arXiv:2602.01479v1 [cs.CL] 1 Feb 2026

EBISU：日本金融领域大语言模型基准测试

Xueqing Peng[2], Ruoyu Xiang[3],Fan Zhang[4],Mingzi Song[5],Mingyang Jiang[2], Yan Wang[2],Lingfei Qian[2],Taiki Hara[1],Yuqing Guo[2],Jimin Huang[1,2], 辻井润一[6],Sophia Ananiadou[1,7]

[1]曼彻斯特大学， [2]*The Fin AI*， [3]纽约大学，[4]东京大学， [5]明治学院大学，[6]产业技术综合研究所 *(AIST)*，[7]国家文本挖掘中心

联系邮箱：jimin.huang@postgrad.manchester.ac.uk

摘要

日本金融融合了黏着语、中心语后置的语言结构、混合书写系统以及依赖间接表达和隐含承诺的高语境交际规范，这对大语言模型构成了巨大挑战。我们推出EBISU，这是一个用于评测日语母语金融语言理解能力的基准，包含两项以语言和文化为基础、由专家标注的任务：*JF-ICR*，用于评估面向投资者的问答中的隐含承诺与拒绝识别；以及*JF-TE*，用于评估从专业信息披露中层次抽取并排序嵌套财务术语的能力。我们评估了多种开源和专有大语言模型，涵盖通用模型、适配日语的模型以及金融模型。结果表明，即使是最先进的系统，在两项任务上也表现不佳。尽管扩大模型规模带来的提升有限，针对语言和领域的适配也未能可靠地改善性能，仍有大量差距亟待弥补。EBISU 提供了一个专门的基准，以推动立足语言和文化的金融自然语言处理发展。所有数据集和评测脚本均已公开发布。 [1]

1 引言

即使是最先进的大语言模型，也可能在日本金融领域迷失方向。这在语言学上有其根源，因为日语是黏着语，且句子中心语后置（McCawley, 1975），因此情态和否定往往通过句子



图1：*JF-ICR* 和 *JF-TE*的表示示例。

基准	JA	Fin	Ling	Prag	FC
JGLUE（Kurihara 等人，2022）	✓	✗	✗	✗	✗
JBLiMP（Someya 和 Oseki，2023年）	✓	✗	✓	✗	✗
FinBen（Xie 等人，2024年）	✗	✓	✗	✗	✗
FinQA（Chen 等人，2022）	✗	✓	✗	✗	✗
MultiFinBen（Peng 等人，2025年b）	✓	✓	✗	✗	✓
Japanese-LM-Fin-Harness（Hirano，2024b）	✓	✓	✗	✗	✗
EDINET-Bench（Sugiura 等人，2025年）	✓	✓	✗	✗	✗
惠比须金融（我们提出）	✓	✓	✓	✓	✓

表1：与现有基准的比较。JA：日语覆盖范围；Fin：金融领域覆盖范围；语言学：语言现象；语用学：语用理解；事实一致性：金融文化知识。

1数据：https://huggingface.co/datasets/TheFinAI/JF-ICRhttps://huggingface.co/datasets/TheFinAI/JF-TE代码：https://github.com/The-FinAI/Ebisu

最终（Hasegawa，2018；Narrog，2007）以及拒绝回答，常通过语用线索间接表达（Ide，1989年）。日语金融术语的构词方式进一步增加了解读难度：嵌套的名词复合词和混合文字变体（汉字、平假名和片假名（Bond andBaldwin，2016；Joyce etal.，2017年；Tamaoka，2014年））会使语义范围取决于术语边界和指代。鉴于日本企业披露和投资者沟通已在全球范围内被接收和处理，评估这项能力对跨境金融而非仅限于国内应用十分重要；尤其是2024财年，外国投资者持有日本上市股票市值的32.4%（Tsutsumi，2025年）。

然而，现有的多语言与金融大语言模型评测针对跨语言可比性和可扩展监督进行了优化（Peng 等人，2025年b；Xie 等人，2024年），这使其偏向短上下文、明确目标，以及在不同语言和标注者之间保持稳定的标签体系。因此，多语言基准集中于问答、自然语言推理和多项选择式格式（Qian 等人，2025年a；Zhang 等人，2024年；Peng 等人，2025年a），而金融基准通常被实例化为具有明确输入—输出监督的规范任务模板（Wang 等人，2025年a、b；Qian 等人，2025年b），这使得二者都不太适合那些难以大规模操作化的现象，例如以间接方式传达的话语层面的意图，以及由句末形态判定的立场（Nasukawa 等人，2025年）。日语金融资源（Hirano，2024年a）往往继承同样的限制，因为高质量的母语金融文本难以整理并取得许可证，而语用意图标签的可靠标注成本高昂，因此现有数据集往往侧重情感（Nakatsuka 和 Suimon，2025年）或考试式问答（Sugiura 等人，2025年）。

在本研究中，我们提出EBISU，一个面向原生日语金融语言理解的基准，针对真实企业沟通中两项相互关联的要求：关于承诺与拒绝的立场推断，以及在混合文字文本中对金融术语进行边界敏感型落地（表2）。*JF-ICR*（日语金融隐含承诺识别）由4家公司、跨越3年的94组面向投资者的问答对组成，并标注为5个隐性立场类别，用以在高语境话语下区分同意、保留态度与拒绝。*JF-TE*（日语金融术语抽取）包含202个附注-级实例，提取自10份专业披露文件，包含2,412处专家整理的术语提及，经规范化后覆盖777个独特金融术语；其中嵌套复合词与文字变体外来语使术语边界和语义范围变得模糊。两个数据集均由接受过金融培训的日语母语水平标注员依照详细指南进行标注，且每个条目均经独立双重标注，分歧由资深专家裁定；我们报告标注者间同意（表3），并辅以基于审核的立场一致性与术语边界有效性质量检查。我们评估*JF-ICR*时以准确率作为主要指标，并将*JF-TE* 作为排序检索任务，使用F1值和HitRate@K（K=[1/5/10]），在精确匹配下进行诊断性评估，以检验模型能否在日语金融话语中还原所讨论的内容以及所作出的承诺。

我们在EBISU上评估了22个大语言模型，其中包括开放权重模型和专有系统，结果发现各模型的整体表现仍然较低，凸显了该基准的挑战性。我们在同系列模型内部进行规模匹配比较，以分析模型规模、日语适配和金融领域适配的影响。模型规模扩大 consistently improves performance on both *JF-ICR*(+0.33 准确率) and *JF-TE* (+0.38 F1值)，但这些提升仍然有限。相比之下，语言适配和领域适配并不能稳定带来改进：在规模匹配的情况下，日语适配模型并未始终优于其英语对应模型，而持续进行金融预训练甚至可能导致性能下降，尤其是在金融术语抽取任务上（-0.12 F1值）。在两项任务中，错误主要集中于由语言学和语用学定义的现象，而非领域知识缺失。在*JF-ICR*上，模型难以处理通过分句末尾助动词或间接拒绝策略来确定立场的情况，经常将拒绝与中性或含蓄的回答混淆。在*JF-TE*上，随着复合词深度和混合文字变异性的增加，排序质量会下降；而HitRate@1与HitRate@10之间的差距表明，模型面临的是术语边界解析和变体处理方面的困难，而不是金融词汇不足。总体而言，这些发现表明，当前的规模扩展和适配策略虽能提高平均准确率，却无法稳定捕捉日语金融交流所依赖的形态句法和语用机制。

我们的贡献体现在四个方面：

• **EBISU 基准**。我们推出EBISU，这是一个面向日语母语金融语言理解的基准，旨在日语特有的语言和话语条件下，评估语用立场推断与金融术语落地能力。

•**基于日本金融的新任务和数据集**。 我们定义了两项由专家标注的新任务和数据集，充分反映日本金融实践：*JF-ICR* 用于识别面向投资者的问答中的隐含承诺，*JF-TE* 用于提取并排序专业信息披露中的金融相关术语。

• **全面评估与诊断分析**。 我们对 22 个大语言模型进行了广泛评估，并提供诊断分析，以分离规模扩展和适配的影响。结果表明，性能提升有限，仅凭专有训练并不能保证出色表现。

