面向保险业务人员线上培训绩效评估的多维度打分方法

李 昊，陈 昱，许冰玉，袁鹏宇，徐秀云

（人保信息科技有限公司研发中心，北京 100010）

摘要：针对保险业务人员线上培训效果量化评估难的问题，提出一种基于自然语言处理与机器学习技术的多维度打分方法。该方法首先构建了多维度评分体系，以精准捕捉并量化培训过程中的关键指标，在此基础上引入二级动态权重分配策略，通过智能融合多维度指标实现对学员培训效果的全面综合评价。实验结果表明，该方法能够高度模拟专家评价，确保评估结果的客观性与准确性，从而有效提高保险业务人员的线上培训成效。

关键词：保险在线培训；量化评估；自然语言处理；多维度评分

中图分类号：TP389.1 **文献标志码**：A **文章编号**：1671－1807(2025)04－0009－07

近年来，随着信息技术的不断发展，传统线下培训模式正加速向灵活便捷的在线培训模式转型。相较于线下培训常面临的课程内容固化、资源分配不均等问题，在线培训支持学员随时随地访问课程，进行个性化的专业知识学习，并通过语音互动功能参与实战模拟训练。然而，随着在线培训模式的普及，新的问题也浮现出来——如何以高效且可量化的方式评估学员的在线学习成效。传统人工专家评分虽能确保评估的精准度，但在面对大规模学员群体时，其时效性却难以保证。因此，亟须构建智能化、自动化的线上培训评估方法，在保证评估准确性的同时，提升整体评估效率。

当前，在线教育领域针对学员课堂表现的评价方法主要包括两大核心模块：一是构建评价体系，通过数据分析、挖掘等技术确定一系列能从学员学习数据中提取并量化的评价指标；二是设定指标权重，采用专家打分、层次分析或数学统计等科学方法，建立各评价指标与学员表现间的量化关系，明确各指标在整体评价中的权重。例如，魏顺平[1]从学习投入与学习产出两大维度出发，设定出勤率、互动频率等具体指标，并借助层次分析法合理分配各指标权重；罗凤娥等[2]结合模糊层次分析法和三角模糊函数，提出针对航空安全检查员的培训能力指标体系与评估模型；庄倩倩[3]设计了基于层次分析法的线上教学评价体系，用于评估和识别影响线上教学效果的关键因素；周宇等[4]设计了涵盖参与类、交互类、自律类三个维度共11项的学习评价指标，并利用多元线性回归模型确定各指标权重，以精准评估学员表现；燕贤青等[5]设计了一套针对在线教育学习过程的评价系统，通过收集学员学习记录，为教学管理提供参考。以上研究主要侧重于对在线学习或培训效果的辅助性评价分析，在面对离散或复杂多变的评价数据时，效果往往不尽如人意，且大多研究聚焦于特定应用场景，难以直接适用于保险行业培训效果量化评估的需求。

本文针对保险业务人员在线培训的实战模拟环节，创新性地设计了一种多维度实训表现评价算法。该算法首先融合多种自然语言处理（natural language processing，NLP）模型，构建了针对学员回复答案的全方位量化评价指标体系，接着采用基于监督学习的二级动态权重分配算法，实现对各个维度评价指标的加权求和，以提高评分结果的客观性和准确性，达到高度模拟专家评分的效果。

1 方案介绍

保险业务人员在实际工作中需要频繁面对客户，展现出强互动与强交流的特点，这使得保险业

收稿日期：2024-08-20

作者简介：李昊（1985—），男，吉林白城人，硕士，高级架构师，研究方向为人工智能；陈昱（1997—），男，江西上饶人，硕士，初级工程师，研究方向为自然语言处理；通信作者许冰玉（1994—），女，河南驻马店人，博士，工程师，研究方向为计算机视觉；袁鹏宇（1998—），男，甘肃天水人，硕士，初级工程师，研究方向为自然语言处理；徐秀云（1994—），女，安徽合肥人，硕士，架构师（三级），研究方向为技术创新。

