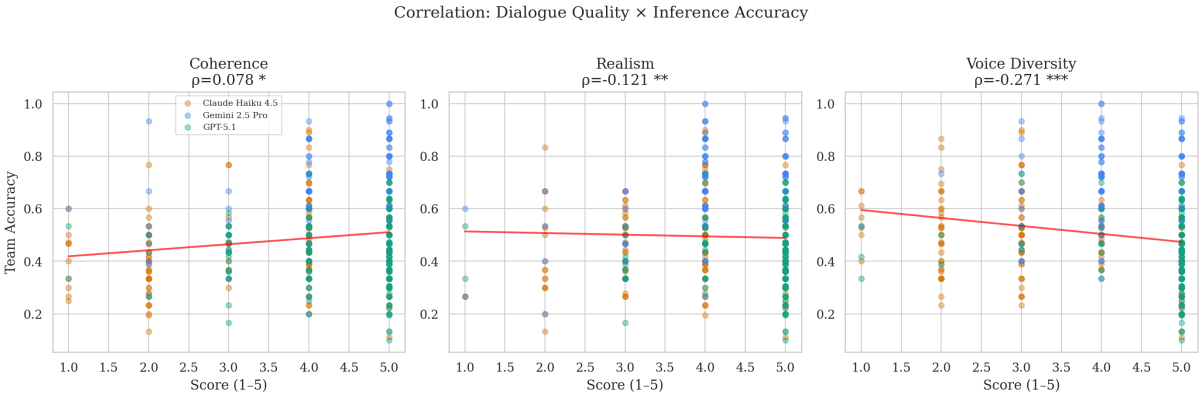

Figura 4.14 – Correlação entre rubricas de qualidade e acurácia de inferência por time.

4.8.1 Taxa de validação por gerador

A taxa de validação indica a proporção de diálogos considerados válidos (`is_valid=true`) pelo avaliador Gemini. A Tabela 4.13 apresenta os resultados.

Tabela 4.13 – Taxa de validação por provedor gerador.

Gerador	**Total**	**Válidos**	**Inválidos**	**Taxa (%)**
Claude Haiku 4.5	210	189	21	90,00
Gemini 2.5 Pro	210	205	5	97,62
GPT-5.1	210	208	2	99,05
Total	630	602	28	95,56

O GPT-5.1 apresenta a maior taxa de validação (99,05%), seguido pelo Gemini (97,62%) e Claude (90,00%). A taxa global de 95,56% indica que a grande maioria dos diálogos gerados atende aos critérios mínimos de qualidade estrutural. A diferença de ≈9pp entre Claude e GPT-5.1 sugere que o Claude Haiku tem maior dificuldade em cumprir as restrições estruturais do *prompt* de geração. A Figura 4.15 ilustra essa comparação.

4.8.2 Distribuição de *issues* por código

Os *issues* identificados durante a validação são categorizados por código e severidade. A Tabela 4.14 apresenta a distribuição e a Figura 4.16 a visualização correspondente.

O *issue* mais frequente é `COUNT` (54,26%), referente a violações de contagem, *threads* ou mensagens fora dos limites especificados. A alta prevalência indica que os LLMs têm dificuldade em controlar precisamente a quantidade de conteúdo gerado.

O segundo mais frequente é `ARTIFACT_MISSING` (22,95%), indicando referências a artefatos que deveriam aparecer nos diálogos, mas estão ausentes. Todos os 453 casos são de

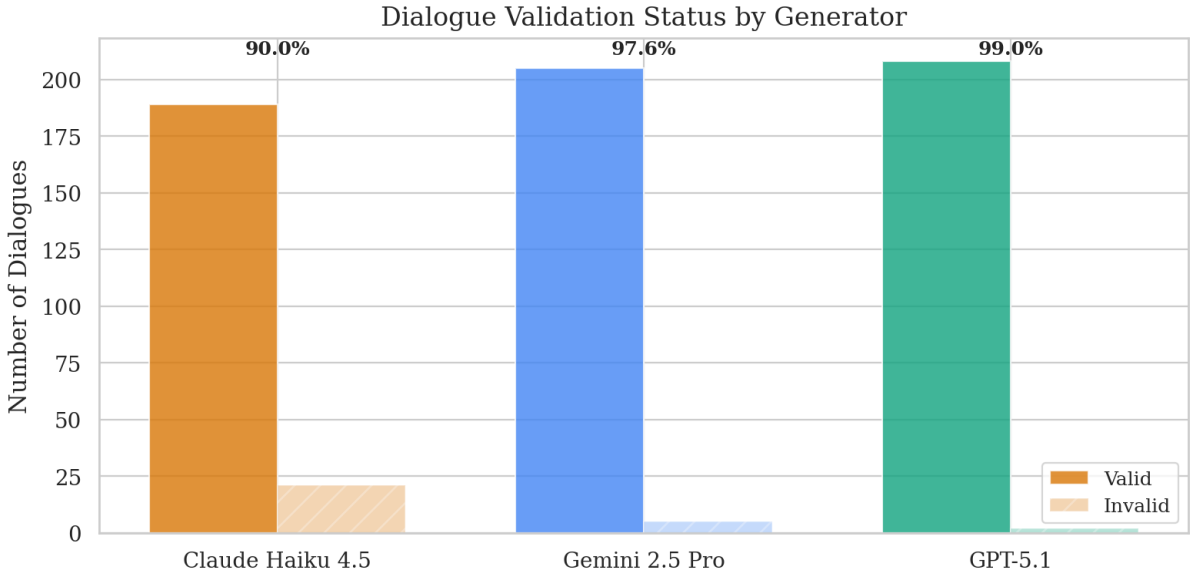

Figura 4.15 – Taxa de validação por provedor gerador.

Tabela 4.14 – Distribuição de *issues* por código e severidade.

Código	**Total**	**HIGH**	**MID**	**LOW**	**%**
COUNT	1071	982	24	65	54,26
ARTIFACT_MISSING	453	453	0	0	22,95
LEAK	241	240	0	1	12,21
VOICE_SIMILARITY	57	0	55	2	2,89
INVALID_REF	46	29	13	4	2,33
PARTICIPATION	46	46	0	0	2,33
INCONSISTENCY	28	4	12	12	1,42
DECISION_MISSING	23	23	0	0	1,17
SCHEMA	5	5	0	0	0,25
INCOMPLETE	4	2	1	1	0,20

severidade HIGH, indicando que este é um problema crítico para a fidelidade do diálogo ao *blueprint*.

O *issue* `LEAK` (12,21%) é particularmente relevante para a validade do experimento, pois representa vazamento de informação, termos proibidos ou indicações explícitas de níveis de atributos latentes que não deveriam aparecer no texto. A presença de 241 casos sugere que os LLMs ocasionalmente mencionam explicitamente características que deveriam ser apenas implícitas.

4.8.3 *Issues* por gerador

A distribuição de *issues* varia significativamente entre geradores. A Tabela 4.15 apresenta os resultados e a Figura 4.17 a visualização.

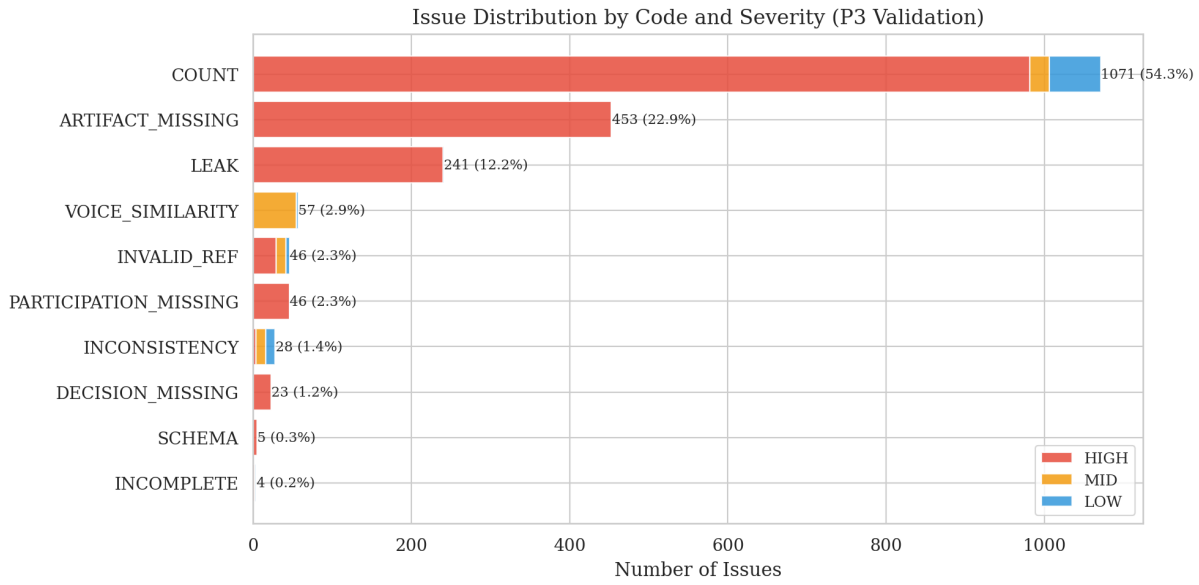

Figura 4.16 – Distribuição de *issues* por código e severidade.