• **公开资源**。 我们发布 EBISU 及配套的标注指南和评估代码，以促进兼顾语言与文化的金融自然语言处理研究。

2 相关工作

2.1 日语金融大语言模型

日语大语言模型的发展已从通用适配转向深度金融领域专业化。早期的 Swallow 系列（Fujii 等人，2024年）建立了一套成功流程，通过持续预训练和扩展词汇量来提升日语句法能力与推理能力。随后，（Tanabe 等人，2024年；Hirano 和 Imajo，2024年）等研究进一步推动了金融领域对齐，提供来自日本银行和金融厅等权威机构的高质量指令，确保模型回答符合专业监管标准。产业界的创新包括软银的 Sarashina[2]，，其目标是打造拥有 4600 亿参数、面向法律语言理解的“主权人工智能”；以及 Stockmark-LLM[3]，它通过在商业场景中抑制幻觉并通过海量新闻与专利语料库进行训练，来处理商业场景。

[2]https://huggingface.co/sbintuitions/sarashina2.1-1b [3] https://huggingface.co/stockmark/stockmark-100b

2.2 日语金融基准

日语金融大语言模型的评估框架已从基于知识的测试发展到专家级推理任务。Japanese-LM-Fin-Harness（Hirano，2024b）率先整合证券分析师考试、注册会计师审计理论和二元情感分析，为基准测试提供了初步基础。近期，为评估专家级推理能力，EDINET-Bench（Sugiura 等人，2025年）基于十年来的官方申报文件，推出了会计舞弊检测和盈利预测等专家任务。SIG-FIN UFO-2024（Kimura 等人，2024年）等专项评估聚焦于证券报告中的表格检索和问答，而 pfgen-bench（Imajo 等人，2025年）则衡量金融文本生成的质量。其他关于小盘股（Suzuki 等人，2025年）和业绩说明会对话（Nakat-suka 和 Suimon，2025年）的研究进一步表明，专业化大语言模型能够识别传统方法所忽略的预测信号和细微的投资者立场。现有评估涵盖通用日语语言基准（Kurihara 等人，2022；Someya 和 Oseki，2023年）以及全球金融基准（Xie 等人，2024年；Chen 等人，2022），但前者缺少金融领域的覆盖，后者则主要以英语为中心。跨语言研究（Peng 等人，2025年b）在一定程度上弥合了日本金融领域的差距，但通常仍过于粗略，无法进行细粒度语言分析。

3 EBISU基准

本节将介绍EBISU，这是一个旨在评估大语言模型对日语母语金融语言理解能力的基准。EBISU 包含两个互补任务，分别针对：（1）高语境日语金融交流中的日语金融隐含承诺识别（*JF-ICR*）；（2）从专业金融文档中进行日语金融术语抽取（*JF-TE*）（表2）。

3.1 *JF-ICR*：日语金融隐含承诺识别

日语交流的特点是间接性、礼貌程度以及意图的隐含表达，这使得同意或拒绝难以推断。这一困难因日语的黏着语特性而进一步加剧，其中关键的语义

任务 / 数据集	来源	测试规模	指标	许可证	测试能力
JF-ICR	公司问答[1]	94	准确率	公开	日本金融中的隐含意图理解
JF-TE	年度证券报告[2]	202	F1值和HitRate@K[3]	公开	金融术语抽取与排序

[1] https://www.mufg.jp/ir/presentation/2024/index.html; https://www.smfg.co.jp/investor/financial/presentation.html; https://www.itochu.co.jp/ja/ir/financial_statements/2025/index.html; https://www.mufg.jp/ir/stock/meeting/2023/; https://www.mitsubishicorp.com/jp/ja/ir/

[2] https://disclosure2.edinet-fsa.go.jp/

[3] 我们报告 HitRate@1、HitRate@5 和 HitRate@10，以评估不同排序深度下金融术语的准确率。

表 2： EBISU 基准概览。对于每项任务，我们报告数据来源、数据规模、评估指标、许可证信息以及目标能力。

和语用信息往往编码于句末位置，这与英语或汉语形成鲜明对比，从而导致在极性、承诺和风险解读方面频繁出错。在日语金融话语中，由于普遍使用模糊而抽象的表达、显性否定断言有限、模糊与假设性写作风格、层叠否定，以及含义可能偏离其来源语言的片假名外来语，意图推断变得更加困难；这些表达往往以中性或有利的方式呈现潜在风险。

由于大多数大语言模型主要使用以英语为中心的语料库和美式风格的财务披露进行训练，因此在准确推断日语金融交流背后的真实意图时可能会遇到困难。*JF-ICR* 任务旨在评估这一能力，考察模型能否在此类高语境场景中正确识别隐含同意和拒绝。

任务定义。 *JF-ICR* 任务被构建为针对日语金融交流的意图分类问题。给定一个单轮金融问题—回答对（q， a），其中 q 表示金融问询，a 表示相应的日语公司回复，模型需要推断回复 a所表达的潜在意图。输出为五级意图标签 $y \in \{-2, -1, 0, +1, +2\}$，分别对应强承诺 ($+2$)、弱承诺或有条件承诺 ($+1$)、中性或保留态度 (0)、弱拒绝 ($-1$)和强拒绝 ($-2$)。这种构建方式侧重于从回复$a$中推断意图，而非进行情感分类或表层极性检测。

数据来源。 我们根据真实世界中公开可获取的日本企业披露材料构建了*JF-ICR* ， 这些材料记录了企业在高语境场景下回应财务咨询的情况。具体而言，我们收集了以下问答实录：(i) 盈利电话会议——/投资者说明会问答[4],(ii) 股东大会问答[5],以及 (iii) 财务业绩说明会（决算说明会）问答[6]。我们的语料库总计包含来自4家公司的8份来源文档，时间跨度为2023年–2026年，每份文档平均包含10组问答。

为确保意图标注明确无歧义，我们进行了细致的人工筛选。我们仅保留聚焦于单一主题的单轮问答对，并排除包含多个部分的问题、交织的后续提问，或需要更广泛对话背景才能理解的交流。这一筛选流程减少了标注歧义并提升了数据质量。筛选后，我们获得了 94 个单轮日语金融问答实例，供专家标注。

专家标注。我们与日本金融专家合作，制定了严格的标注指南（附录A）。每个实例均采用五级同意／拒绝量表进行标注 $\{+2, +1, 0, -1, -2\}$，并为每个标签制定了详细标准，以确保一致性。我们通过预标注轮次反复完善指南，并特别注意纳入针对模糊情形的规则。只有表达出强烈、明确、清晰且肯定的同意或拒绝的回答才会被赋予 ± 2 标签。所有标注均由两位日语达到母语水平且拥有丰富行业经验的日本金融专家完成（附录C）。标注流程使用 Label Studio 平台开展（图3，附录D），确保工作流程简洁高效且可复现。

质量验证。 为评估 *JF-ICR*标注的可靠性，我们使用多类别分类任务的标准指标来衡量标注者间的一致性，包括宏平均 F1（Sokolova和Lapalme，2009年）、Cohen's κ （Co-

[4] https://www.mufg.jp/ir/presentation/2024/index.html
https://www.smfg.co.jp/investor/financial/presentation.html
https://www.itochu.co.jp/ja/ir/financial_statements/2025/index.html

[5] https://www.mufg.jp/ir/stock/meeting/2023/

[6] https://www.mitsubishicorp.com/jp/ja/ir

数据集	宏平均 F1	Cohen 的 κ	Krippendorff 的 α
JF-ICR	0.9215	0.8769	0.8768
JF-TE	0.8230	0.7786	0.7832

表 3：使用宏平均 F1、Cohen's κ 和 Krippendorff's α 评估的 *JF-ICR*和 *JF-TE*标注者间一致性结果。

hen，1960年）以及 Krippendorff's α （Krippendorff，2011）（附录 E）。宏平均 F1 反映了五个意图类别之间的均衡一致性，而 Cohen's Kappa 和 Krippendorff's Alpha 则针对偶然一致性和标签分布偏斜进行了调整。一致性结果表明，标注者之间具有高度一致性（表3），说明标注指南定义清晰且得到了一致应用。

指令数据转换。为支持基于指令的评估，我们使用专家编写的特定任务提示，将每个已标注实例转换为结构化的指令式格式，如下所述。

评估指标。我们将意图识别建模为一个多分类问题，并使用准确率（Makri-dakis，1993）（附录F）评估模型性能。准确率度量预测意图标签与标注的真实标签完全匹配的实例所占比例，从而直接评估整体分类的正确性。

JF-ICR 的任务指令

你是一位精通日语商务沟通的日本金融专家。给定一个财务问题及相应的公司回复（两者均为日语），你的任务是确定公司的潜在意图级别，并从以下集合中准确输出一个标签：{"+2", "+1", "0", "-1", "-2"}。