务人员在线培训需求更加注重实践模拟和语言沟通技能的培养。因此，在设计针对保险业务人员的实训表现评价算法时，需要重点考虑对学员回复文本内容和情感状态的分析和评价。

保险业务人员在线实战课程示例如表1所示。课程中包含若干标准问文本 $Q_i(i=1,2,3,\cdots)$ 及其对应的标准答文本 $A_i^{\text{Truth}}(i=1,2,3,\cdots)$，这些文本能够模拟客户与业务人员之间的真实对话流程。在培训过程中，学员需要针对系统顺序提出的标准问进行语音回答 $\text{Audio}_i(i=1,2,3,\cdots)$。随后，系统基于语音识别(automatic speech recognition，ASR)技术，将学员回答音频转换为文本结果 A_i，以便后续的文本分析和评估。

表1　业务人员在线实训模拟课程样例

标准问	标准答
我现在想要购买XX类的产品，用于XX，有什么推荐吗？	我们公司有产品A、B、C、…
相比于Y公司的产品，你们产品有什么亮点？	面向于您这样的客户，该产品……

本文设计的多维度培训绩效评估算法包含多维度评价指标模块与二级动态权重分配模块两个模块，首先通过综合考虑多个评分维度，为学员的每次问答进行评分，得到单次问答分数 $s_i(i=1,2,3,\cdots)$，接着通过加权计算得出学员在整个课程的最终得分。整体算法流程如图1所示。

1.1　多维度评价指标模块

为了准确评价学员在培训过程中的成绩和实效，结合保险行业培训专家的专业意见，提炼出五个核心评价指标：要点命中率（I_1）、表达准确性（I_2）、语言流畅度（I_3）、用语规范性（I_4）、情感一致性（I_5）。采用NLP等技术将学员输入系统的音频和文本映射到各评价指标中，进而计算得到学员在各项指标上的具体得分，实现对学员培训表现的量化评价。

1.1.1　要点命中率

要点命中指标是衡量学员在回答标准问时，其回复内容是否全面覆盖了标准答中所列举的所有要点，强调了保险业务员在与客户沟通时，必须能够准确且全面地表达业务关键信息。具体而言，设标准答 A_i^{Truth} 中包含若干关键点 $K_j^i(j=1,2,3,\cdots)$，则学员的回答文本 A_i 需要尽可能涵盖 K_j^i，且回答命中的关键点数量越多表明其回答质量越高。

结合Maarten[6]、TF-IDF等关键词抽取算法，实现对标准答文本关键词的提取。首先提取标准答中长度较大的关键信息，利用KeyBERT，将标准答文本 A_i^{Truth} 转化为稠密向量 $\mathbf{V}_i^{\text{Truth}}$，随后在 A_i^{Truth} 中寻找与 V_i^{Truth} 余弦相似度最高的词汇 n 元组（n-gram），将其作为 A_i^{Truth} 的关键词。接着采用基于词频统计的传统关键词发现算法TF-IDF，通过统计 A_i^{Truth} 中各词的词频以及该词在其他标准答文本中出现的次数，判断该词的重要性。

设学员答文本 $A_i=\{\text{word}_1^{A_i},\text{word}_2^{A_i},\text{word}_3^{A_i},\cdots\}$ 由一系列词组成，为了评估学员答文本对标准答中关键要点的覆盖程度，计算学员答词组与标准答词组的交集数量，计算公式为

