# 条件付き対数尤度ベクトルに基づく視覚言語モデルのモデルマップの初期検討

田中 大　中村 純也　衣川 和尭　岡田 拓也
遠藤 伶　美野 秀弥　河合 吉彦
NHK 放送技術研究所
{tanaka.m-oc, nakamura.j-hy, kinugawa.k-jg, okada.t-im,
endou.r-mm, mino.h-gq, kawai.y-lk}@nhk.or.jp

## 概要

言語モデルを対数尤度ベクトルで表し、モデル間の関係を可視化するモデルマップ手法が提案されている。この手法では、多数のモデルを共通空間に配置しモデル同士の関係を俯瞰することができる。本稿では、このモデルマップ手法を視覚言語モデル（Vision-Language Model; VLM）へ拡張する枠組みを示す。複数の VLM に画像とテキストのペアデータセットを与えて、モデルが出力するテキストの生成分布の差をモデル間の距離として計測する。画像キャプションペアデータに関する条件付き対数尤度から各 VLM の対数尤度ベクトルを構成し、次元削減による可視化を行った結果、モデルファミリーや世代差に対応するまとまりがマップ上に形成されることを確認した。また、データの負の対数尤度の平均と 8 種類の公開ベンチマークスコアの相関を調べたところ、全体平均スコアとの相関は弱く、OCRBench のみ中程度の相関が観察された。

## 1　はじめに

近年、多数の視覚言語モデル（Vision-Language Model; VLM）が公開され、画像キャプション生成や視覚質問応答などの幅広いタスクで高い性能を示している。一方で、モデルの種類が急増し、複数のベンチマークを用いた評価をすべてのモデルに対して行うことは計算コストが高い。新しいモデルを採用する際、ベンチマークスコアに加えて特性の近い既存モデルが分かると、適切な比較対象を選びやすい。

大規模言語モデル（Large Language Model; LLM）に対しては、固定したテキスト集合に対する対数尤度ベクトルを用いて、多数のモデルを共通空間に配置することで、特性の近いモデルを可視化する手法が提案されている [1]。この論文では、モデルを確率分布として扱い、対数尤度ベクトルから得られる距離でモデルの分布間距離（KL ダイバージェンス）が近似できるという理論的証明がなされている。また、対数尤度ベクトルはベンチマークスコアを一定程度説明・予測できることも報告されている。

本研究の主目的は、このモデルマップ手法 [1] を VLM に拡張することである。VLM の入力は画像を含むため、LLM の議論をそのまま適用できない。本稿では、VLM の出力するテキスト $x$ を画像 $I$ で条件付けられたテキスト生成分布と捉え、モデルの差を条件付き分布 $p(x|I)$ の比較として定式化する。

実験では、OpenVLM Leaderboard[2] に掲載される 56 種類の実際のモデルを対象とし、MS-COCO[3] の画像キャプションペアデータを用いてに 2 次元マップによる可視化を行う。これにより、モデルファミリーや世代差に対応するまとまりがマップ上に形成されるかを確認する。また、公開されている 8 種類のベンチマークスコアと対数尤度ベクトルの相関を調べ、VLM における対数尤度ベクトルの性質を考察する。

## 2　モデルの対数尤度ベクトル

以下では LLM のモデルマップ作成手法を説明し、VLM への拡張を試みる。

### 2.1　LLM マップにおける定式化

従来手法である入出力モダリティがテキストに限定された LLM のモデルマップ手法を示す。テキスト $x$ をトークン列 $x = (y_1, \ldots, y_n)$ とすると、言語モ

デル $p_i$ は

$$p_i(x) = \prod_{t=1}^{n} p_i(y_t \mid y_{<t}) \tag{1}$$

を与える自己回帰型確率モデルである。

言語モデル $p_i$ に対し、対数尤度関数

$$\ell_i(x) \coloneqq \log p_i(x) \tag{2}$$

を定義する。$N$ 個のテキスト集合 $D = \{x_s\}_{s=1}^{N}$ を用意し、$\ell_i(x)$ を各サンプルで評価することで、モデル $p_i$ を

$$\boldsymbol{\ell}_i \coloneqq (\ell_i(x_1), \ldots, \ell_i(x_N))^\top \in \mathbb{R}^N \tag{3}$$

という対数尤度ベクトルで表現できる。

ただし、$\boldsymbol{\ell}_i$ の単純な距離は、モデルごとの平均的な尤度や、サンプルごとの次トークンの予測しやすさの影響を強く受ける。そこで文献 [1] では、モデル方向およびサンプル方向の平均を除去する二重中心化を施したベクトル $\boldsymbol{q}_i$ を用いる。基準分布 $p_0(x)$ の近傍にモデル $p_i(x)$ があるという仮定の下で、モデル $p_i$ とモデル $p_j$ 間の KL ダイバージェンスは

$$2\,\mathrm{KL}(p_i \| p_j) \simeq \frac{1}{N} \|\boldsymbol{q}_i - \boldsymbol{q}_j\|_2^2 \tag{4}$$

で近似できることが示されている。

## 2.2 VLM における定式化

VLM は画像キャプショニングだけでなく、視覚質問応答、画像検索、OCR、マルチモーダル推論など多様なタスクに対応する。本稿では、これらのタスクのうち、画像 $I$ が与えられたときのテキスト生成分布 $p_i(x \mid I)$ に着目し、モデル間距離を定義する。実装上は、画像を視覚エンコーダでトークン列に変換し、画像トークン列をテキストトークン列の先頭に連結した入力列として扱う。画像トークンに対応する損失はマスクし、テキスト部分のみの尤度を計算することで、条件付き対数尤度を評価できるため、容易に LLM の枠組みを VLM に拡張できる。条件付き分布と画像周辺分布の扱いを明確に整理するため、以下で定式化する。