标签含义仅用于说明（请勿输出这些文本）："+2"："强承诺" "+1"："弱承诺或有条件承诺" "0"："中性或保留态度" "-1"："弱拒绝" "-2"："强拒绝"

财务问题：{Japanese question} 公司回复：{Japanese answer}

直接输出所选的意图标签，不要提供任何解释。
回答：{Intent label}

3.2 *JF-TE*：日语金融术语抽取

日语金融术语受到三种书写系统、层层嵌套的名词复合结构，以及大量借自汉语和英语的词汇共同影响。金融术语主要以汉字和片假名呈现，其中外来词往往会偏离其来源语言中的语义；而外形相同的词在不同语言中也可能表达截然不同的含义（例如，*kent¯o*"对价"，*bi my¯o* "模糊的"，*tai¯o* "回应"），这限制了跨语言词典方法的有效性（Goworeket al.，2025年；Li et al.，2020）。除语言因素外，日本财务披露遵循的惯例也与美式报告大相径庭：虽然披露名义上面向投资者，但其主要受众是股东，因此形成了独特的写作风格和信息组织方式。年度申报文件高度标准化，重要更新往往嵌在看似不起眼的附注部分，而非醒目的数字中，容易被语言模型忽略。因此，*JF-TE* 任务旨在从日本披露附注中抽取金融术语，评估主要在美国金融文件上训练的大语言模型，能否在日本金融语境中识别领域专用术语。

任务定义 该*JF-TE* 任务被定义为针对日语金融文本的金融术语提取与排序问题，并明确对嵌套术语结构进行建模。受 C-value 等以术语性为导向的原则启发（Frantzi 等人，2000年），该任务评估模型能否在日语披露文件中识别并优先排列特定领域的金融术语。每条输入均为一段取自专业日语财务文件的附注 n。模型需要识别一组最大金融术语 $M = \{m_1, m_2, \ldots, m_p\}$，其中每个 m_i 表示一个代表金融概念的最长连续表达 n。对于每个最大术语 m_i，模型还需进一步识别一组候选嵌套术语 $T_i = \{t_{i1}, t_{i2}, \ldots, t_{ik}\}$，其中每个 $t_{ij} \subseteq m_i$ 都是潜在的金融术语（包括 $t_{ij} = m_i$）；如果不存在嵌套术语，则 $T_i = \{m_i\}$。模型输出一个有序列表 T_i'，按照这些候选项代表特定领域金融术语的可能性从高到低进行排序。非金融表达

将从输出中排除。最终输出为一个结构化 JSON 对象，汇总所有最大术语中的 T_i'。

数据来源。 为构建*JF-TE* 数据集，我们收集了通过 EDINET 披露的日本年度证券报告（*Yūkashōken hōkokusho*）。EDINET 是[7],由日本金融厅运营的官方电子披露系统。我们的语料库包含 10 份于 2025年由 10 家上市公司发布的专业信息披露文件，每份报告在 EDINET 原始文件中最长达 261 页。

按照日本财务披露惯例，重要更新和领域专属信息往往包含在细分的附注中，而非正文叙述部分。为捕捉这些信息密集的内容单元，我们手动识别并提取每份财务披露文件中的单条附注，并将每条附注视为独立实例，用于专家标注和金融术语提取。这一流程共生成 202 条经过整理的附注级数据实例。

专家标注。 我们遵循与*JF-ICR*类似的专家标注流程，并根据层级金融术语制定了针对具体任务的指南（附录 B）。每条附注级实例均通过识别最长的金融术语跨度，以及其中包含的嵌套金融术语（如适用）进行标注。标注以跨度级别进行，因此可以明确标记多词表达、复合名词和嵌套结构。标注指南经过多轮预标注迭代完善，并制定了明确规则，以规范术语边界判定以及金融术语与一般表达之间的区分。所有标注均由同一批日语母语水平的日本金融专家使用 Label Studio 平台完成（附录 C）（图 4，附录 D），从而确保一致性和可复现性。

质量验证。 我们使用标注者间一致性指标 Macro-F1值（Sokolova 和 Lapalme，2009）、Cohen 的 κ（Cohen，1960年）和 Krippendorff 的 α（Krippendorff，2011）（附录E）来评估标注质量。结果表明，在识别财务术语跨度方面具有较高的一致性（表3）。这种一致性表明，规范术语边界和金融术语性的标注规则具有可操作性和可复现性，从而支持所得高质量 JF-TE 数据集的可靠性。

指令数据转换。 为支持基于指令的评估，我们将每个已标注实例重新表述为指令式格式，并针对金融术语提取任务设计专属提示，如下所示。

评估指标。 该*JF-TE* 任务采用两级指标设计进行评估，以反映日本披露文件中金融术语的层级结构。我们首先通过精确匹配预测结果与真实标注，评估最大范围财务术语提取，并报告 F1 分数（Tjong Kim Sang 和 De Meulder，2003）（附录 F）。随后，我们使用 HitRate@K 评估每个最大范围财务术语中的嵌套金融术语排名（Bor- des 等人，2013）（附录 F），衡量黄金嵌套术语是否出现在前K个预测结果中。为考虑不同实例中嵌套术语数量的差异，我们报告 HitRate@1、HitRate@5 和 HitRate@10。综合来看，这些指标既能衡量完整财务术语边界识别的准确性，也能评估嵌套金融术语排序的有效性。

JF-TE 的英语翻译任务说明

你是一位专门从事财务披露分析的日语金融专家。请仔细阅读以下日语财务附注。

识别文本中所有最长的金融术语。对于每个最长术语，识别其中包含的所有嵌套金融术语；如适用，也包括该最长术语本身。

对于每个最长金融术语，按与领域专用金融术语相关的可能性从高到低排列其相关金融术语（包括该术语本身及所有嵌套术语）。不要包含非金融表达。

请以 JSON 格式返回结果，其中每个元素对应一个最长金融术语，并包含一个按相关性排序的金融术语列表。

文本：{Japanese financial disclosure note} 回答：{JSON with ranked financial terms}

4 实验结果

4.1 评估模型

我们对EBISU 基准中的 22 个大语言模型进行了全面评估，涵盖多种模型类型

[7] https://disclosure2.edinet-fsa.go.jp/

和训练范式。评估的模型包括：(1) 11 个不同规模的开源通用模型（Llama-4-Scout-17B（Meta AI，2025年）、Llama-3.3-70B-Instruct（Dubey 等人，2024年）、Qwen3-235B-A22B-Instruct-2507-FP8、Qwen3-32B、Qwen3-14B、Qwen3-8B（Qwen 团队，2025年）、Qwen2.5-32B-Instruct（Qwen 团队，2024年）、Qwen-14B（Qwen 团队，2023年）、Ministral-3-14B-Instruct-2512（Mistral AI，2025年）、DeepSeek-V3.1（DeepSeek-AI，2024年）、Kimi-K2-Instruct（Kimi 团队，2025年））；(2) 通过官方 API 调用的 4 个闭源专有模型（GPT-5（OpenAI，2025年）、gpt- 4o（OpenAI，2024年）、gemini-3-flash（Google DeepMind，2025年）、claude-sonnet-4-5（Anthropic，2025年））；(3) 1 个英语金融领域专用模型（FinMA-7B（Xie 等人，2023年）；(4) 5 个日语通用模型（Llama-3.3-Swallow-70B-Instruct-v0.4（Swallow Team，2025年）、TinySwallow-1.5B-Instruct（Shing 等人，2025年）、japanese-stablelm-instruct-beta-70b、japanese-stablelm-instruct-beta-7b（Stability AI，2023年）、nekomata-14b（Sawada 等人，2024年））以及 (5) 1 个日本金融领域专用模型（nekomata-14b-pfn-qfin（Hirano 和 Imajo，2024年））。