$$I_1=\frac{|A_i\cap K_j^i|}{|K_j^i|} \tag{1}$$

式中：I_1 为要点命中率，取值范围为 $[0,1]$，I_1 越接近1代表命中的要点越多；运算符 $|\cdot|$ 代表计算列表元素数量。

1.1.2　表达准确性

表达准确性衡量了培训中学员回答与标准答之间总体的相似程度，关注学员的回答是否传达了标准答的内在意义。鉴于自然语言的复杂性，评估表达准

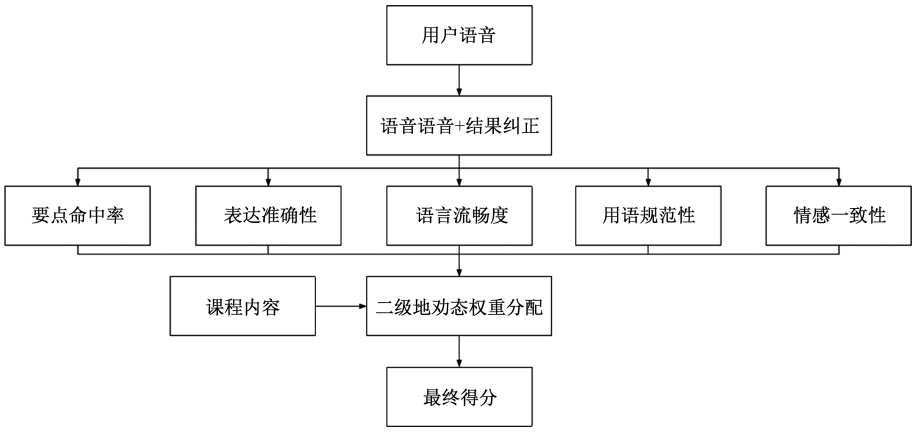

图1　多维度培训绩效评估算法流程

确性时不仅需要考虑学员遣词造句与标准答间的重合程度,还应该考量两者在语义层面上的相似度。

为量化学员答 A_i 与标准答 A_i^{Truth} 在语义上的近似程度,引入文本嵌入技术,将高维的自然语言文本嵌入可计算的向量空间中。首先采用智源文本嵌入模型(BAAI general embedding,BGE)文本嵌入模型[7],将输入语段转化为多级向量表示,然后分别提取学员答和标准答的语段标记向量 $\boldsymbol{V}_i$ 和 $\boldsymbol{V}_i^{\text{Truth}}$,并利用余弦相似度计算两个向量间的相似程度。

记学员表达准确性指标为 I_2,其计算公式为

$$I_2 = \max\left(\frac{\boldsymbol{V}_i \boldsymbol{V}_i^{\text{Truth}}}{\|\boldsymbol{V}_i\| \times \|\boldsymbol{V}_i^{\text{Truth}}\|}, 0\right) \tag{2}$$

式中:$\|\cdot\|$ 为模长运算符。需要注意的是,余弦相似度值域为[-1,1],其中值小于 0 代表语义相反,等于 0 代表语义无关,大于 0 代表语义相似。为将相似度映射为分数,令[-1,0)区间的值为 0,表示当学员答与标准答语义相反则不得分。

1.1.3　语言流畅度

表达水平的高低直接影响着保险从业人员与客户之间的沟通效果和客户体验。因此在实战模拟培训中,还需要考察学员回答问题时语言的通顺性和流畅度。本文采用困惑度指标来评估一段文本的语言流畅程度。困惑度最初被设计用于评估语言模型性能,其核心理念在于,当语言模型能够准确预测文本中每个词的出现概率时,说明该模型性能越好,生成的文本也更加流畅自然,困惑度也会越低。为了实现对学员回答语言流畅度的评估,采用 WWM-BERT 语言模型[8]预测学员答 A_i 中每个词出现的概率,然后基于这些概率值计算学员回答的困惑度指标:

$$\text{PPL} = \sqrt[N]{\prod_{i=1}^{N} \frac{1}{p(\text{word}_i \text{word}_1 \text{word}_2 \cdots \text{word}_{i-1})}} \tag{3}$$

式中:PPL 为困惑度指标,理论阈值为[1,+∞);N 为文本长度;$p(\cdot)$ 为模型输出的概率分布。如果学员回答文本 A_i 中出现较多的停顿、语气助词或语序逻辑错乱等现象,模型输出的概率分布会变得分散,导致计算出的困惑度指标 PPL 增大。

学员答的困惑度指标 PPL_i,实际会受到标准答用词的影响。设 A_i^{Truth} 的困惑度为 $\text{PPL}_i^{\text{Truth}}$,在计算语言流畅指标时,将其定义为归一化后的 PPL_i 的 $\text{PPL}_i^{\text{Truth}}$ 比值,即

$$I_3 = \text{Nor}\left(\frac{\text{PPL}_i}{\text{PPL}_i^{\text{Truth}}}\right) \tag{4}$$