Tabela 4.15 – Distribuição de *issues* por provedor gerador.

Gerador	**Times**	**Total**	**Issues/Time**	**HIGH**	**MID**
Claude Haiku 4.5	105	958	9,12	875	60
Gemini 2.5 Pro	105	559	5,32	519	20
GPT-5.1	105	457	4,35	390	25

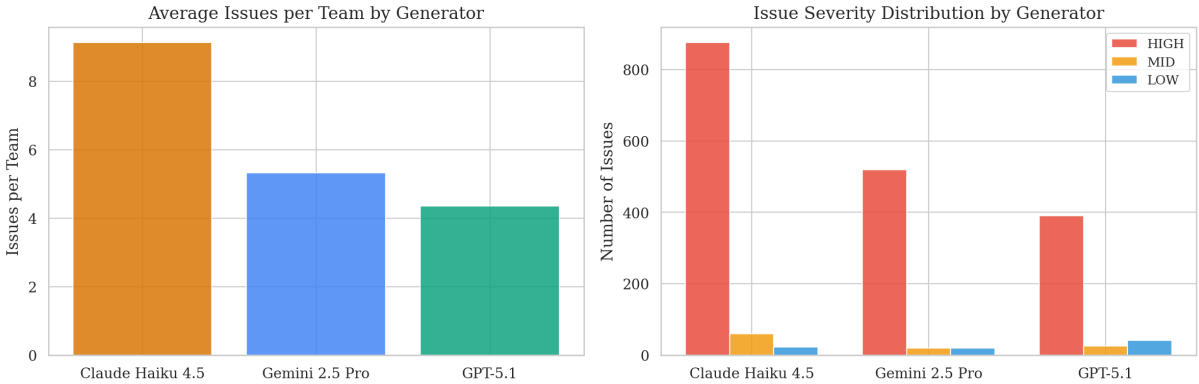

Figura 4.17 – Distribuição de *issues* por provedor gerador.

O Claude Haiku gera mais do que o dobro de *issues* por time (9,12) comparado ao GPT-5.1 (4,35). Isso é consistente com sua menor taxa de validação e sugere que o Claude Haiku tem maior dificuldade em seguir as instruções estruturais complexas do *prompt* de geração.

A Figura 4.18 apresenta o *heatmap* de códigos de *issues* por gerador, evidenciando que o Claude Haiku lidera em todas as categorias, com destaque para `LEAK`.

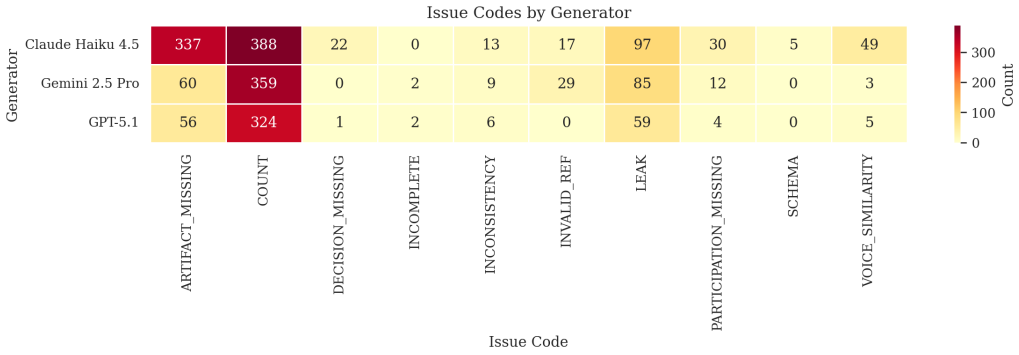

Figura 4.18 – *Heatmap* de códigos de *issues* por provedor gerador.

4.8.4 Rubricas detalhadas por gerador

A Figura 4.19 apresenta o gráfico radar das médias das rubricas por gerador, e a Figura 4.20 os *boxplots* comparativos.

O GPT-5.1 e Gemini apresentam medianas de 5 em todas as rubricas, enquanto o Claude tem mediana de 4. A variabilidade do Claude (mínimo = 1 em todas as rubricas) indica que alguns de seus diálogos são de qualidade muito baixa.

4.8.5 Impacto da validação na acurácia de inferência

Uma questão central é: diálogos válidos produzem inferências mais precisas? A Tabela 4.16 apresenta a comparação e a Figura 4.21 a visualização.

Tabela 4.16 – Métricas de inferência: diálogos válidos vs. inválidos.

Status	Dimensão	n	Acc.	Macro F1	MCC
Válido	Habilidade	5247	0,465	0,434	0,239
	Produtividade	5247	0,456	0,428	0,200
	Desempenho	5247	0,483	0,453	0,254
Inválido	Habilidade	261	0,441	0,405	0,174
	Produtividade	261	0,544	0,500	0,298
	Desempenho	261	0,521	0,482	0,281

Diálogos válidos apresentam métricas consistentemente superiores aos inválidos em habilidade: *accuracy* +2,4pp; *macro F1* +2,9pp; MCC +6,5pp. Interessantemente, para produtividade

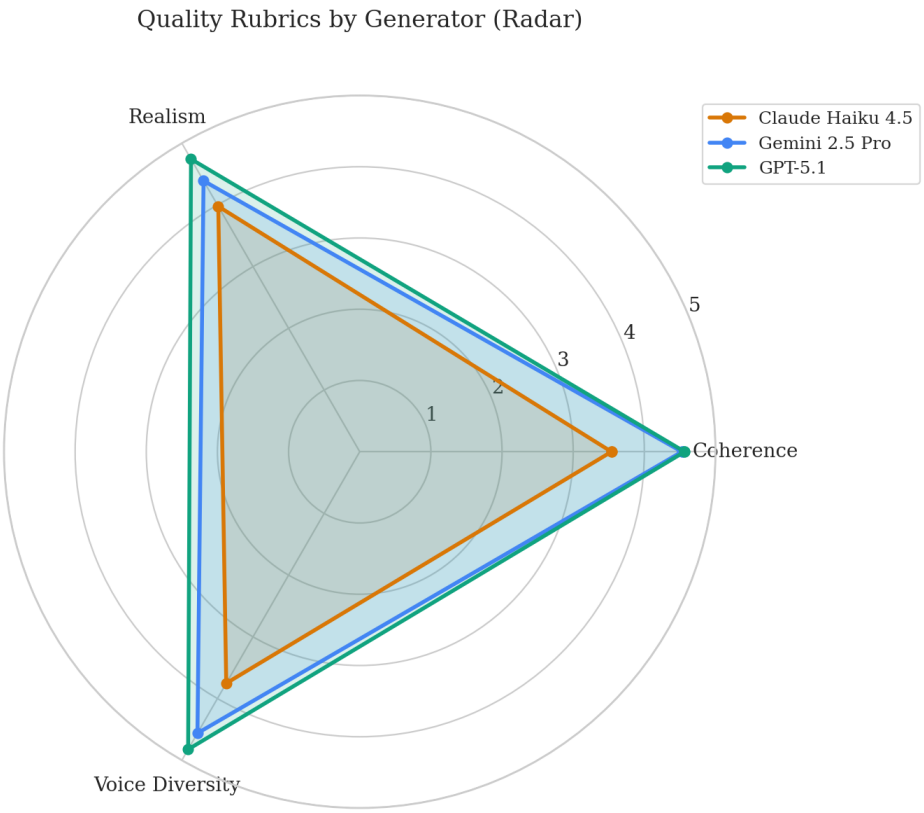

Figura 4.19 – Gráfico radar das rubricas de qualidade por provedor gerador.

e desempenho, os diálogos inválidos apresentam métricas superiores, o que pode indicar que alguns tipos de "violações" (como vazamentos) podem paradoxalmente facilitar a inferência ao tornar os sinais mais explícitos. Essa diferença confirma que a qualidade estrutural dos diálogos impacta a recuperabilidade dos atributos latentes de forma complexa.

4.8.6 Correlação entre rubricas, *issues* e acurácia

A matriz de correlação de Spearman (Figura 4.22) revela relações importantes entre as variáveis de qualidade e a acurácia de inferência.

Os principais achados são:

- **Rubricas vs. *issues*** ($\rho = -0{,}53$): Forte correlação negativa entre a média das rubricas e o número de *issues*. Diálogos de maior qualidade percebida apresentam menos problemas estruturais.
- **Diversidade de voz vs. acurácia** ($\rho = -0{,}317$): Confirmação do paradoxo qualidade-recuperabilidade. Diálogos com vozes mais distintas são mais difíceis de inferir.
- ***Issues* vs. acurácia** ($\rho \approx 0$): Surpreendentemente, o número de *issues* não se correlaciona diretamente com a acurácia, sugerindo que o impacto é mediado por outros fatores.