模型	*JF-ICR*	*JF-TE*				Mean
	Acc	F1	HR@1	HR@5	HR@10	Avg
开源通用模型						
Llama-4-Scout-17B	**0.6064**	0.4133	0.1085	0.3094	0.4513	0.3778
Llama-3.3-70B-Instruct	0.5638	0.4370	**0.1277**	**0.3657**	**0.5111**	**0.4011**
Qwen3-235B-A22B-Instruct	0.4574	0.4355	0.0971	0.3053	0.4301	0.3451
Qwen3-32B	0.1915	0.2860	0.0285	0.1052	0.1763	0.1575
Qwen3-14B	0.1809	0.2038	0.0165	0.0787	0.1097	0.1179
Qwen3-8B	0.1277	0.0548	0.0050	0.0142	0.0233	0.0450
Qwen2.5-32B-Instruct	0.3830	0.4137	0.0982	0.3462	0.4786	0.3439
Qwen-14B	0.0638	0.1613	0.0308	0.0650	0.1076	0.0857
Ministral-3-14B-Instruct	0.3723	0.3756	0.0875	0.2377	0.3704	0.2887
DeepSeek-V3.1	0.5106	0.3653	0.0794	0.2663	0.4121	0.3267
Kimi-K2-Instruct	0.4574	**0.4396**	0.1048	0.3251	0.4680	0.3590
专有模型						
GPT-5	0.5532	0.0000	0.0000	0.0000	0.0000	0.1106
gpt-4o	0.0319	0.3997	0.0890	0.2845	0.4255	0.2461
gemini-3-flash	0.3936	0.0000	0.0000	0.0000	0.0000	0.0787
claude-sonnet-4-5	0.5213	0.4362	0.0743	0.3022	0.4752	0.3618
英语金融模型						
FinMA-7B	0.1596	0.0000	0.0000	0.0000	0.0000	0.0319
日语通用模型						
Llama-3.3-Swallow-70B-Instruct	0.5957	0.4070	0.1258	0.3270	0.4701	0.3851
TinySwallow-1.5B-Instruct	0.0638	0.1937	0.0264	0.0712	0.1012	0.0913
japanese-stablelm-instruct-beta-7b	0.0426	0.0143	0.0069	0.0079	0.0079	0.0159
japanese-stablelm-instruct-beta-70b	0.0319	0.1826	0.0388	0.1090	0.1461	0.1017
nekomata-14b	0.0213	0.0059	0.0173	0.0330	0.0479	0.0251
日语金融模型						
nekomata-14b-pfn-qfin	0.0213	0.0468	0.0012	0.0118	0.0181	0.0198

表 4：大语言模型在EBISU 基准上的表现。*JF-ICR* 使用准确率（Acc）进行评估。*JF-TE* 使用 F1 值和 HitRate@K（HR@1/5/10）进行评估。粗体数值表示各列中的最高分，下划线数值表示各列中的次高分。

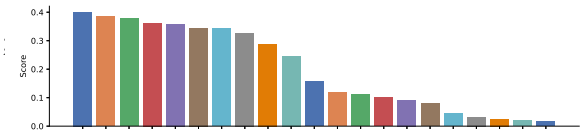

图 2：EBISU基准上各模型的排名表现。

4.2 实施细节

为确保评估完整性和一致性，我们基于 LM Evaluation Harness（Gao 等人，2024年）开发了一套统一的评估流程。专有模型通过其官方 API 进行评估，并将温度固定为 0，以确保输出具有确定性。通过 TogetherAI 提供的开源模型在 TogetherAI 平台上进行部署和评估，其他所有开源模型则使用 vLLM（Kwon 等人，2023年）在图形处理器上进行本地部署和评估。内部评估在一个由四块 H100 图形处理器组成的集群上进行，每块显存为 80GB，总计使用超过 200 GPU 小时。我们将最大生成长度统一设为 1,024。

4.3 主要结果

表4 和图 2 总结了 EBISU 基准的模型表现。我们的分析围绕以下研究问题展开。

RQ1：在日本金融环境下，EBISU 对最先进的大语言模型而言是否具有挑战性？总体而言，结果表明当前的大语言模型面临在EBISU 基准上仍面临重大挑战，包括最先进的系统。领先的专有模型，如 GPT-4o（0.2461）和 Claude-Sonnet-4.5（0.3618），表现有限，而开源模型 Llama-3.3-70B-Instruct（0.4011）取得了总体最佳成绩，这表明要稳健地建模日语金融文本的语言和语用特征，仍然存在持续的困难。模型规模增大通常会带来性能提升。例如，在 Qwen系列中，较大的模型，如 Qwen3-235B-A22B-Instruct（0.3451）和 Qwen3-32B（0.1575），表现优于较小的 Qwen3-14B（0.1179）和 Qwen3-8B（0.0450）。然而，这些提升仍然有限，表明仅凭模型容量无法解决EBISU 所揭示的挑战。相反，持续存在的性能差距表明，现有训练方法未能充分捕捉日语金融交流的语言和文化特性。

范式。总的来说，这些结果凸显了日语金融语言理解的难度，并使EBISU 成为评估现代大语言模型此类能力的高难度基准。

RQ2：模型在日本金融领域跨越语言和领域边界进行知识迁移的效果如何？ 不同模型系列在跨语言和跨领域知识迁移方面存在明显局限。英语金融领域专用模型（*FinMA-7B*：*0.0319*）的表现不如同等规模的通用模型，例如 Qwen3-8B（0.0450）。这一现象表明，主要从英语语料库中学习的金融知识，在缺乏充分日语训练的情况下，无法轻易迁移至日语金融任务。日语通用模型的表现仍然不及其英语对应模型。 例如，Llama-3.3-Swallow-70B-Instruct（0.3851）的表现落后于 Llama-3.3-70B-Instruct（0.4011），这表明语言适配带来的性能提升并不一定具有普适性。与预期相反，持续进行语言和领域专用预训练并不能始终改善性能。日语金融模型 nekomata-14b-pfn-qfin（0.0198）的表现不如其日语通用版本 nekomata-14b（0.0251）以及原始骨干模型 Qwen-14B（0.0857）。这表明，仅在日语金融文本上持续进行下一词元预训练并不足够，甚至可能妨碍模型在需要理解日本金融领域隐含意图的任务上的表现。总体而言，这些发现揭示了现有英语金融模型、日语通用模型和日语金融模型的关键不足，并为推动模型更好地捕捉日本金融领域的语言和文化语境提供了有意义的基准。

RQ3：大语言模型在识别日语金融回复中的隐含承诺时，是否表现出系统性偏差？ 在所有受评估模型中，*JF-ICR* 的表现均不尽如人意，凸显出模型在日语金融语境中的意图理解方面持续面临困难。 即使是表现最强的模型 Llama-4-Scout-17B（准确率：0.6064），在准确识别日本金融中的隐含同意与拒绝时仍然举步维艰，这表明该能力仍是一项尚待攻克的挑战。除整体准确率外，我们还进一步考察了各模型预测的平均承诺评分。在模型规模相近的情况下，英语中心型模型始终给出高于日语模型的承诺评分，例如 Llama-3.3-70B-Instruct（0.9787）高于 Llama-3.3-Swallow-70B-Instruct （0.8404），Qwen-14B（0.3830）高于 nekomata-14b（-0.8511）。这一模式表明，主要使用英语数据训练的模型存在系统性偏差：它们经常将日语金融回复中的间接拒绝和策略性不承诺误解为接受或弱同意。这些发现体现了*JF-ICR*本身的难度；其难点源于日语语言特征与金融沟通规范的共同影响：前者通过句末表达、层叠否定等方式隐含地传达意图，后者则不鼓励明确拒绝。因此，在高语境日语金融话语中，*JF-ICR* 要求模型超越表层语义，区分真正的承诺与策略性不承诺。

RQ4：由于语言特有的词汇属性，日语金融术语对于大语言模型来说是否特别难以抽取？ 在所有评估模型中，*JF-TE* 的表现仍然较弱，揭示出当前大语言模型在抽取和优先排序日语金融术语方面存在显著局限。即使是表现最强的模型 Llama-3.3-70B-Instruct，在不同截断值下也只能取得一般的命中率（HitRate@1：0.1277，HitRate@5：0.3657，HitRate@10：0.5111），这表明，在日本披露文件中准确识别并排序金融术语，仍远未得到解决。*JF-TE* 强调，日语金融术语抽取并不仅仅是表层匹配问题，而是在一个由含义发生偏移的外来词以及受日本金融披露惯例塑造、金融表达高度密集的环境中，识别并排序与领域相关的术语。

5 结论

我们推出EBISU，这是一个用于评估大语言模型理解日语金融语言能力的基准，旨在明确反映日本金融的语言结构和交流规范。该基准由*JF-ICR* 和*JF-TE*组成，EBISU 针对日语使用方式与日本金融实践相互作用所带来的两项核心挑战：隐含承诺识别和金融术语抽取。实验