式中:$I_3 \in [0,1]$,表示语言流畅度指标;Nor(·)为归一化操作。

1.1.4　用语规范性

为了评估学员回答文本的用语规范性,采用 DFA 算法检测文本 A_i 中的违规用语。DFA 算法的优点是在读取违规词典后,能够快速且高效地对大规模文本数据和复杂的违规词匹配模式进行处理。首先构建状态转换图,该图以违规词典中的每个字为节点构建一棵状态树,若 A_i 中的字符能依次匹配树上的节点,则认为该文本包含违规用语。

设用语规范性评价指标为 I_4,为避免学员在实战模拟中使用违规用语,根据实际场景需求将 I_4 设定为 0 或 1,即若文本中检测到违规用语,则 $I_4 = 0$,反之则 $I_4 = 1$。

1.1.5　情感一致性

在人际沟通中,情感作为非言语信息的重要组成部分,影响着信息的传达效果。本文不仅考虑对学员回复文本的评估,还考虑学员回复语音情感与标准答情感的一致性评估。

设学员输入音频为 Audio_i,采用多模态情感分析方法从 Audio_i 中提取情感特征,并将这些特征与 A_i^{Truth} 中的情感进行对比,判断学员在回答问题时所持情绪是否与标准答中应当传递的情感相同。

对于 Audio_i,采用 emotion2vec 模型[9]将音频域数据中频率、音量等特征转化为向量表示,随后建立分类层将这些向量映射到各情感类别,得到 Audio_i 在各情感类别上的概率分布 D_i。同时采用经过领域数据微调的 BERT 架构模型提取标准答文本情感特征,得到 A_i^{Truth} 的情感类别分布 D_i^{Truth}。为衡量 D_i 与 D_i^{Truth} 间的近似程度,引入分布间距离函数 KL 散度,计算公式为

$$\text{KL}(D_i^{\text{Truth}} D_i) = \sum_{C \in \text{情感类别}} D_i^{\text{Truth}}(c) \ln \frac{D_i^{\text{Truth}}(c)}{D_i(c)} \tag{5}$$

KL 散度的取值范围为[0,+∞],KL 越大代表分布间差异越多。接着将 KL 散度转化为区间[0,1]的指标得分,计算公式为

$$I_5 = \max\left(1 - \frac{\text{KL}(D_i^{\text{Ttuth}} \mid D_i)}{\text{thr} - 1}, 0\right) \tag{6}$$

式中:thr 为人为设定的阈值。

1.2　二级动态权重分配模块

在得到学员单次答案的多维度评分结果(I_1,I_2,I_3,I_4,I_5)后,还需要计算学员单次互动的得分 s_i,并且对所有单次互动分数 s_i 进行加权求和,得到评价学员参与完整课程的总成绩(Score)。分别采

用基于岭回归的多维评价指标权重分配和基于隐含狄利克雷分布(latent Dirichlet allocation,LDA)的对话权重分配方法来求解 s_i 和 Score。

1.2.1　基于岭回归的多维评价指标权重分配

设要点命中率(I_1)、表达准确性(I_2)、语言流畅度(I_3)、用语规范性(I_4)、情感一致性(I_5)5 个指标对应的权重为 $w_i^j, j=1,2,3,4,5$,学员每次互动得分为 s_i ,则 s_i 的计算公式为

$$s_i = \sum_{j=1}^{5} w_i^j I_j \tag{7}$$

式中:s_i 的取值为[0,1]。

由于各评价指标的计算过程相对独立,因此多维评价指标的权重分配可简化为线性回归权重拟合问题。为克服传统多元线性回归或 Lasso 回归可能导致的权重稀疏性问题,采用岭回归模型对权重进行拟合,同时还考虑了课程类型的不同对权重的影响。设课程类型为 T_n 的实战模拟课程,存在 N 次学员实训互动历史数据、专家对每次互动的评分 s_i' ,以及各维度评价指标得分 I_j ,基于岭回归的多维评价指标权重分配算法学习目标为 Loss,则 Loss 的计算公式为

$$\text{Loss} = \frac{1}{N}\sum_{i=1}^{N}\left(\sum_{j=1}^{5} w_i^j I_j - s_i'\right)^2 + \lambda \| w \|_2^2 \tag{8}$$