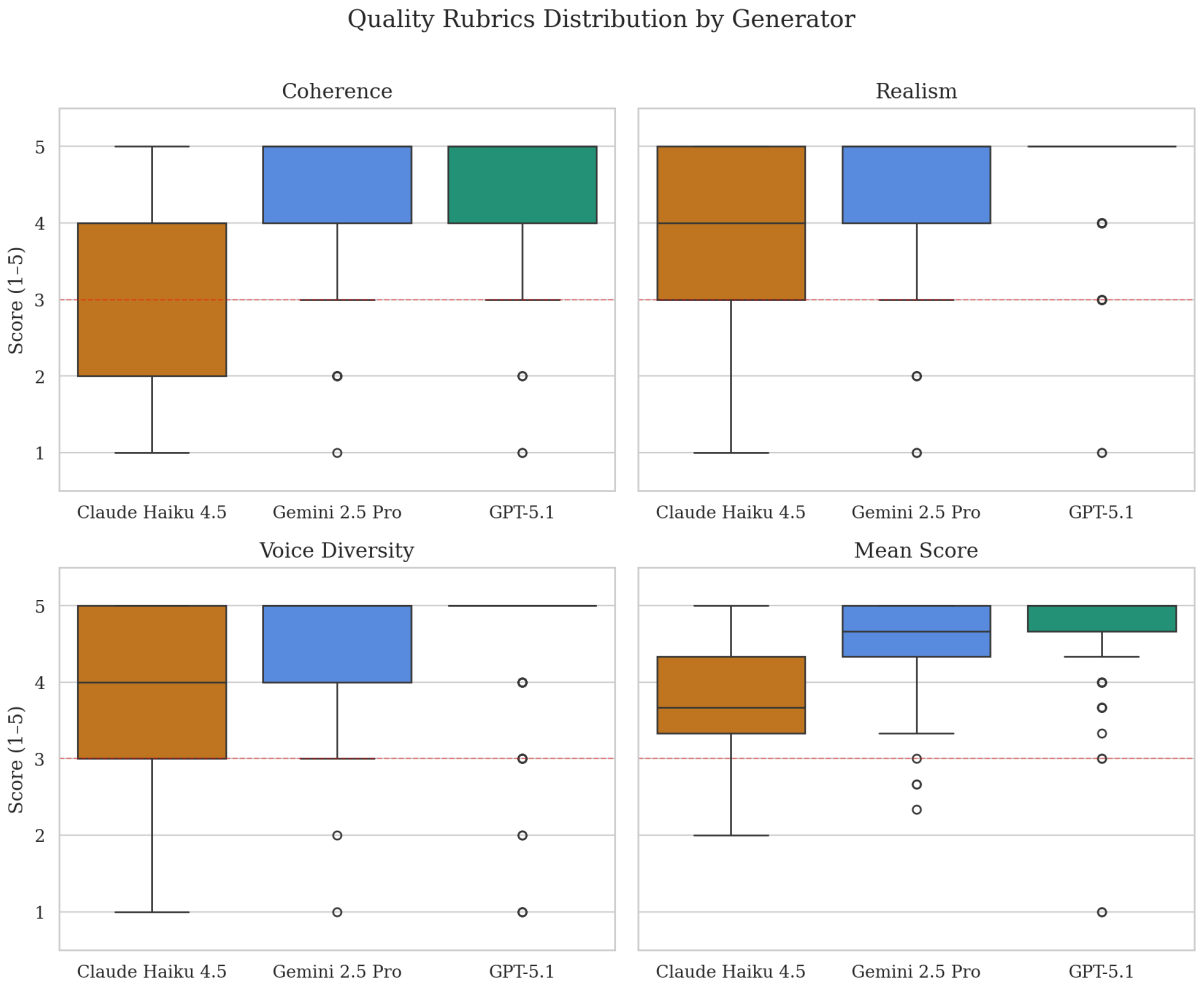

Figura 4.20 – *Boxplots* das rubricas de qualidade por provedor gerador.

A Figura 4.23 apresenta os *scatter plots* de número de *issues* e média das rubricas versus acurácia por time.

4.8.7 Implicações para o *pipeline*

Os resultados da análise de validação sugerem várias melhorias potenciais:

1. **Seleção de gerador**: O GPT-5.1 produz diálogos de maior qualidade estrutural e com menos vazamentos, enquanto o Gemini produz diálogos com sinais latentes mais recuperáveis. A escolha do gerador deve considerar esse *trade-off*.

2. **Filtragem de diálogos**: Excluir diálogos com *issues* de severidade HIGH (especialmente `LEAK`) pode melhorar a qualidade do *corpus* para treinamento de modelos de inferência.

3. **Ajuste de *prompts***: A alta prevalência de *issues* `COUNT` sugere que os *prompts* de geração devem ser ajustados para especificar limites de contagem de forma mais explícita ou flexível.

Figura 4.21 – Comparação de métricas entre diálogos válidos e inválidos.

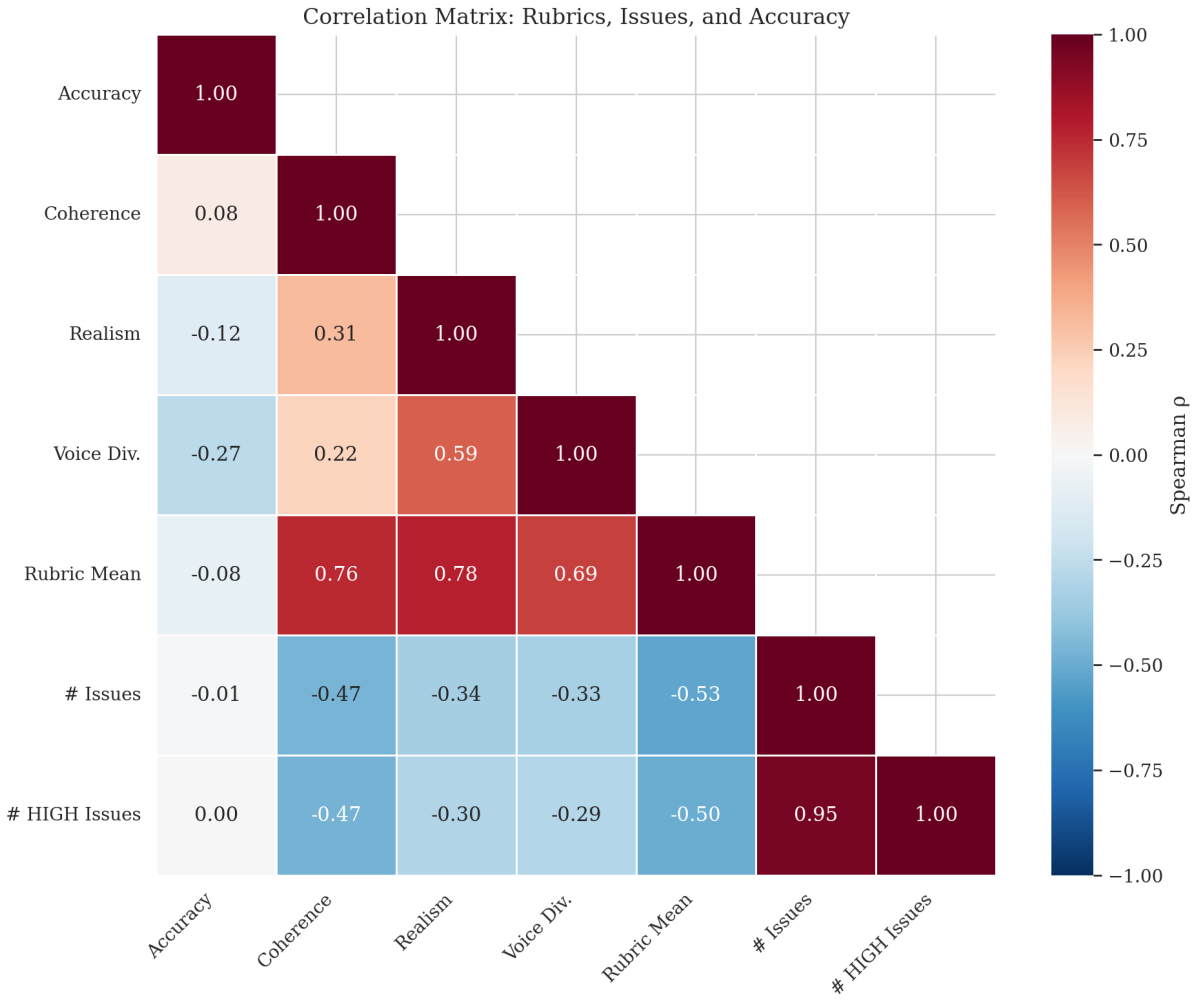

Figura 4.22 – Matriz de correlação de Spearman entre rubricas, *issues* e acurácia.