对不同模型系列的评估表明，当前的大语言模型（包括最先进的专有系统和开源系统）在这两项任务上都表现吃力。尽管扩大模型规模能带来一定改善，但针对语言和领域的训练收益有限，甚至可能降低性能。这些结果揭示了日本日语金融交流的语言与文化特征所导致的持续性差距，而当前模型训练尚未解决这些问题。通过揭示承诺识别和术语排序方面的系统性难点，EBISU 提供了一个专门的基准，助力大语言模型更深入、更可靠地理解日本金融。

局限性

我们需要承认，EBISU 存在若干局限。首先，受实际因素制约，数据集的范围有限：*JF-ICR*包含来自4家公司的问答对话记录，时间跨度有限（2023年–2026年）；而*JF-TE* 则仅关注来自EDINET的年度证券报告。这种覆盖范围可能无法充分体现不同规模公司、不同行业或不同披露格式下日语金融交流的多样性。其次，鉴于当前数据集规模有限，本次发布未提供训练/测试集划分。待数据集扩充后，我们将引入标准的训练/验证/测试集划分，以支持模型训练和评估。第三，EBISU 评估了两项互为补充的任务，旨在应对日语金融语言理解的核心挑战，但并未涵盖金融自然语言处理能力的全部范围。这两项任务均针对相对较短的文本片段，因此对长上下文理解和多轮话语推理的评估能力有限。最后，评估指标本身也存在局限：*JF-ICR* 的准确率将所有误分类一视同仁，而*JF-TE*的精确匹配则可能会惩罚语义正确但略有偏差的文本片段。

伦理方面的考量

发布EBISU 引发了若干值得认真关注的伦理问题。首要担忧是，评估日语金融大语言模型的能力时，可能会过度依赖基准分数。基准表现未必能充分预测模型在生产系统中的实际效用，因为延迟、应对分布偏移的稳健性以及与现有工作流程的集成等因素同样至关重要。EBISU 范围较为集中，因此可能存在数据集偏差和代表性方面的局限。由于数据集取自数量有限的大型上市公司，可能无法反映日本金融领域更广泛范围内的沟通风格、术语或披露惯例。其时间范围（2023年–2026年）也可能无法涵盖不断演变的金融术语或监管变化。此外，与许多基准一样，模型可能会被优化为在EBISU上表现出色，却未能相应提升泛化能力，从而可能助长针对特定基准的过拟合。

EBISU 旨在通过增进对大语言模型日语金融语言处理能力的理解，服务于研究与评估目的。所有数据集均由公开可用的企业披露文件和财务文件构建，我们尊重原始数据来源的使用条款和许可要求。该基准旨在用于学术研究、模型评估和方法开发。在未经广泛的额外验证、人工监督和领域专家审查的情况下，即使模型在EBISU 上取得高性能，也不应认为其适合在金融系统中进行生产环境部署、自动化投资分析、监管合规或其他高风险应用。EBISU 所评估的任务聚焦于隐含承诺识别和金融术语抽取，这些任务体现了日语金融语言理解的特定方面，但并不能保证准确的金融推理、风险评估或监管解释。我们鼓励负责任地使用该基准，并欢迎研究界提出反馈，以改进其覆盖范围、公平性和实用性。

参考文献

Anthropic. 2025年。 Claude sonnet 4.5 系统卡。技术报告，Anthropic。访问日期：2026-01-03。

Francis Bond 和 Timothy Baldwin。2016。日语计算语言学导论。Antoine Bordes、Nicolas Usunier、Alberto Garcia- Duran、Jason Weston 和 Oksana Yakhnenko。2013。 用于多关系数据建模的嵌入转换。载于 第*26*届神经信息处理系统进展，第2787–2795页。

Zhiyu Chen、Wenhu Chen、Charese Smiley、Sameena Shah、Iana Borova、Dylan Langdon、Reema Moussa、Matt Beane、Ting-Hao Huang、Bryan Routledge 和 William Yang Wang。2022年。 FinQA：一个用于金融数据数值推理的数据集。 预印本，arXiv:2109.00122。

Jacob Cohen. 1960年。 名义量表的一致性系数。《教育与心理测量》，20(1):37–46。

DeepSeek-AI. 2024年。 Deepseek-v3 技术报告。预印本，arXiv:2412.19437。

Abhimanyu Dubey、Abhinav Jauhri、Abhinav Pandey、AbhishekKadian、AhmadAl-Dahle、AieshaLetman、Akhil Mathur、Alan Schelten、Amy Yang、Angela Fan 等1人。2024年。 Llama 3 模型系列。*arXiv* 预印本 *arXiv:2407.21783*。

Katerina T. Frantzi、Sophia Ananiadou 和 Hideki Mima。2000年。 多词术语的自动识别：C-value/NC-value 方法。 《国际数字图书馆期刊》，3(2):115–130。

Kazuki Fujii、Taishi Nakamura、Mengsay Loem、Hi- roki Iida、Masanari Ohi、Kakeru Hattori、Shota Hi- rai、Sakae Mizuki、Rio Yokota 和 Naoaki Okazaki。2024年。用于跨语言大语言模型适配的持续预训练：增强日语能力。*arXiv* 预印本 *arXiv:2404.17790*。Leo Gao、Jonathan Tow、Baber Abbasi、Stella Biderman、Sid Black、Anthony DiPofi、Charles Foster、Laurence Golding、Jeffrey Hsu、Alain Le Noac'h、Haonan Li、Kyle McDonell、Niklas Muennighoff、Chris Ociepa、Jason Phang、Laria Reynolds、Hailey Schoelkopf、Aviya Skowron、Lintang Sutawika 等另外 5 人。2024年。少样本语言模型评测框架。Google DeepMind。2025年。Gemini 3 flash：模型卡。技术报告，Google DeepMind。访问日期：2026-01-03。Roksana Goworek、Olivia Macmillan-Scott 和 Eda B Özyi˘git。2025年。弥合语言鸿沟：多语言大语言模型跨语言信息检索的进展。*arXiv* 预印本 *arXiv:2510.00908*。Yoko Hasegawa。2018年。《剑桥日语语言学手册》。剑桥大学出版社。Masanori Hirano。2024年a。用于大语言模型的日语金融基准构建。*arXiv* 预印本 *arXiv:2403.15062*。Masanori Hirano。2024b。Japanese-LM-Fin-Harness：面向日语大语言模型的金融评测框架。*arXiv* 预印本 *arXiv:2403.15062*。Masanori Hirano 和 Kentaro Imajo。2024年。通过持续预训练构建面向金融领域的专用日语大语言模型。*arXiv* 预印本 *arXiv:2404.10555*。

井出祥子。1989年。正式形式与辨别：语言礼貌普遍性中被忽视的两个方面。

Kentaro Imajo、Masanori Hirano、Shuji Suzuki 和 Hiroaki Mikami。2025 年。pfgen-bench：用于评估日语预训练模型文本生成性能的基准。

Terry Joyce、Bor Hodoš˘cek 和 Hisashi Masuda。2017年。构建日语词汇属性本体和数据库：处理日语书写系统的正字法复杂性。《书面语言与读写能力》，20(1):27–51。

Kimi 团队。2025年。Kimi K2：开放式智能体智能。*arXiv* 预印本 *arXiv:2507.20534*。

Yasutomo Kimura、Eisaku Sato、Kazuma Kadowaki 和 Hokuto Ototake。2024年。理解财务文件中的表格。

Klaus Krippendorff。2011年。计算 Krippendorff α 信度系数。技术报告，宾夕法尼亚大学安嫩伯格传播学院。文献更新于 2013-09-13。访问于 2026-01-04。

Kentaro Kurihara、Daisuke Kawahara和 Tomohide Shibata。2022。 JGLUE：日语通用语言理解评测。载于第十三届语言资源与评测会议论文集，第2957–2966页，法国马赛。欧洲语言资源协会。Woosuk Kwon、Zhuohan Li、Siyuan Zhuang、Ying Sheng、Lianmin Zheng、Cody Hao Yu、Joseph E. Gonzalez、Hao Zhang和Ion Stoica。2023年。基于PagedAttention的大语言模型服务高效内存管理。载于*ACM SIGOPS*第*29*届操作系统原理研讨会论文集。Yanyang Li、Yingfeng Luo、Ye Lin、Quan Du、Huizhen Wang、Shujian Huang、Tong Xiao和Jingbo Zhu。2020年。一种面向稳健无监督双语词典归纳的简单有效方法。 *arXiv*预印本 *arXiv:2011.14874*。Spyros Makridakis。1993年。准确度度量：理论与实践问题。 国际预测期刊，9(4):527–529。Noriko A McCawley。1975年。久野暲所著《日语的结构》。 *Journal ofJapaneseLinguistics*，4(1-2):209–232。Meta AI。2025年。Llama 4族群：原生多模态人工智能创新新时代的开端。https://ai.meta.com/blog/llama-4-multimodal-intelligence/。访问日期：2025-05-09。Mistral AI。2025年。推出 Mistral 3。 https://mistral.ai/news/mistral-3。访问日期：2026-01-03。