在确定损失函数后,采用梯度下降法不断迭代求解 5 个评价指标的权重值 w_i^j 。

1.2.2　基于 LDA 的对话权重分配

课程 $C=\{QA_1^{\text{Truth}}, QA_2^{\text{Truth}}, QA_3^{\text{Truth}}, \cdots\}$ 由若干标准问答文本 QA_i^{Truth} 组成,设学员针对课程 C 的得分为 Score,其计算公式为

$$\text{Score} = \sum_{i=1}^{N} w_i s_i \tag{9}$$

式中:学员总得分 Score 为单轮对话得分 s_i 的加权和,取值为[0,1]。为了量化课程 C 在各个主题上的“关注程度”,采用 LDA 分析课程总体主题分布情况[10],获取课程 C 文本的文档主题分布 $D_{\text{Topic}} = p(\text{Topic} = \text{topic}_i C)$ 和主题 topic_i 的词分布 $D_{\text{word}} = p(\text{word} = \text{word}_i \text{Topic} = \text{topic}_i)$,接着利用该分布情况量化各个 QA_i^{Truth} 在总体课程中的重要性,作为单次互动得分的权重 w_i 。

设 QA_i^{Truth} 由一系列词 word_j 组成,通过计算可以得到每个词在 D_{word} 中不同 topic_i 的概率值 $p_j^{\text{topic}_i}$,以及 topic_i 在 D_{Topic} 中出现的概率 p_{topic_i} ,则 QA_i^{Truth} 所占总课程的权重 w_i 计算公式为

$$w_i = \frac{w_i' - \min w'}{\max w' - \min w'} \tag{10}$$

式中:w_i' 为 QA_i^{Truth} 中每个词重要性的加权和;$\max w'$、$\min w'$ 分别为所有 w_i' 中最大值和最小值。w_i' 的计算公式为

$$w_i' = \sum_{j=1}^{5}\sum_{i=1}^{N} p_j^{\text{topic}_i} p_{\text{topic}_i} \tag{11}$$

2　实验与分析

为验证本文提出的多维度打分方法的合理性与有效性,创建了保险业务人员在线培训评分数据集:首先收集保险业务人员在真实对练培训课程中所产生的对话语音,并从中随机采样得到 700 条有效训练记录,涉及 7 种不同类型的课程各 100 条样本,每条训练记录样本包含若干次问答互动记录,总共 3 512 次互动记录;接着邀请行业内具有丰富培训经验的专家导师对这 700 条训练记录进行独立打分,包括对学员参与课程的总分进行评价,以及对学员单次互动的评分,并取专家打分的成绩平均值作为学员训练最终得分。数据集的统计特征如表 2 所示,其中采用的分词方法为 jieba 分词。

从表 2 可以看出,专家打分结果的标准差较大,表明该数据集的样本分散性较强。为了更直观地展示这一特点,绘制了单次问答互动专家打分结果的范围分布如图 2 所示。从图 2 中可以看出,专家打分结果呈现两极分化特性:在 0～15 分段和 85～100 分段的频数较高,其他分段频数较低且分布不均匀。针对该类型数据集,传统基于层次分析或线性回归的打分算法可能难以平衡各项评价指标,导致评分结果不准确。本文设计了基于 NLP、监督学习等技术的多维度打分方法,在面对离散且复杂的数据集时具有一定的鲁棒性。

2.1　实验参数设置

在多维度评价指标计算公式中,针对要点命中率指标,设置标准答分词数量的 10%(向上取整)作为关键词数量,允许提取 4 元组的词组,表达准确性的文本嵌入模型选择为 768 维向量模型,情感一致性计算时的阈值设置为 5。

在二级动态权重分配算法中,将岭回归的正则

表 2　数据集统计特征

样本数量	标准问		标准答		学员答		专家打分	
	平均长度	分词长度	平均长度	分词长度	平均长度	分词长度	平均值	标准差
700(课程)/3 512(问答)	17.06	10.16	121.22	73.66	84.90	50.60	41.20	43.10