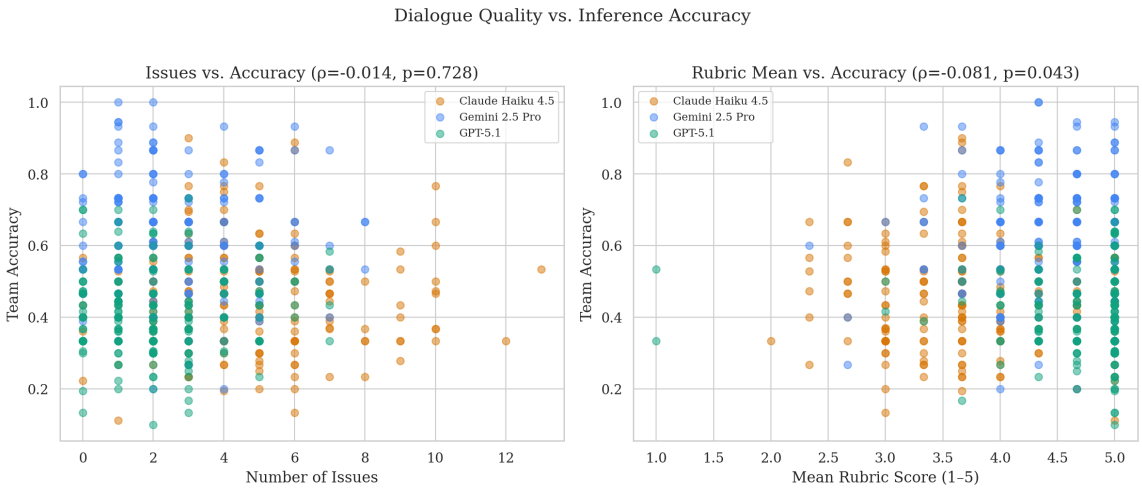

Figura 4.23 – Relação entre qualidade do diálogo e acurácia de inferência.

4. **Mitigação de vazamentos**: Os 241 casos de `LEAK` indicam necessidade de reforçar as instruções antivazamento nos *prompts* de geração, especialmente para o Claude.

4.9 Análise por Banda de Afinidade

Os times foram formados com três bandas de afinidade baseadas na similaridade de trajetória profissional (Weighted Jaccard). A Tabela 4.17 apresenta as métricas por banda e a Figura 4.24 o *heatmap* correspondente.

Tabela 4.17 – Métricas por banda de afinidade e dimensão.

Banda	Dimensão	n	Acc.	Macro F1	MCC	κ quad.
HIGH	Habilidade	3220	0,510	0,480	0,289	0,418
	Produtividade	3220	0,481	0,458	0,230	0,360
	Desempenho	3220	0,484	0,465	0,255	0,419
MID	Habilidade	2118	0,397	0,364	0,156	0,255
	Produtividade	2118	0,434	0,397	0,169	0,256
	Desempenho	2118	0,478	0,433	0,242	0,330
LOW	Habilidade	170	0,418	0,394	0,218	0,243
	Produtividade	170	0,388	0,355	0,130	0,196
	Desempenho	170	0,571	0,501	0,373	0,301

A banda HIGH (times com trajetórias muito similares) apresenta os melhores resultados em habilidade e produtividade, sugerindo que a homogeneidade de trajetória facilita a diferenciação de atributos latentes, pois quando os participantes têm *backgrounds* similares, as diferenças comportamentais são mais atribuíveis às variáveis latentes do que a diferenças de *expertise*. A banda LOW em desempenho (0,571 de *accuracy*) é um *outlier* positivo, mas com apenas 170 observações.

4.10 Exemplo de Diálogo Gerado e Disponibilização do *corpus*

Para ilustrar concretamente o tipo de conteúdo produzido pelo *pipeline*, esta seção apresenta um exemplo resumido de diálogo gerado pelo Gemini 2.5 Pro para um time da banda de afinidade HIGH.

4.10.1 Contexto do time

O time `HIGH_00186` foi formado com 5 participantes para o projeto "Lançamento de Programa de Vendas de Membros por Níveis". A composição do time inclui:

- **Participante 54019**: Gerente de membros

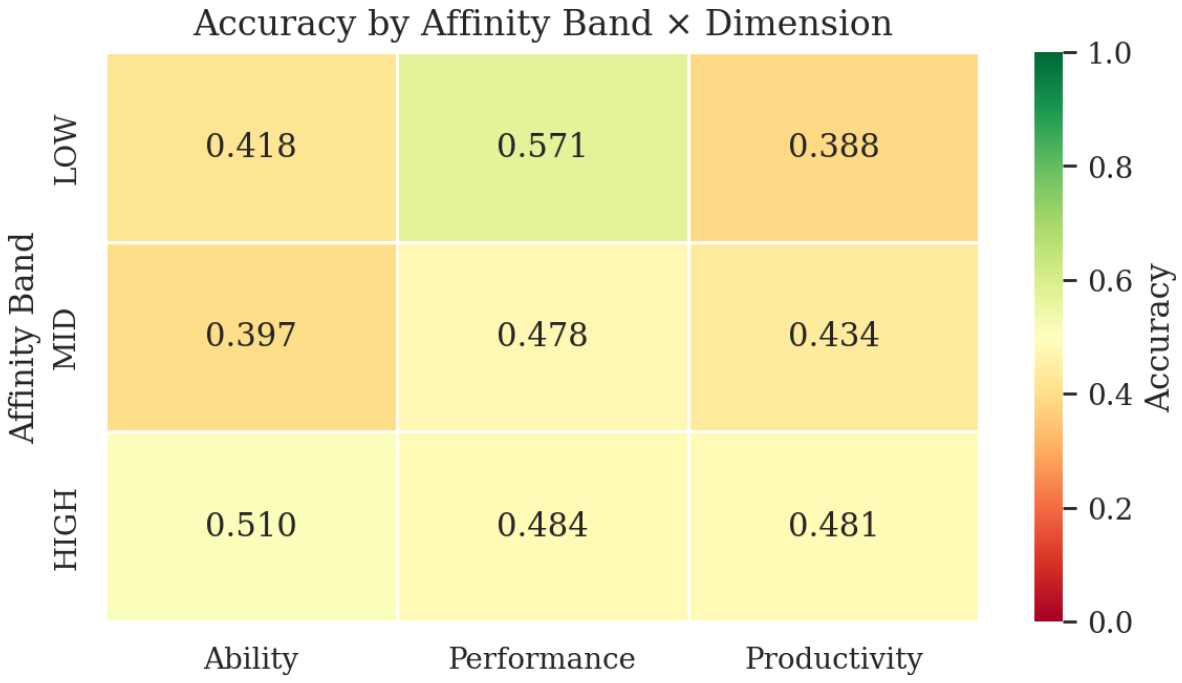

Figura 4.24 – *Heatmap* de métricas por banda de afinidade e dimensão.

- **Participante 46183**: Representante comercial de vendas
- **Participante 52934**: Representante comercial de vendas
- **Participante 488135**: Representante comercial de vendas
- **Participante 31942**: Representante comercial de vendas

O projeto tinha como objetivo definir uma estrutura de níveis de membros com preços e benefícios, criar um roteiro de vendas padronizado, e estabelecer métricas de acompanhamento.

4.10.2 Trecho do diálogo

O diálogo foi estruturado em 5 *threads* temáticas. A seguir, um trecho traduzido da primeira *thread*, cujo objetivo era alinhar a estrutura de níveis, preços e restrições de política:

> **[54019 - Gerente de membros]:** “Equipe, preparei o primeiro rascunho dos níveis de membros para T1. O documento está aqui: A1. Por favor, deem uma olhada e compartilhem suas opiniões.”
>
> **[46183 - Rep. comercial]:** “Obrigado. Lendo agora. Pensamento inicial sobre o Nível 3 (‘Platinum’): o benefício de ‘horas ilimitadas de consultoria’ parece arriscado e pode ser mal interpretado. Precisamos definir limites claros para evitar problemas

de conformidade, que faz parte de T2. Sugiro mudarmos para ‘10 horas dedicadas de consultoria por mês’. É mais concreto e seguro.”

[52934 - Rep. comercial]: “Boa observação, @46183. Isso está alinhado com as melhores práticas de proteção ao consumidor. Sobre preços, o salto do Nível 1 para o Nível 2 parece pequeno para o valor agregado. Podemos canibalizar as vendas do Nível 2. E se aumentarmos o preço do Nível 2 em 15% mas adicionarmos mais um recurso exclusivo para justificar?”

[31942 - Rep. comercial]: “Desculpe, onde posso encontrar o documento?”

[488135 - Rep. comercial]: “@31942 Está no link da primeira mensagem: A1.”

[46183 - Rep. comercial]: “Ok, vamos resumir para fechar isso. Estamos avançando com 3 níveis. Os benefícios principais estão aprovados com a mudança nas horas de consultoria do Nível 3. Isso é D1. Para preços, estamos aumentando o Nível 2 em 15% e adicionando 10% de desconto anual. Isso é D2. Todos alinhados?”

4.10.3 Padrões comportamentais observados

O trecho ilustra como os atributos latentes se manifestam nos comportamentos:

- **Participante 46183** demonstra alta proatividade, identificando riscos, propondo soluções concretas e conduzindo a tomada de decisão. Esse padrão é consistente com atributos latentes elevados.
- **Participante 52934** contribui com análises estratégicas e sugestões de melhoria, indicando engajamento ativo.
- **Participante 31942** apresenta participação mais passiva, com perguntas básicas, padrão associado a atributos latentes mais baixos.
- **Participante 488135** atua como facilitador, respondendo dúvidas e oferecendo suporte aos colegas.

Esse exemplo demonstra como o *pipeline* gera diálogos estruturados que simulam interações profissionais realistas, onde os comportamentos dos participantes refletem, de forma implícita, seus atributos latentes de habilidade, produtividade e desempenho.

O *corpus* completo de diálogos sintéticos, incluindo os dossiês de times, cenários experimentais, *blueprints* de planejamento e resultados de inferência, está disponível publicamente no repositório GitHub: <https://github.com/BeatrizOrlandi/behavioral-dialog-corpus>. O repositório contém todos os artefatos gerados pelo *pipeline*, permitindo a reprodutibilidade dos experimentos e a reutilização dos dados em pesquisas futuras.