Hiromasa Nakatsuka和Yoshiyuki Suimon。2025年。财务业绩说明会对话中的信息抽取与情感分析。*Digital Finance*，7(4):605–621。

Heiko Narrog。2007年。日语中的情态与语法化。*Journal ofHistorical Pragmatics*，8(2):269–294。

Kuniya Nasukawa、Ge Song和Sachiko Kiyama。2025年。日语中具有共情功能的句末助词ne：其音系表征研究。*Journal ofJapanese Linguistics*，41(1):95–113。

OpenAI。2024年。GPT-4o 系统卡。技术报告，OpenAI。访问于 2026-01-03。

OpenAI。2025年。GPT-5 系统卡。技术报告，OpenAI。访问于 2026-01-03。

Xueqing Peng、Triantafillos Papadopoulos、Efstathia Soufleri、Polydoros Giannouris、Ruoyu Xiang、Yan Wang、Lingfei Qian、Jimin Huang、Qianqian Xie 和 Sophia Ananiadou。2025年a。Plutus：低资源希腊语金融领域大语言模型基准测试。*arXiv* 预印本 *arXiv:2502.18772*。

Xueqing Peng、Lingfei Qian、Yan Wang、Ruoyu Xiang、Yueru He、Yang Ren、Mingyang Jiang、VincentJim Zhang、Yuqing Guo、Jeff Zhao、Huan He、Yi Han、Yun Feng、Yuechen Jiang、Yupeng Cao、Haohang Li、Yangyang Yu、Xiaoyu Wang、Penglei Gao 和其他 28 位作者。2025年b。MultiFinBen：面向多语言多模态金融应用的大语言模型基准测试。预印本，arXiv:2506.14028。

Lingfei Qian、Xueqing Peng、Yan Wang、Vincent Jim Zhang、Huan He、Hanley Smith、Yi Han、Yueru He、Haohang Li、Yupeng Cao及其他1人。2025年a。当智能体进行交易：面向大语言模型智能体的实盘多市场交易基准。*arXiv*预印本 *arXiv:2510.11695*。

Lingfei Qian、Weipeng Zhou、Yan Wang、Xueqing Peng、Jimin Huang和Qianqian Xie。2025年b。Fino1：论推理增强大语言模型向金融领域的可迁移性。*arXiv*电子预印本，页码arXiv–2502。

Qwen 团队。2023年。Qwen 技术报告。*arXiv* 预印本 *arXiv:2309.16609*。Qwen 团队。2024年。Qwen2.5 技术报告。*arXiv* 预印本 *arXiv:2412.15115*。Qwen 团队。2025年。Qwen3 技术报告。*arXiv* 预印本 *arXiv:2505.09388*。Kei Sawada、Tianyu Zhao、Makoto Shing、Kentaro Mitsui、Akio Kaga、Yukiya Hono、Toshiaki Wakatsuki 和 Koh Mitsuda。2024年。日语预训练模型发布。Makoto Shing、Kou Misaki、Han Bao、Sho Yokoi 和 Takuya Akiba。2025年。TaID：用于语言模型高效知识迁移的时序自适应插值蒸馏。预印本，arXiv:2501.16937。

Marina Sokolova 和 Guy Lapalme。2009年。对分类任务性能度量的系统性分析。*Information Processing& Management*，45(4):427–437。Taiga Someya 和 Yohei Oseki。2023年。JBLiMP：日语语言最小对基准。载于计算语言学协会论文集：*EACL 2023* 研究成果，第1581–1594页，克罗地亚杜布罗夫尼克。计算语言学协会。Stability AI。2023年。《推出日语 Stable LM Beta》。
h
t
t
p
s://
stability.ai/news/japanese-stable-lm-beta-language-models。访问日期：2026年1月4日。

Hirohisa Sugiura、Yusuke Kawamura和Hiroshi Hamaguchi。2025年。EDINET-Bench：日本金融领域大语言模型基准评测。*arXiv*预印本 *arXiv:2506.08762*。

Masahiro Suzuki、Yasushi Ishikawa、Masayuki Teraguchi 和 Hiroki Sakaji。2025年。小盘股中的情感研究：利用语言模型分析日本股票情绪。*International Journal of Informa-tion ManagementData Insights*，5(1)：100318。

Swallow Team。2025年。Llama-3.3-swallow-70b-instruct-v0.4。模型卡。访问日期：2026年1月4日。

Katsuo Tamaoka。2014年。日语书写系统与词汇理解。*JapaneseLanguageandLiterature*，第431–471页。

Kota Tanabe、Masahiro Suzuki、Hiroki Sakaji 和 Itsuki Noda。2024年。Jafin：日语金融指令数据集。*arXiv* 预印本 *arXiv:2404.09260*。

Erik F. Tjong Kim Sang 和 Fien De Meulder。2003年。CoNLL-2003 共享任务简介：与语言无关的命名实体识别。载于*HLT-NAACL 2003* 第七届自然语言学习会议论文集，第142–147页。

Kentaro Tsutsumi。2025年。日本个人持股量创历史最大增幅。彭博新闻。外国股东在2024财年持有的日本股票总市值占比创历史新高，达到32.4%。Yan Wang、Yang Ren、Lingfei Qian、Xueqing Peng、Keyi Wang、Yi Han、Dongji Feng、Xiao-Yang Liu、Jimin Huang 和 Qianqian Xie。2025年a。Fintagging：一个面向大语言模型的基准，用于提取和结构化金融信息。*arXiv* 预印本 *arXiv:2505.20650*。Yan Wang、Keyi Wang、Shanshan Yang、Jaisal Patel、Jeff Zhao、Fengran Mo、Xueqing Peng、Lingfei Qian、Jimin Huang、Guojun Xiong 及另外1位作者。2025年b。Finauditing：用于评估大语言模型的金融分类法结构化多文档基准。*arXiv* 预印本 *arXiv:2510.08886*。

Qianqian Xie、Weiguang Han、Zhengyu Chen、Ruoyu Xiang、Xiao Zhang、Yueru He、Mengxi Xiao、Dong Li、Yongfu Dai、Duanyu Feng、Yijing Xu、Haoqiang Kang、Ziyan Kuang、Chenhan Yuan、Kailai Yang、Zheheng Luo、Tianlin Zhang、Zhiwei Liu、Guojun Xiong 及另外15位作者。2024年。FinBen：面向大语言模型的全面金融基准。载于神经信息处理系统进展，第37卷，第95716–95743页。Curran Associates, Inc.出版。

Qianqian Xie、Weiguang Han、Xiao Zhang、Yanzhao Lai、Min Peng、Alejandro Lopez-Lira 和 Jimin Huang。2023年。Pixiu：面向金融的大语言模型、指令数据和评估基准。预印本，arXiv:2306.05443。

Xiao Zhang、Ruoyu Xiang、Chenhan Yuan、Duanyu Feng、Weiguang Han、Alejandro Lopez-Lira、Xiao- Yang Liu、Meikang Qiu、Sophia Ananiadou、Min Peng 等人。2024年。Dólares 还是 dollars？揭示西班牙语与英语之间金融大语言模型的双语能力。载于第*30*届 *ACM SIGKDD*知识发现与数据挖掘会议论文集，第 6236–6246 页。

一份*JF-ICR*标注指南

标注理念。 *JF-ICR* 关注管理层回应所表达的承诺程度，而非问题表面上的意图或内容的事实正确性。标注仅依据回答是否传达了前瞻性行动、决策或政策立场，以及如何传达这些内容，并反映财务披露中的常见沟通惯例。

由于日本财务披露谨慎且受惯例驱动的特点，明确的强承诺相对少见。因此，大量回答自然归入弱承诺或中性类别，反映了现实中的管理层沟通实践，而非标注歧义或不一致性。

A.1 分类类别

每条回答均根据说话者对回答中所表达的未来行动、决策或政策立场的承诺程度进行标注。

- *+2*：强承诺——对具体行动或结果作出明确、清晰且毫无歧义的承诺。
- *+1*：弱承诺或有条件的承诺——通过条件、保留态度或有限范围表达的意图或方向性立场。
- *0*：中立或带有保留态度的意图——澄清、解释、评估或感知，不包含承诺或拒绝。
- *-1*：弱拒绝——对拟议行动的隐含或有条件的拒绝。
- *-2*：强拒绝——明确且坚决的拒绝，不留任何回旋余地。