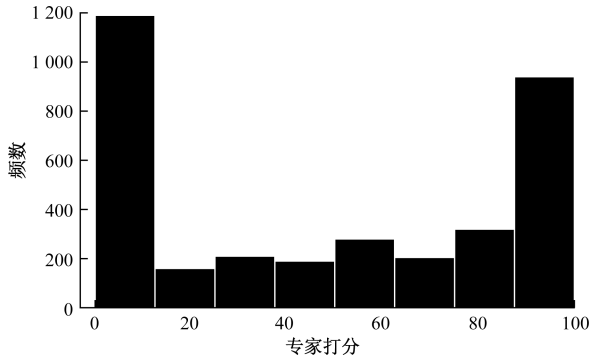

图 2　单次问答互动专家打分结果分布

化参数设置为 0.1,考虑到各个指标都有意义且指标间独立性较大,对于权重的稀疏性不做要求。此外,将对于对话权重分配中的 LDA 模型主题数设定为 5。

本文提出的多维度打分方法中,仅基于岭回归的多维评价指标权重分配算法需要监督训练,但所需拟合参数量较少,因此采用 5:5 的比例随机划分训练集与测试集。

2.2　实验结果评价指标

为了评价多维度打分方法的准确性,采用均方根误差(root mean squared error,RMSE)来衡量算法打分结果和专家打分结果之间的差别,其计算公式为

$$\mathrm{RMSE}=\sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i-\hat{y}_i)^2} \tag{12}$$

式中:N 为样本数量;y_i、$\hat{y}_i$ 分别为真实值(专家评分结果)和预测值(算法打分结果)。RMSE 越小表示算法效果与真实预测值越接近。

为了评估多维度打分方法的鲁棒性,还引入中位绝对误差(median absolute error,MedAE)来计算预测值和真实值之间的中位数绝对差异。MedAE 指标考虑了误差的异常值和极端值,数值越小说明模型性能越优秀,其计算公式为

$$\mathrm{MedAE}=\mathrm{median}(|y_i-\hat{y}_i|) \tag{13}$$

式中:median(·)为中位数计算函数。

2.3　实验结果分析

接下来通过实验分别验证多维度评价指标和二级动态权重分配算法的合理性及必要性。

2.3.1　多维度评价指标消融实验

为了评估多维度打分方法中各维度评价指标在整个评估体系中的重要性和必要性,采用消融实验的方式,逐一从多维度评价指标中移除某个维度并进行打分,然后将实验结果进行对比。此外,还采用单维度评价指标作为基线模型,观察多维度指标能否通过综合各维度信息,产生比单一维度更准确的评估效果。需要注意的是,由于用语规范性指标在单独使用时无实际意义,在实验中不考虑该维度单独评价结果。消融实验整体实验结果如表 3 所示,其中加粗数字代表最佳评估效果,短横杠“—”表示相应模型无需训练且不参与训练集结果比较。

从表 3 中可以看出,利用单个评价指标为学员回答进行打分时,算法表现结果均不如综合 5 个维度评价指标的综合评分结果。其中,“要点命中率”“表达准确性”维度在单独评分时的误差较小,说明专家在对学员回答进行评价时,比较重视学员答案是否包含考题要点以及所表述内容的含义是否贴近标准答案。逐一排除各维度指标时,其评分结果相对于完整多维度评价方法均有不同程度的下降。

为了进一步分析各评价指标对于总体评分结果的影响,本文统计了在不同指标排除后打分误差增强情况,如图 3 所示。从图 3 中可以看出,当排除“要点命中率”指标时,打分结果的 RMSE 和 MedAE 误差项分别增加了 24%和 68%,排除“表达准确性”指标时,RMSE 和 MedAE 误差项分别增加了 27.9%和 32%。这说明,要点命中率指标和表达准确性指标在本文提出的多维度评价指标体系中较为重要。此外,在排除“语言通顺度”和“情感一致性”指标时,打分结果的 RMSE 和 MedAE 误差项也有 8%～10%的上升,说明语言通顺度、情感一致性指标同样对多维度评价体系存在积极贡献。各个指标相辅相成,共同在构建能够替代人工专家的智能化线上评分方法中起到了重要作用。