4.11 Discussão

4.11.1 Principais achados

Os resultados permitem sintetizar sete achados centrais:

1. **O sinal latente é parcialmente recuperável**: a *accuracy* global de 46-48% supera o *baseline* de chance (33,3%), e o cenário estruturado supera o aleatório em 9-20pp, confirmando que LLMs conseguem codificar e recuperar atributos latentes em diálogos sintéticos.

2. **Viés sistemático para HIGH**: todos os inferidores sobre-predizem HIGH (62-77% das predições vs. 29-31% de prevalência real), resultando em *recall* elevado para HIGH mas *recall* mais baixo para LOW e MID. Esse *positivity bias* continua sendo um limitador da acurácia (WANG et al., 2023).

3. **Gemini domina como gerador e inferidor**: o Gemini 2.5 Pro produz diálogos com sinais mais recuperáveis (*accuracy* 57-64% como gerador) e é o inferidor mais equilibrado (melhor *macro F1* e MCC).

4. ***Trade-off* qualidade vs. controlabilidade**: diálogos com maior diversidade de voz e realismo são mais difíceis de inferir ($\rho = -0{,}271$), sugerindo que naturalidade e controlabilidade são objetivos parcialmente conflitantes (DATHATHRI et al., 2020).

5. ***Self-evaluation* supera *cross-evaluation***: a vantagem de 2-3pp em *macro F1* sugere viés de autoavaliação ou consistência interna de vocabulário comportamental (ZHENG et al., 2023a).

6. **Qualidade estrutural impacta a inferência**: a análise de validação P3 mostra que a qualidade dos diálogos afeta a recuperabilidade dos atributos latentes, embora a relação seja complexa.

7. **Vazamentos comprometem a validade**: os 241 casos de `LEAK` identificados na validação P3 indicam que os LLMs ocasionalmente vazam informações sobre os atributos latentes, o que pode contaminar a inferência inversa.

4.11.2 Limitações

Algumas limitações devem ser consideradas na interpretação dos resultados:

- **Viés do avaliador P3**: as rubricas de qualidade e a validação estrutural foram atribuídas exclusivamente pelo Gemini, introduzindo potencial viés de avaliador único (GAO et al., 2024).

- ***Design* assimétrico de inferência**: OpenAI e Claude fazem apenas *self-evaluation*, enquanto Gemini avalia todos os geradores. Isso impede comparações pareadas completas (o teste Q de Cochran não pôde ser computado).
- **Tamanho amostral desbalanceado por banda**: a banda LOW tem apenas 170 observações (vs. 3.220 para HIGH), limitando a potência estatística das comparações por afinidade.
- **Prevalência de *issues* `COUNT`**: a alta frequência de violações de contagem (54% dos *issues*) sugere que os *prompts* de geração podem ser refinados para especificar limites de forma mais robusta.
- **Custo computacional**: a execução completa do *pipeline* totalizou USD 283,18 em chamadas de API e mais de 102 milhões de tokens processados (Tabela 3.3), o que limita a reprodutibilidade e a escalabilidade dos experimentos. O custo elevado restringe a possibilidade de ampliar o número de cenários, replicações ou provedores avaliados, e pode inviabilizar a adoção da metodologia em contextos com recursos financeiros limitados.

5 Considerações Finais

Este capítulo sintetiza as conclusões decorrentes do desenvolvimento e dos resultados obtidos nesta monografia, explicitando em que medida os objetivos foram atingidos (Seção 5.1), sumarizando os achados centrais e apontando direções para trabalhos futuros (Seção 5.2).

5.1 Conclusão

O presente trabalho teve como objetivo geral a construção e avaliação de um *corpus* sintético de conversas empresariais, gerado por LLMs, que representasse de forma coerente e diversificada as interações entre colaboradores em distintos perfis e contextos organizacionais, com variáveis latentes controláveis e recuperáveis por inferência inversa.

A seguir, avalia-se o grau de atendimento de cada objetivo específico definido no Capítulo 1, à luz das evidências produzidas nos Capítulos 3 e 4.

1. **Metodologia de simulação de variáveis latentes.** Este objetivo foi alcançado. Desenvolveu-se uma abordagem parametrizada para simulação controlada de desempenho, produtividade e habilidade por experiência profissional, combinando traço individual, ajuste por trajetória (*skill_fit*), senioridade, efeito de time, termo temporal e componente serial AR(1) (Seção 3.3.1). A validação estatística foi conduzida com $R = 25$ replicações Monte Carlo por cenário, e a estabilidade das estimativas foi confirmada por MCSE (Equação 3.1). A seleção de repetição representativa por *medoid* garantiu coerência interna dos perfis utilizados na geração. A análise prévia demonstrou que a agregação por média suaviza as flutuações de ruído, justificando a escolha de apenas dois cenários para a etapa gerativa.

2. **Taxonomia de *personas* corporativas.** Este objetivo foi alcançado. Os perfis profissionais foram derivados da fusão do *Karrierewege* (SENGER et al., 2024) com o *Karrierewege_plus*, consolidados via DuckDB com deduplicação por chave composta e materialização em Parquet (Seção 3.2). Cada perfil contém sequência de experiências com cargos ESCO, habilidades e metadados de trajetória. As variáveis latentes simuladas foram agregadas por pessoa na tabela `per_id`, com médias, probabilidades empíricas de classe e rótulo final por dimensão (LOW/MID/HIGH), constituindo atributos funcionais e comunicacionais quantificáveis para condicionamento da geração. O *corpus* final compreende 541 membros únicos distribuídos em 105 times.

3. **Mecanismo de formação de times por similaridade textual.** Este objetivo foi alcançado. Implementou-se a representação de trajetórias por TF-IDF com n-gramas (1,2) e filtragem por `max_df/min_df`, com similaridade definida por Weighted Jaccard (Equação 3.2). Os

limiares foram calibrados por percentis de similaridade amostrada, definindo três bandas de afinidade: HIGH ($\geq P_{95}$), MID (P_{60}-P_{75}) e LOW (P_{35}-P_{45}), conforme descrito na Seção 3.4. O algoritmo guloso com índice invertido, checagem intra-grupo e limites de tamanho ($\mathrm{MIN} = 5$, $\mathrm{MAX} = 15$) produziu 105 times com relatórios de descarte para auditoria de cobertura. A análise por banda de afinidade (Seção 4.9) confirmou que a homogeneidade de trajetória (banda HIGH) facilita a diferenciação de atributos latentes, com MCC de 0,230-0,289 contra 0,130-0,218 na banda LOW.

4. **Protocolo reprodutível de geração condicionada.** Este objetivo foi alcançado. O *pipeline* de engenharia de *prompt* foi estruturado em três etapas: planejamento (Prompt A.1, com dossiê de projeto e roteiro de canal fixos por time), geração (Prompt A.2, com execução do roteiro sob condições latentes específicas) e validação (Prompt A.3, com checagem estrutural e regeneração quando necessário), conforme detalhado na Seção 3.5. A conversão de rótulos latentes em descritores comportamentais sem vazamento explícito foi implementada com regras interpretáveis e restrição dura contra menção de termos proibidos. A execução com três provedores LLM (GPT-5.1, Gemini 2.5 Pro e Claude Haiku 4.5) e registro completo de identificadores, modelos, versões e parâmetros assegura a reprodutibilidade. A taxa global de validação estrutural de 95,56% (Seção 4.8) confirma a viabilidade do protocolo, embora a presença de 241 casos de `LEAK` (12,21% dos *issues*) indique que a restrição antivazamento não é perfeitamente cumprida pelos geradores, constituindo uma limitação parcial deste objetivo.

5. ***Framework* de avaliação multinível do *corpus*.** Este objetivo foi alcançado em sua totalidade. O protocolo de avaliação em três níveis foi implementado e executado conforme especificado: (i) validação estrutural com checagens determinísticas de *schema*, contagem, distribuição de participação e testes de vazamento (Seção 4.8); (ii) avaliação de qualidade por rubricas (coerência, realismo e diversidade de voz) atribuídas por LLM avaliador, com agregação por mediana e comparação entre provedores (Seção 4.7); e (iii) inferência inversa dos atributos latentes com cálculo de acurácia, *balanced accuracy*, *macro F1*, MCC, kappa quadrático, MAE ordinal e *within-1 accuracy*, executada por três inferidores com comparação contra *ground truth* (Seção 4.2). A comparação entre modelos avaliadores de diferentes provedores foi realizada por meio de testes de McNemar com correção de Holm-Bonferroni e intervalos de confiança por *bootstrap* (Seção 4.6).

Em síntese, os cinco objetivos específicos foram alcançados, com ressalvas pontuais relativas ao vazamento parcial de informação latente nos diálogos gerados e ao *design* assimétrico de inferência que impediu comparações pareadas completas entre todos os provedores.