A.2 一般标注规则

- 标注关注的是**意图和承诺强度**，而非事实正确性。
- 只有表达**未来行动、决策、政策、目标或战略方向**的陈述才会被考虑赋予非零标签。
- 标注依据**回答的内容**，而非问题的表面形式或意图。即使问题是回顾性的或解释性的，回答也可能提出前瞻性承诺或拒绝。

• 如果回答未表达任何面向未来的行动、决策或政策立场，则应将其标注为*0*，无论问题是什么。

• 模糊表达（例如不确定性、对条件的依赖、持续审查）会降低承诺强度。

• 当出现多个信号时，标签体现回答中表达的**最强适用承诺或拒绝**。

• 标注基于**完整回答的整体立场**，而非孤立的句子或单个表达；标签反映说话者通过回答整体所传达的主导承诺或拒绝。

• 标注关注的是**承诺强度**，而非回答的语气、长度或礼貌程度。

• 日语金融语境中的惯用礼貌表达（例如「努めてまいります」）本身并不意味着承诺。

A.3 特定标注规则

• *+2*（强承诺）：当回答包含表明坚定承诺的明确、果断陈述时适用。典型的语言信号包括：

– 「将实施～」「将执行～」「将实现～」– 「已经决定～」「～已经确定」–「没有改变方针的打算」（直接回答决策或政策问题时）

代表性示例包括：

– 「下期将实施增配股息。」– 「将在中长期实现12%的ROE。」– 「我们坚信，通过这项举措能够实现利润增长。」

• *+1*（弱承诺或有条件承诺）：适用于回答表达了对未来行动或结果的肯定承诺或倾向性承诺，但附带限定条件、谨慎态度或具体性有限的情况。这类回答表明的是倾向性意图或可能性，而非最终决策。典型的语言信号包括：

– 「～していきたい」「～を目指しています」– 「我认为……」「我预期……」– 条件或假设表达（例如「如果……」「取决于……」）

代表性示例包括：

– 「我们计划推进增长投资。」– 「我们认为今后收益也将持续扩大。」–「如果条件成熟，我们将推进讨论。」

• *0 (*中性或保留态度*)*：当回答确实存在歧义，或提供澄清、解释或背景信息，但未表达对任何未来行动、决策或政策立场的承诺或拒绝时，适用此标签。此类回答可能描述事实、背景、看法或推理，但不会传达对未来行为的倾向性或评价性立场。典型的语言信号包括：

– 「无法断定……」「无法给出明确的展望」– 「正在考虑」「有必要观察形势」 – 仅提供事实或背景的描述性或解释性陈述

代表性示例包括：

–「目前无法给出明确的展望。」– 「存在各种看法，因此不予置评。」– 「过去我们开展过这样的举措。」（仅作背景说明）

• *-1*（弱拒绝）：适用于回答对拟议的未来行动或变更表达否定立场，但这种表达带有限定性、条件性或时间限制，为未来重新考虑留下余地的情况。此类回答拒绝在当前语境下采取该行动，但并未永久排除这一可能性。典型的语言信号包括：

– 「现阶段不会～」「目前尚未考虑立即～」– 「今后仍有考虑的余地，但」 – 以暂时性、条件性或取决于未来情况为框架的拒绝

代表性示例包括：

– 「目前没有改变股息政策的打算。」– 「预计不会在当前中期经营计划期间进行调整。」– 「我们认为眼下难以做到，但今后会考虑。」

• *-2*（强拒绝）：当回答明确且坚决地拒绝某项拟议的未来行动、决策或政策立场，且没有留下任何重新考虑的余地时，适用此项。此类回答表达了坚定且最终的否定立场。典型的语言信号包括：

– 「没有计划～」– 「不会进行～」「否认～」

代表性示例包括：

– 「没有计划进行股票分割。」– 「不会对相关业务进行投资。」–「这并非本公司的设想。」

B *JF-TE* 标注指南

标注理念。 *JF-TE* 旨在识别日本年度证券报告的附注和标注（注記・※）中明确定义或提及的财务与会计术语。这些部分主要发挥功能性和解释性作用，用于定义会计概念、明确披露范围，或具体说明计算与监管条件，而非提供叙述性或管理层讨论。

标注旨在记录披露附注中出现的领域相关财务术语，以反映真实的会计和报告惯例。

B.1 文本范围

所有输入文本均仅摘自非叙述性说明部分，包括：

• 脚注和注释区块（注记・※）

• 表格下方的解释性说明

• 会计项目、指标或披露条件的定义

• 对计算方法、范围或监管处理方式的澄清

这些文本应在其定义语境和监管语境中进行解读。

B.2 标注单元

每个标注单元对应一个文本跨度（命名实体识别式跨度）。标注者应：

• 仅选择代表金融或会计术语的**最小跨度**

• 优先选择名词或名词短语片段

• 避免包含周围的解释性文本、动词或修饰语

如果某个跨度不代表金融或会计概念，则不应对其进行标注。

B.3 金融术语的定义

金融术语是指表示以下内容的词语或短语：

•会计项目或财务报表组成部分

• 金融指标、比率或估值概念

• 披露类别、报告单位或合并范围

• 由法律或监管规定界定的金融或会计概念

除非通用语言或程序性表达属于公认的金融术语，否则不应对其进行标注。

B.4 嵌套和重叠术语

金融术语可能以嵌套或重叠的形式出现。标注者应遵循以下规则：

• 如果嵌套和非嵌套金融术语各自独立代表有意义的金融或会计概念，都应予以标注。

• 当较长的术语包含较短的术语时，只要每个术语片段都具有独立的金融含义，就应分别标注。

• 对于每个术语，标注者都应始终选择**最小范围**，即使多个标注范围彼此重叠也是如此。

• 除非较短的术语脱离较长表达后不具有独立的金融含义，否则不应合并嵌套术语。

B.5 排除项

以下内容**不应**被标注：

•一般动词、形容词或解释性短语

• 纯粹的数值或单位（除非属于已命名的金融术语）

• 不具有金融含义的时间表达

• 交叉引用、脚注符号或格式标记

B.6 代表性示例

示例1：单一金融术语。

> （注）本公司为确保即使在金融危机等情况下也能稳定筹措资金，已与交易银行设定承诺授信额度。根据该合同，未动用借款余额如下。

标注片段：

- 交易银行
- 承诺授信额度
- 未动用借款余额

本例未采用嵌套标注，因为识别出的金融术语是并列披露单元，而非组合式或层级式表达。

示例2：嵌套金融术语。

> (※) 由于不存在潜在股票，因此未列示稀释后每股当期净利润。

标注片段：

- 稀释后每股当期净利润
- 潜在股票
- 1每股当期净利润
- 当期净利润
- 净利润

示例 3：未标注内容。

> （注）主要通过邮购销售的功能性标示食品“芝麻豆乳风味的大家的伙伴 DHA”、特定保健用食品“イマークS”等保健食品。

由于该句未引入特定的金融或会计术语，因此不进行标注。

C 标注者概况

所有 EBISU 中的标注均由具备扎实金融专业知识、丰富从业经验且日语能力出众的标注者完成，从而确保领域忠实度和标注可靠性。

其中一位标注者是日本的博士生，具备扎实的金融数学学术背景，并在日本学习和居住了数年。其研究聚焦于人工智能在金融领域的应用，并曾参与针对金融推理任务。在攻读博士学位之前，其作为金融行业的量化研究员积累了专业经验，从而实际接触真实世界的金融数据和分析工作流程。

另一位标注者是某日本金融科技公司的研究员，在日本金融市场拥有近二十年的经验。他拥有经济学博士学位，在日本公司金融、财务报告和投资者沟通方面具备深厚的专业知识。作为日语母语水平的使用者，他熟悉日本财务披露中的语言、文化和制度惯例，包括业绩报告、管理层评论以及股东大会材料。

标注者的学术训练、行业经验和语言能力相结合，使他们能够构建高质量的日本金融基准。他们的专业知识确保标注既准确无误，又扎根于具体语境，从而为评估日语金融语言理解提供可靠基础。