2.3.2　二级动态权重分配算法对比实验

为了验证本文提出的二级动态权重分配算法的

表 3　消融实验结果

模型	RMSE(训练集)	MedAE(训练集)	RMSE(测试集)	MedAE(测试集)
要点命中率	—	—	24.88	7.56
表达准确性	—	—	31.59	15.54
语言流畅度	—	—	58.57	54.50
情感一致性	—	—	54.37	45.13
多维度评价指标(排除要点命中率)	19.83	6.39	21.51	9.73
多维度评价指标(排除表达准确性)	20.59	7.37	22.15	7.64
多维度评价指标(排除语言流畅度)	18.03	6.77	19.11	6.41
多维度评价指标(排除情感一致性)	17.49	6.31	19.08	6.25
多维度评价指标	14.62	5.94	17.31	5.78

有效性，以平均权重法为参考，分别进行多维评价指标权重分配和对话权重分配两个模块的对比实验。其中，多维评价指标权重分配方法针对单次对话进行训练和测试，对话权重分配方法则针对课程总分进行评估。实验结果分别如表 4 和表 5 所示，其中加粗数字代表最佳评估效果。

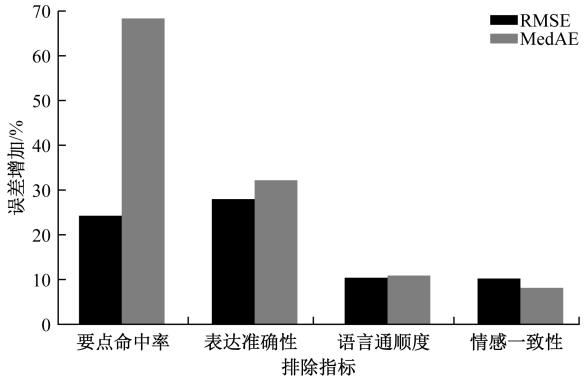

图 3　不同指标排除后误差增加情况统计

表 4　单次互动得分对比实验结果

权重分配方法	RMSE（训练集）	MedAE（训练集）	RMSE（测试集）	MedAE（测试集）
平均权重	—	—	23.19	9.48
基于岭回归的多维评价指标权重	14.62	5.94	17.31	5.78

表 5　课程总分对比实验结果

权重分配方法	RMSE	MedAE
平均权重	30.24	12.39
基于 LDA 的对话权重分配	22.16	8.10

表 4 和表 5 显示，相比于基线方法，本文所设计的基于岭回归的多维评价指标权重分配算法，以及基于 LDA 的对话权重分配算法，在 RMSE 指标上均表现更优效果。这一结果表明，本文提出的权重分配策略能够更加精准地识别不同评价指标以及不同课程间的差异性，从而使评价结果更加接近专家打分。

另外，从中位数绝对误差 MedAE 来看，本文提出的权重分配方法远小于基线方法的 MedAE 值，说明在面对极端值或者异常数据时，本文提出的权重分配方法具备更强的鲁棒性。为进一步分析验证，统计了专家打分、本文提出方法、平均权重法分别对单次互动的打分结果，如图 4 所示。从图 4 可以看出，这三种方法的中位数和上四分位数较为相近，但下四分位数差异较大。其中，专家打分的下界（0 分）和下四分位数非常接近，这表明测试样本中大量分数聚集于低分区。采用动态权重法计算