No que concerne ao objetivo geral, os resultados quantitativos confirmam a hipótese central do trabalho: LLMs conseguem gerar diálogos corporativos nos quais atributos latentes ocupacionais são parcialmente codificados no comportamento linguístico e recuperáveis por

inferência inversa. O cenário estruturado (`S_full_high_noise`) superou consistentemente o *baseline* aleatório (`S_random_with_latents`) em todas as dimensões e métricas, com ganhos de 9 a 20 pontos percentuais de MCC (Seção 4.3), validando que a correlação entre variáveis latentes e trajetória profissional é preservada no texto gerado e não constitui artefato metodológico. A *accuracy* global de 46-48% supera o *baseline* de chance (33,3%), e a *within-1 accuracy* de 90-92% indica que erros de duas classes de distância são raros.

Não obstante, três limitações qualificam o alcance dos resultados. Primeiro, o *positivity bias* dos inferidores, que sobre-predizem a classe HIGH em 62-77% dos casos contra uma prevalência real de 29-31% (Seção 4.5), compromete a discriminação das classes LOW e MID e limita o MCC global a 0,20-0,50. Segundo, o *trade-off* entre qualidade percebida e controlabilidade, evidenciado pela correlação negativa entre diversidade de voz e acurácia de inferência ($\rho = -0{,}317$, $p < 0{,}001$; Seção 4.7), indica que diálogos mais naturais e estilísticamente diversificados diluem os sinais comportamentais codificados. Terceiro, o custo computacional de USD 283,18 e mais de 102 milhões de tokens processados restringe a escalabilidade e a reprodutibilidade da metodologia em contextos com recursos limitados.

O Gemini 2.5 Pro destacou-se como o provedor mais adequado para o *pipeline*, tanto na geração (*accuracy* de 57-64% e MCC de 0,38-0,50) quanto na inferência com o melhor *macro F1* e MCC em todas as dimensões, enquanto o GPT-5.1 produziu os diálogos de maior qualidade percebida, mas com sinais latentes menos recuperáveis, reforçando o *trade-off* identificado.

5.2 Trabalhos futuros

Com base nas limitações identificadas e nos achados do presente trabalho, propõem-se as seguintes direções para continuidade:

1. **Calibração de predições:** implementar ajuste de limiares de decisão baseado na distribuição do *ground truth* (*Platt scaling*, *temperature scaling*) para mitigar o viés HIGH dos inferidores.

2. **Avaliação humana:** adicionar avaliação humana como *gold standard* para validar as rubricas do Prompt 3 e as inferências do Prompt 4, permitindo quantificar o viés do avaliador LLM.

3. ***Few-shot prompting*:** explorar *prompting* com exemplos de cada classe para mitigar o viés HIGH, fornecendo ao inferidor referências explícitas de comportamentos LOW e MID.

4. ***Chain-of-thought*:** investigar se o raciocínio explícito (*chain-of-thought prompting*) melhora a discriminação entre classes, forçando o inferidor a justificar cada atribuição antes de emitir o rótulo final.

5. ***Design* simétrico de inferência:** executar avaliação cruzada completa (todos os inferidores avaliando diálogos de todos os geradores) para permitir testes estatísticos pareados completos (Cochran's Q) e eliminar confusões entre efeito do gerador e efeito do inferidor.

6. **Ampliação do *corpus*:** expandir o número de cenários (incluindo variações de ruído e composição de times) e o número de replicações para aumentar a potência estatística, especialmente para a banda LOW de afinidade.

7. **Avaliação de modelos abertos e gratuitos:** investigar o desempenho de LLMs de código aberto e gratuitos (como LLaMA, Mistral e Gemma) como geradores e inferidores no *pipeline*, visando reduzir os custos de execução e viabilizar a reprodutibilidade e escalabilidade da metodologia sem dependência de APIs comerciais.

Referências

ABDULLIN, Y.; MOLLA-ALIOD, D.; OFOGHI, B.; YEARWOOD, J.; LI, Q. Synthetic dialogue dataset generation using LLM agents. In: *Proceedings of the Third Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*. Singapore: Association for Computational Linguistics, 2023. p. 181–191. Disponível em: <https://aclanthology.org/2023.gem-1.16/>.

ALENEZI, M.; BANITAAN, S.; MALETIC, J. I. Extracting and analyzing issue reports from open source issue tracking systems. In: IEEE. *Proceedings of the 10th Working Conference on Mining Software Repositories*. [S.l.], 2013. p. 235–238.

ANTHROPIC. *Models overview (Claude API documentation)*. 2026. Disponível em: <https://platform.claude.com/docs/en/about-claude/models/overview>.

BALARAMAN, V.; SHEIKHALISHAHI, S.; MAGNINI, B. Recent neural methods on dialogue state tracking for task-oriented dialogue systems: A survey. In: *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Singapore and Online: Association for Computational Linguistics, 2021. p. 239–251. Disponível em: <https://aclanthology.org/2021.sigdial-1.25/>.

BLEIH, H.; BELAID, A. Data fusion. *International Journal of Computer Science and Information Technology*, v. 8, n. 1, p. 1–13, 2016.

BROWN, T. B.; MANN, B.; RYDER, N.; SUBBIAH, M.; KAPLAN, J. D.; DHARIWAL, P.; NEELAKANTAN, A.; SHYAM, P.; SASTRY, G.; ASKELL, A. et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020. Disponível em: <https://arxiv.org/abs/2005.14165>.

BUDZIANOWSKI, P.; WEN, T.-H.; TSENG, B.-H.; CASANUEVA, I.; ULTES, S.; RAMADAN, O.; GAŠIĆ, M. MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In: RILOFF, E.; CHIANG, D.; HOCKENMAIER, J.; TSUJII, J. (Ed.). *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, 2018. p. 5016–5026. Disponível em: <https://aclanthology.org/D18-1547/>.

CAMILLERI, M. A. *Strategic Corporate Communication in the Digital Age*. Cham: Springer, 2021. Disponível em: <https://link.springer.com/book/10.1007/978-3-030-69791-9>.

CHICCO, D.; JURMAN, G. The advantages of the matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, v. 21, n. 1, p. 6, 2020.

COHEN, J. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, v. 70, n. 4, p. 213–220, 1968.

DATHATHRI, S.; MADOTTO, A.; LAN, J.; HUNG, J.; FRANK, E.; MOLINO, P.; YOSINSKI, J.; LIU, R. Plug and play language models: A simple approach to controlled text generation. In: *International Conference on Learning Representations (ICLR)*. [S.l.: s.n.], 2020.

DEMŠAR, J. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, v. 7, p. 1–30, 2006. Disponível em: <https://jmlr.csail.mit.edu/papers/v7/demsar06a.html>.

DEVLIN, J.; CHANG, M.-W.; LEE, K.; TOUTANOVA, K. Bert: Pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. [S.l.: s.n.], 2019. p. 4171–4186.

DIETTERICH, T. G. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, v. 10, n. 7, p. 1895–1923, 1998.

EFRON, B.; TIBSHIRANI, R. J. *An Introduction to the Bootstrap*. New York: Chapman & Hall/CRC, 1993.

European Commission. *European Skills, Competences, Qualifications and Occupations (ESCO)*. 2025. Multilingual classification of skills, competences, qualifications and occupations – ESCO portal. Disponível em: <https://esco.ec.europa.eu/en>.

FENG, Y.; LIU, J.; GLASS, M.; JIN, M.; MA, Y. Msdialog: A dataset for multi-turn question answering in online technical support. In: EUROPEAN LANGUAGE RESOURCES ASSOCIATION. *Proceedings of the 12th Language Resources and Evaluation Conference (LREC)*. 2020. p. 1967–1977. Disponível em: <https://aclanthology.org/2020.lrec-1.243>.

FERRARA, E. Should ChatGPT be biased? Challenges and risks of bias in large language models. *First Monday*, v. 28, n. 11, 2023. Disponível em: <https://firstmonday.org/ojs/index.php/fm/article/view/13346>. Disponível em: <https://firstmonday.org/ojs/index.php/fm/article/view/13346>.

GAO, J.; GALLEY, M.; LI, L. Neural approaches to conversational ai. *Foundations and Trends in Information Retrieval*, Now Publishers, v. 13, n. 2-3, p. 127–298, fev. 2019. ISSN 1554-0669. Disponível em: <https://doi.org/10.1561/1500000074>.

GAO, M.; HU, X.; RUAN, J.; PU, X.; WAN, X. *LLM-based NLG Evaluation: Current Status and Challenges*. 2024. ArXiv:2402.01383. Disponível em: <https://arxiv.org/abs/2402.01383>.

GOOGLE. *Gemini API: Available models*. 2026. Disponível em: <https://ai.google.dev/gemini-api/docs/models>.

GOYAL, M.; MAHMOUD, Q. H. A systematic review of synthetic data generation techniques using generative ai. *Electronics*, v. 13, n. 17, p. 3509, 2024. Open Access; disponível em: <https://www.mdpi.com/2079-9292/13/17/3509>.

HANDEL, M. J. The occupational information network (o*net): strengths and limitations. *Journal for Labour Market Research*, v. 49, p. 157–176, 2016.

HOLM, S. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, v. 6, n. 2, p. 65–70, 1979.

Hugging Face. *Karrierewege Dataset*. 2025. <https://huggingface.co/datasets/ElenaSenger/Karrierewege>. Acesso em: 6 jun. 2025.

JARKE, J.; BREITER, A. Data sovereignty and data literacy: Two sides of the same coin? *International Review of Information Ethics*, IRIE, v. 29, n. 1, p. 1–13, 2020. Disponível em: <https://www.i-r-i-e.net/inhalt/029/IRIE-29-01.pdf>.