D 标注流程

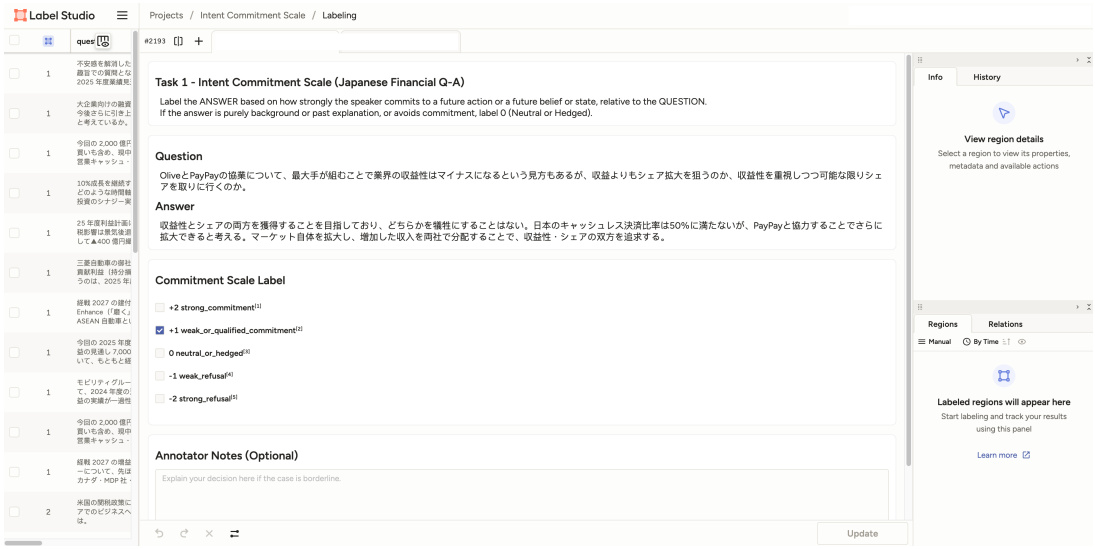

图 3：*JF-ICR* 标注流程的 Label Studio 界面。

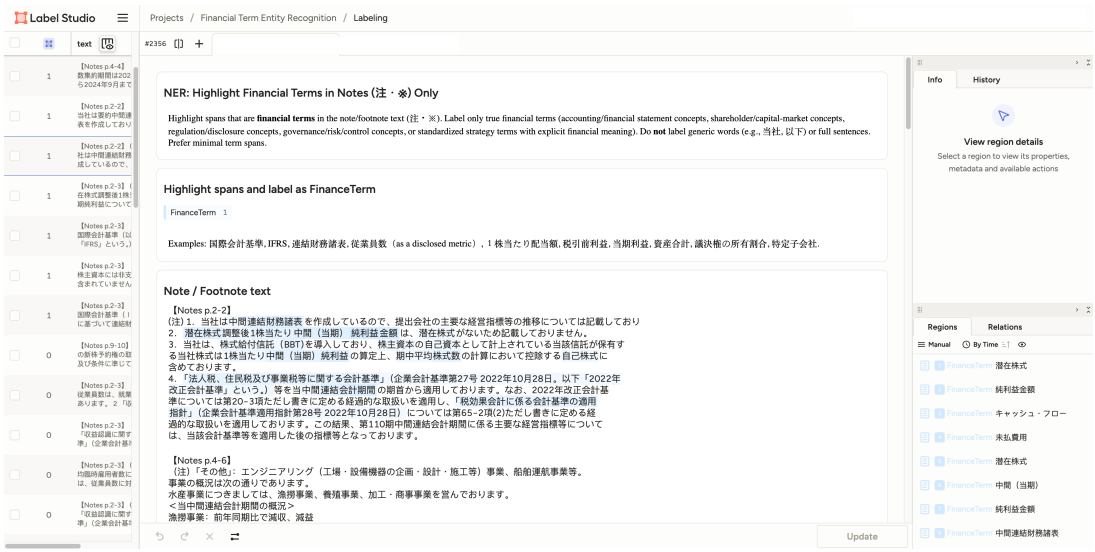

图 4：*JF-TE* 标注流程的 Label Studio 界面。

E 质量验证

为评估*JF-ICR* 和 *JF-TE* 标注的可靠性，我们计算了 3 种标准的**标注者间一致性指标**：宏平均 F1、Cohen 的 κ 和 Krippendorff 的 α。下面概述它们的正式定义。

宏平均 F1。令 $\mathcal{C}$ 表示类别集合，对于每个类别 $c \in \mathcal{C}$，我们计算

$$\text{Precision}_c = \frac{TP_c}{TP_c + FP_c}, \tag{1}$$

$$\text{Recall}_c = \frac{TP_c}{TP_c + FN_c}. \tag{2}$$

类别级F1分数为

$$\text{F1}_c = \frac{2 \cdot \text{Precision}_c \cdot \text{Recall}_c}{\text{Precision}_c + \text{Recall}_c}. \tag{3}$$

宏平均 F1 对逐类F1分数进行等权平均:

$$\text{Macro-F1} = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \text{F1}_c. \tag{4}$$

与准确率相比，宏平均 F1 对所有类别赋予相同权重，因此对偏斜的标签分布更加敏感。

Cohen 的 κ。对于两位标注者 r_1 和 r_2，令

$$P_o = \sum_{c \in \mathcal{C}} p_{cc}, \tag{5}$$

$$P_e = \sum_{c \in \mathcal{C}} p_{c\cdot}\, p_{\cdot c}, \tag{6}$$

其中，p_{cc} 是两位标注者都将类别 c 赋予标注项的经验概率，$p_{c\cdot}$ 和 $p_{\cdot c}$ 是相应的边缘概率。

Cohen 的 κ 通过减去偶然一致性来校正观察一致性:

$$\kappa = \frac{P_o - P_e}{1 - P_e}. \tag{7}$$

值为 $\kappa = 1$ 表示完全一致，而 $\kappa = 0$ 则对应于偶然水平的同意。

Krippendorff 的 α （名义量表）。
Krippendorff 的 α 能够适用于多个标注者和缺失标签。其定义为

$$\alpha = 1 - \frac{D_o}{D_e}. \tag{8}$$

对于名义类别，距离函数为

$$\delta(a, b) = \begin{cases} 0, & a = b, \\ 1, & a \neq b. \end{cases} \tag{9}$$

令 N 为项目数量，n_i 为项目 i的标注数量，n_{ia} 为项目 i 获得标签 a的次数，n_a 为标签 a 出现的总次数。那么

$$D_o = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{n_i - 1} \sum_{a \neq b} n_{ia}\, n_{ib}, \tag{10}$$

$$D_e = \frac{1}{N(N-1)} \sum_{a \neq b} n_a\, n_b. \tag{11}$$

当标注者完全同意时，$\alpha = 1$；数值越大，表示标注可靠性越高。

F 评估指标

*JF-ICR*任务是一个意图识别问题，被视为在标签集 $\{+2, +1, 0, -1, -2\}$ 上进行的多分类任务。给定预测标签 $\hat{y}_i$ 和第 i 个实例的标注金标准标签 y_i，**准确率** 定义如下:

$$\text{Acc} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}(\hat{y}_i = y_i)\,, \tag{12}$$

其中，$\mathbf{1}(\cdot)$ 是指示函数，N 是实例总数。

JF-TE 任务采用两级指标设计，以体现日本披露文件中金融术语的层级结构。

最大金融术语F1。对于每个实例，我们从金标准标注（M_{gold}）和预测结果（M_{pred}）中提取最大金融术语集合。我们计算:

$$TP = \left| M_{\text{gold}} \cap M_{\text{pred}} \right|. \tag{13}$$

$$\text{Precision} = \frac{TP}{\left| M_{\text{pred}} \right|} \tag{14}$$

$$\text{Recall} = \frac{TP}{\left| M_{\text{gold}} \right|} \tag{15}$$

$$\text{F1} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{16}$$

嵌套金融术语的 HitRate@K。对于第 i 个实例，令 G_i 表示金标准嵌套术语集合，$P_i^{(K)}$ 表示排名前 K 的预测术语：

$$\mathrm{HR@K}(i) = \frac{\left|G_i \cap P_i^{(K)}\right|}{|G_i|}. \tag{17}$$

最终分数是对所有有效实例取平均所得：

$$\mathrm{HR@K} = \frac{1}{N}\sum_{i=1}^{N} \mathrm{HR@K}(i). \tag{18}$$

我们报告 HR@1、HR@5 和 HR@10，以衡量不同排名容忍度下的表现。