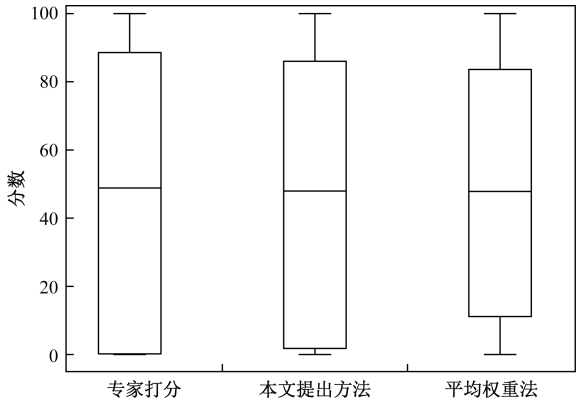

图 4　单次互动各打分方式结果箱线图

学员参与完整课程的总成绩，其下四分位数与专家打分统计结果较为接近，而平均权重法则表现不佳。这说明，面对极端离散数据，多维度评价指标需要结合二级动态权重分配算法才能够更好地发挥效果。

3　结论

针对保险业务人员在线培训中效果量化难、精准度不足等痛点问题，本文创新性地提出一种融合多维度评价指标和动态权重分配机制的在线绩效评估方法。该方法综合考量学员回答中要点命中率、表达准确性、语言流畅度、用语规范性、情感一致性等多个关键指标，对学员的每次回答进行单维度评分，在此基础上设计了二级动态权重分配算法，通过加权计算得出学员在整个培训课程的最终得分。实验结果表明，本文提出的多维度评分方法在保险业务人员的在线实战模拟训练中，能够获得接近专家打分的效果，具有较高的准确性和鲁棒性。

参考文献

[1] 魏顺平. 在线学习自动评价模式构建与应用研究[J]. 中国远程教育，2015(3)：38-45.

[2] 罗凤娥，俎振洲，周广杯，等. OJT 视角下基于 FAHP 的航空安全监察员培训能力评估[J]. 科技和产业，2021，21(5)：121-125.

[3] 庄倩倩. 基于层次分析法的线上教学评价研究[J]. 江苏科技信息，2022(18)：64-66.

[4] 周宇，应鑫迪，陈文智. 在线学习过程评价模型研究——以“学在浙大”在线教学平台为例[J]. 现代教育技术，2023，33(7)：118-125.

[5] 燕贤青，陈凤凤，沈丽. 应用型高校在线学习过程评价系统研究[J]. 无线互联科技，2024，21(8)：21-25.

[6] MAARTEN G. KeyBERT：minimal keyword extraction with BERT[DB/OL]. [2024-08-02]. https://doi.org/10.5281/zenodo.4461265.

[7] XIAO S T，LIU Z，ZHANG P T，et al. C-pack：packed

resources for general Chinese embeddings[C]. //In Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. Washington, DC, USA: 2024: 641-649.

[8] CUI Y M, CHE W X, LIU T, et al. Pre-training with whole word masking for Chinese BERT [C]. //ACM Transactions on Audio, Speech and Language Processing. Toronto, Canada: 2021: 3504-3514.

[9] MA Z Y, ZHENG Z S, YE J X, et al. emotion2vec: self-supervised pre-training for speech emotion representation [DB/OL]. [2024-07-30]. https://arxiv. org/pdf/2312. 15185v1.

[10] 李一鸣, 叶中华. 基于 LDA 主题模型的智慧社区新闻文本分析[J]. 科技和产业, 2022, 22(8): 116-122.

Multi Dimensional Scoring Method for Online Training Performance Evaluation of Insurance Business Personnel

LI Hao, CHEN Yu, XU Bingyu, YUAN Pengyu, XU Xiuyun

(Research and Development Center, PICC Information Technology Company Limited, Beijing 100010, China)

Abstract: To address the challenge of quantitatively assessing the effectiveness of online training for insurance professionals, a multidimensional scoring method based on natural language processing and machine learning technology is proposed. Initially, a multidimensional scoring framework is devised to precisely capture and quantify pivotal metrics throughout the training process. Subsequently, a two-tiered dynamic weight allocation strategy is introduced, facilitating a comprehensive evaluation of trainees' effectiveness through the intelligent integration of these multifaceted indicators. The experimental results show that this method closely emulates expert evaluation, guaranteeing the objectivity and precision of assessment outcomes, ultimately enhancing the effectiveness of online training for insurance practitioners.

Keywords: online insurance training; quantitative evaluation; natural language processing; multi-dimensional rating