JI, Z.; LEE, N. et al. Survey of hallucination phenomena in natural language generation. *Transactions of the Association for Computational Linguistics*, v. 11, p. 249–270, 2023. Disponível em: <https://aclanthology.org/2023.tacl-1.13>.

JORDON, J.; SZPRUCH, L.; HOUSSIAU, F.; BOTTARELLI, M.; CHERUBIN, G.; MAPLE, C.; COHEN, S. N.; WELLER, A. *Synthetic Data - what, why and how?* [S.l.], 2022. Commissioned by the Royal Society; arXiv:2205.03257. Disponível em: <https://arxiv.org/abs/2205.03257>.

JURAFSKY, D.; MARTIN, J. H. *Speech and Language Processing*. 3. ed. [S.l.]: Prentice Hall, 2023. Versão online: <https://web.stanford.edu/~jurafsky/slp3/>.

KAHN, R. L.; KATZ, D.; JACOBS, R. Communication, roles, and social systems. *Human Organization*, v. 23, n. 4, p. 366–377, 1964.

KAUR, H.; KAUR, K.; BHATIA, S.; SHARMA, S.; SHARMA, S. Ethical implications of conversational ai in healthcare: A review. *Computer Communications*, Elsevier, v. 166, p. 44–54, 2021.

KLIMT, B.; YANG, Y. The enron corpus: A new dataset for email classification research. In: *Machine Learning: ECML 2004*. [S.l.]: Springer, 2004. p. 217–226.

KOZLOWSKI, S. W. J.; ILGEN, D. R. Enhancing the effectiveness of work groups and teams. *Psychological Science in the Public Interest*, v. 7, n. 3, p. 77–124, 2006.

KUMMERFELD, J. K.; GOURAVAJHALA, S. R.; PEPER, J. J.; ATHREYA, V.; GUNASEKARA, C.; GANHOTRA, J.; PATEL, S. S.; POLYMENAKOS, L.; LASECKI, W. S. A large-scale corpus for conversation disentanglement. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*. [S.l.: s.n.], 2019.

LANDIS, J. R.; KOCH, G. G. The measurement of observer agreement for categorical data. *Biometrics*, v. 33, n. 1, p. 159–174, 1977.

LIU, Y.; LI, Y.; YANG, W.; ZHANG, J.; LIU, J. Synthetic dialogues: An empirical study of dialogue corpora construction via pre-trained language models. *arXiv preprint arXiv:2301.11916*, 2023. Disponível em: <https://arxiv.org/abs/2301.11916>.

MANNING, C. D.; RAGHAVAN, P.; SCHÜTZE, H. *Introduction to information retrieval*. [S.l.]: Cambridge university press, 2008.

MCNEMAR, Q. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, v. 12, n. 2, p. 153–157, 1947.

MELLO, C. A. d.; ARIMA, F. A.; NEVES, R. B. d. L. Formação e características de times, equipes ou grupos e características dos membros para atuar em sistemas dinâmicos. In: *Anais do XXXIV Simpósio de Engenharia de Produção*. Foz do Iguaçu, PR: Associação Brasileira de Engenharia de Produção (ABEPRO), 2020. Disponível em: <https://www.abepro.org.br/biblioteca/TN_STO_332_1329_37758.pdf>.

O*NET Resource Center. *The O*NET Content Model*. 2025. Marco conceitual do O*NET para identificar tipos de informações sobre trabalho. Disponível em: <https://www.onetcenter.org/content.html>.

OPENAI. *Models (documentation)*. 2026. Disponível em: <https://platform.openai.com/docs/models>.

PARK, J. S.; O'BRIEN, J. C.; CAI, C. J.; MORRIS, M. R.; LIANG, P.; BERNSTEIN, M. S. Generative agents: Interactive simulacra of human behavior. In: *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. [S.l.: s.n.], 2023. p. 1–22.

RASTOGI, A.; ZANG, X.; SUNKARA, S.; GUPTA, R.; KHAITAN, P. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. [S.l.: s.n.], 2020. v. 34, n. 05, p. 8689–8696.

RASTOGI, A.; ZANG, X.; SUNKARA, S.; GUPTA, R.; KHAITAN, P. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. [s.n.], 2020. v. 34, p. 8689–8696. Disponível em: <https://arxiv.org/abs/1909.05855>.

ROTUNDO, M.; SACKETT, P. R. The relative importance of task, citizenship, and counterproductive performance to global ratings of job performance: A policy-capturing approach. *Journal of Applied Psychology*, v. 87, n. 1, p. 66–80, 2002.

S, A. M.; DESHPANDE, T.; SCIENTISTS, I. D. *IBM HR Analytics Employee Attrition & Performance*. IEEE Dataport, 2023. Disponível em: <https://dx.doi.org/10.21227/2m1g-6v47>.

SENGER, E.; CAMPBELL, Y.; GOOT, R. van der; PLANK, B. KARRIEREWEGE: A large scale career path prediction dataset. *arXiv*, 2024. Disponível em: <https://arxiv.org/abs/2412.14612>.

SENGER, E.; CAMPBELL, Y.; GOOT, R. van der; PLANK, B. KARRIEREWEGE: A large scale career path prediction dataset. In: RAMBOW, O.; WANNER, L.; APIDIANAKI, M.; AL-KHALIFA, H.; EUGENIO, B. D.; SCHOCKAERT, S.; DARWISH, K.; AGARWAL, A. (Ed.). *Proceedings of the 31st International Conference on Computational Linguistics: Industry Track*. Abu Dhabi, UAE: Association for Computational Linguistics, 2025. p. 533–545. Disponível em: <https://aclanthology.org/2025.coling-industry.46/>.

SOKOLOVA, M.; LAPALME, G. A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, v. 45, n. 4, p. 427–437, 2009.

SOUDANI, H.; HASIBI, F.; KANOULAS, E. *A Survey on Recent Advances in Conversational Data Generation*. 2024. ArXiv:2405.13003. Disponível em: <https://arxiv.org/abs/2405.13003>. Disponível em: <https://arxiv.org/abs/2405.13003>.

SYVERSON, C. What determines productivity? *Journal of Economic Literature*, v. 49, n. 2, p. 326–365, 2011.

WANG, P.; LI, L.; CHEN, L.; CAI, Z.; ZHU, D.; LIN, B.; CAO, Y.; LIU, Q.; LIU, T.; SUI, Z. Large language models are not fair evaluators. *arXiv preprint arXiv:2305.17926*, 2023. Disponível em: <https://arxiv.org/abs/2305.17926>.

WINTERTON, J.; DEIST, F. D.-L.; STRINGFELLOW, E. *Typology of knowledge, skills and competences: clarification of the concept and prototype*. Luxembourg, 2006. CEDEFOP Reference series, 64. Disponível em: <https://www.cedefop.europa.eu/en/publications/3048>.

WITTGENSTEIN, L. *Tractatus Logico-Philosophicus*. London: Kegan Paul, Trench, Trubner & Co., 1922.

YAMASHITA, M.; DOM, B.; PUROHIT, H. Openresume: Advancing career trajectory modeling with anonymized and synthetic resume datasets. In: *2024 IEEE International Conference on Big Data (BigData)*. [S.l.]: IEEE, 2024. p. 6697–6706. Disponível em: <https://pike.psu.edu/publications/bigdata24-resume.pdf>.

YANG, K.; KLEIN, D. FUDGE: Controlled text generation with future discriminators. In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*. [S.l.: s.n.], 2021. p. 3511–3535.

YANG, S.; ISLAM, M. T. Ibm employee attrition analysis. *arXiv preprint arXiv:2012.01286*, 2020. Disponível em: <https://arxiv.org/abs/2012.01286>. Disponível em: <https://arxiv.org/abs/2012.01286>.

ZHANG, S.; LI, J.; LI, J.; WANG, H.; ZHANG, C.; LI, J. Guided profile generation for controllable dialogue systems. In: *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. [S.l.: s.n.], 2024. p. 400–410.

ZHANG, Y.; WANG, W.; ZHAO, Y.; WANG, Q. Jira issue data from multiple open source projects: A large-scale dataset for software process analytics. In: IEEE. *Proceedings of the 45th International Conference on Software Engineering: Companion Proceedings*. [S.l.], 2023. p. 213–217.

ZHENG, L.; CHIANG, W.-L.; SHENG, Y.; ZHUANG, S.; WU, Z.; ZHUANG, Y.; LIN, Z.; LI, Z.; LI, D.; XING, E. P.; ZHANG, H.; GONZALEZ, J. E.; STOICA, I. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In: *Advances in Neural Information Processing Systems (NeurIPS)*. [S.l.: s.n.], 2023.

ZHENG, Z.; WANG, L.; LIU, X.; ZHU, J.; REN, X. *Generative Job Recommendations with Large Language Model*. 2023. ArXiv preprint arXiv:2307.02157. Disponível em: <https://arxiv.org/abs/2307.02157>.

ZHOU, C.; LIU, P.; XU, P.; IYER, S.; SUN, J.; MAO, Y.; MA, X.; EFRAT, A.; YU, P.; YU, L.; ZHANG, S.; GHOSH, G.; LEWIS, M.; ZETTLEMOYER, L.; LEVY, O. LIMA: Less is more for alignment. In: *Advances in Neural Information Processing Systems (NeurIPS)*. [S.l.: s.n.], 2023.

ZHU, Q.; ZHANG, Z.; FANG, Y.; LI, X.; TAKANOBU, R.; LI, J.; PENG, B.; GAO, J.; ZHU, X.; HUANG, M. Convlab-2: An open-source toolkit for building, evaluating, and diagnosing dialogue systems. In: CELIKYILMAZ, A.; WEN, T.-H. (Ed.). *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Online: Association for Computational Linguistics, 2020. p. 142–149. Disponível em: <https://aclanthology.org/2020.acl-demos.19/>.

Apêndices

APÊNDICE A – Modelos completos dos prompts

A.1 Prompt P1: Dossiê do projeto + plano de execução

Prompt 1 — Base (Dossiê do projeto + plano de execução

```
[SYSTEM]
Você é um gerador de cenários corporativos realistas e auditáveis.
Retorne APENAS JSON válido conforme o schema especificado.

[INPUT]
- run_meta: {run_id, seed, timebox_days, language}
- affinity_band: {HIGH|MID|LOW}
- participants: lista de objetos, cada um com:
  {
    _id: int,
    trajectory_lastK: [string],
    trajectory_summary: string,
    skills_text: string,
    enriched_titles: [string],
    enriched_descriptions: [string]
  }

[TASK]
1) Defina papéis prováveis no time apenas a partir de trajetória/skills.
2) Gere PROJECT_PACK único contendo: projeto, temas[3-6], backlog[5-12],
    artefatos[4-8].
3) Gere BLUEPRINT do canal contendo: threads[3-6], alvo_mensagens_por_thread,
   distribuição_de_fala_por_papel, decisões_obrigatórias[3-8], rituais.

[STYLE RULES POR AFINIDADE]
- HIGH: mais jargão, menos explicação, decisões rápidas.
- MID: equilíbrio.
- LOW: mais grounding e checagem de entendimento.

[CONSTRAINTS]
- Não use nomes próprios.
- Linguagem corporativa, aplicável a múltiplas áreas.
- Saída deve seguir o schema exato:
```

```
[OUTPUT JSON SCHEMA]
{
  "schema_version": "1.0",
  "run_meta": {"run_id": "...", "seed": 0, "timebox_days": 0, "language": "..."},

  "affinity_band": "HIGH|MID|LOW",
  "participants": [{"_id": 0, "role": "...", "role_rationale": "..."}],
  "project_pack": {
    "project": {
      "title": "...",
      "objective": "...",
      "deliverables": ["..."],
      "constraints": ["..."],
      "stakeholders": ["..."],
      "deadline_days": 0,
      "risks": [{"risk": "...", "mitigation": "..."}]
    },
    "themes": [{"theme": "...", "why_relevant": "..."}],
    "backlog": [{"task_id": "T1", "task": "...", "owner_role": "...", "
    definition_of_done": ["..."]}],
    "artifacts": [{"artifact_id": "A1", "type": "doc|sheet|ticket|checklist|
    guideline|incident|other", "title": "...", "used_for": "..."}]
  },
  "blueprint": {
    "channel_name_hint": "...",
    "threads_plan": [{
      "thread_id": "TH1",
      "thread_goal": "...",
      "trigger": "...",
      "target_messages_range": [15, 25],
      "mandatory_decisions": [{"decision_id": "D1", "decision": "..."}],
      "artifacts_to_reference": ["A1","A2"]
    }],
    "talk_distribution_by_role": [{"role": "...", "share": 0.0}],
    "channel_rituals": ["..."]
  }
}

[FINAL OUTPUT]
Retorne somente o JSON.
```

A.2 Prompt P2: Geração do canal por cenário

Prompt 2 — Por cenário (gerar canal + threads)

```
[SYSTEM]
Você é um simulador de conversa corporativa estilo Slack (multi-party, multi-
    thread).
Retorne APENAS JSON válido conforme o schema.

[INPUT]
- run_meta: {run_id, seed, scenario_id, group_id, affinity_band}
- project_pack: (fixo, NÃO MODIFICAR)
- blueprint: (fixo, NÃO MODIFICAR)
- participants: lista com:
  {
    _id: int,
    role: string,
    trajectory_signature: string,
    behavioral_overlay: {
      initiative: {LOW|MED|HIGH},
      clarity: {LOW|MED|HIGH},
      responsiveness: {LOW|MED|HIGH},
      coordination: {LOW|MED|HIGH},
      error_proneness: {LOW|MED|HIGH},
      confidence: {LOW|MED|HIGH}
    }
  }

[TASK]
Gere um canal contendo as threads do blueprint, respeitando decisões e artefatos.


[CONSTRAINTS]
- Proibido mencionar: performance, produtividade, habilidade, scores, low/med/
    high, percentuais.
- Não mudar PROJECT_PACK/BLUEPRINT.
- Garantir vozes distintas.
- Saída no schema:

[OUTPUT JSON SCHEMA]
{
  "schema_version": "1.0",
  "run_meta": {"run_id": "...", "seed": 0, "scenario_id": "...", "group_id":
    "...", "affinity_band": "..."},
  "project_pack_ref": {"project_title": "..."},
  "participants": [{"_id": 0, "role": "..."}],
```

```
  "channel": {
    "channel_id": "...",
    "threads": [{
      "thread_id": "TH1",
      "thread_goal": "...",
      "messages": [{
        "msg_id": "M1",
        "speaker_id": 0,
        "reply_to": null,
        "text": "...",
        "mentions": [0,1],
        "artifact_refs": ["A1","T2"],
        "decision_refs": ["D1"],
        "reactions": [{"emoji": "+1", "count": 2}]
      }],
      "thread_outcome": {
        "decisions_made": ["D1"],
        "next_steps": [{"task_id": "T2", "owner_id": 0, "when": "..."}]
      }
    }]
  }
}

[FINAL OUTPUT]
Retorne somente o JSON.
```

A.3 Prompt P3: Validação e regeneração

Prompt 3 — Validador (rubricas + correção)

```
[SYSTEM]
Você é um verificador de qualidade de dataset.
Retorne APENAS JSON no schema.

[INPUT]
- expected: {min_threads, max_threads, min_msgs_per_thread, max_msgs_per_thread}
- blueprint: (threads_plan com mandatory_decisions e artifacts_to_reference)
- generated_json: (saída do Prompt 2)

[TASK  CHECKLIST]
A) Estrutura: JSON parseável, chaves obrigatórias, contagens nos limites,
    speaker_id válido.
B) Conteúdo invariável: mandatory_decisions presentes; artefatos referenciados
    conforme plano.
```

```
C) Anti-vazamento: não contém termos proibidos nem notas.
D) Rubricas (15): coerência, realismo, diversidade de voz.

[REPAIR POLICY]
- Erros estruturais menores: devolver corrected_json.
- Falha grave: ok=false e regen_instructions.

[OUTPUT JSON SCHEMA]
{
  "ok": true,
  "issues": [{"severity": "LOW|MED|HIGH", "code": "SCHEMA|COUNT|DECISION_MISSING
    |LEAK|INCONSISTENCY|VOICE_SIMILARITY", "detail": "..."}],
  "rubrics": {"coherence_1_5": 0, "realism_1_5": 0, "voice_diversity_1_5": 0},
  "corrected_json": null,
  "regen_instructions": null
}

[FINAL OUTPUT]
Retorne somente o JSON.
```

A.4 Prompt P4: Inferência inversa (estimar atributos a partir do texto)

Prompt 4 — Inferência inversa (atributos por participante), breakable

```
[SYSTEM]
Você é um avaliador que infere sinais de desempenho, produtividade e habilidade
    a partir de conversas corporativas.
Produza APENAS JSON válido conforme o schema.

[INPUT]
- run_meta: (run_id, seed, model_id, scenario_id, group_id)
- participants: lista com (_id, role, trajectory_signature)
- channel: threads e mensagens (msg_id, speaker_id, text, artifact_refs,
    decision_refs)

[TASK]
Para cada participante, estime:
- produtividade_level: LOW|MED|HIGH
- habilidade_level: LOW|MED|HIGH
- desempenho_level: LOW|MED|HIGH
- (opcional) score_0_100 por atributo + confidence_0_1
- evidence: 2-5 evidências com msg_id e trecho curto justificando cada atributo
```

```
[CONSTRAINTS]
- Não mencionar "low/med/high" no texto das evidências; apenas no JSON.
- Não assumir acesso a rótulos de cenário; baseie-se somente no conteúdo do
    diálogo e trajetória.

[OUTPUT JSON SCHEMA]
{
  "ok": true,
  "predictions": [
    {
      "_id": 0,
      "productivity_level": "LOW|MED|HIGH",
      "ability_level": "LOW|MED|HIGH",
      "performance_level": "LOW|MED|HIGH",
      "scores_optional": {
        "productivity_0_100": 0,
        "ability_0_100": 0,
        "performance_0_100": 0,
        "confidence_0_1": 0.0
      },
      "evidence": [
        {"attribute": "productivity|ability|performance", "msg_id": "M12", "
    snippet": "..."}
      ]
    }
  ]
}
```