テキスト集合ではなく、視覚言語データ集合

$$D = \{(I_s, x_s)\}_{s=1}^{N} \tag{5}$$

を用いる。ここで $x_s$ は画像 $I_s$ に対応する教師テキストであり、キャプションや回答文など、モデルが生成する対象のテキストを指す。データは基準分布

$$p_0(I, x) = p_0(I)\, p_0(x \mid I) \tag{6}$$

からサンプリングされているとする。

VLM $p_i$ は、画像が与えられたときの条件付き分布

$$p_i(x \mid I) = \prod_{t=1}^{n} p_i(y_t \mid y_{<t}, I) \tag{7}$$

を表す。本稿で対象とするのは、同一の画像分布の下で、モデル間で条件付き生成分布がどの程度異なるかを表す量

$$\mathbb{E}_{I \sim p_0(I)} \left[ \mathrm{KL}\big(p_i(x \mid I) \,\|\, p_j(x \mid I)\big) \right] \tag{8}$$

である。

一方、通常の VLM は画像の周辺分布 $p_i(I)$ を明示的には持たないため、$I$ と $x$ の同時分布をそのまま比較することはできない。そこで、すべてのモデルが同一の画像経験分布 $p_0(I)$ を共有すると仮定し、拡張同時分布を

$$p_i(I, x) \coloneqq p_0(I)\, p_i(x \mid I) \tag{9}$$

と定義する。このとき、モデル間の違いは条件付き分布 $p_i(x \mid I)$ のみによって表現される。この定義を利用することで、簡単な式変形で

$$\begin{aligned} &\mathbb{E}_{I \sim p_0(I)} \left[ \mathrm{KL}\big(p_i(x \mid I) \,\|\, p_j(x \mid I)\big) \right] \\ &\qquad = \mathrm{KL}\big(p_i(I, x) \,\|\, p_j(I, x)\big) \end{aligned} \tag{10}$$

が成り立つことが確認できる。左辺は画像で条件付けた生成分布の差の平均であり、右辺は定義した同時分布同士の KL 距離である。

## 2.3 LLM の結果を用いた VLM への拡張

式 (4) の近似関係は、確率変数 $z$ 上の分布に対する対数尤度関数 $\ell_i(z)$ の一般的な枠組みとして導かれている [1]。そこで確率変数を $z = x$ から $z = (I, x)$ に置き換え同様の議論を適用する。このとき、本来の同時分布の対数尤度は

$$\log p_i(I, x) = \log p_0(I) + \log p_i(x \mid I) \tag{11}$$

である。ただし $\log p_0(I)$ はモデルに依存せず、モデル間差分やマップを作成する際の距離計算では常に相殺される。したがって以降は簡単のため、条件付き対数尤度

$$\ell_i(I, x) \coloneqq \log p_i(x \mid I) \tag{12}$$

を用いる。データ集合 $D$ に対し、各 VLM の対数尤度ベクトルを

$$\boldsymbol{\ell}_i \coloneqq \big(\ell_i(I_1, x_1), \ldots, \ell_i(I_N, x_N)\big)^\top \in \mathbb{R}^N \tag{13}$$

と定義し、さらに LLM の場合と同様に二重中心化を施したベクトルを $\boldsymbol{q}_i$ と表す。このとき式 (4) を $(I, x)$ に拡張した近似として

$$2\,\mathrm{KL}\big(p_i(I,x)\,\|\,p_j(I,x)\big) \simeq \frac{1}{N}\,\|\boldsymbol{q}_i - \boldsymbol{q}_j\|_2^2 \tag{14}$$

が得られる。これを式 (10) と組み合わせることで、

$$2\,\mathbb{E}_{I\sim p_0(I)}\big[\mathrm{KL}\big(p_i(x\mid I)\,\|\,p_j(x\mid I)\big)\big] \simeq \frac{1}{N}\,\|\boldsymbol{q}_i - \boldsymbol{q}_j\|_2^2 \tag{15}$$

が導かれる。よって、VLM 同士の比較においても、画像で条件付けた対数尤度ベクトル（およびその二重中心化表現）を用いることが妥当だと考えられる。

### 2.4 VLM モデルマップの構成

モデル $p_i$ の表現を $\boldsymbol{q}_i \in \mathbb{R}^N$ とし、これを縦に並べた行列

$$Q = \begin{pmatrix} \boldsymbol{q}_1^\top \\ \vdots \\ \boldsymbol{q}_K^\top \end{pmatrix} \in \mathbb{R}^{K\times N} \tag{16}$$

を作る。$Q$ の行間距離に基づき、t-SNE[4]、UMAP[5] などで二次元平面に埋め込むことで VLM のモデルマップを得る。本稿の実験では、コサイン距離を用いた t-SNE により二次元マップを構成する。

## 3 実験

### 3.1 対象モデルとデータ

提案する対数尤度ベクトルに基づく VLM のモデルマップが、実際のモデルに対してどのような配置を与えるかを確認する。モデルは OpenVLM Leaderboard に掲載されているモデルのうち我々の実行環境で推論可能なモデルで、主要なモデルファミリー、InternVL 系列 [6]、Qwen 系列 [7]、LLaVA 系列 [8]、Gemma 系列 [9] などが含まれるように 56 種類の VLM を選定した。

対数尤度ベクトルの計算には MS-COCO[3] から無作為に抽出した画像キャプションペアデータを用いた。推定の安定性と計算コストのバランスを考慮し、$N = 5000$ 個の画像キャプションペアデータを用いた。

各モデルの入出力形式に従い画像を前処理し、教師キャプション $x$ に対する負の対数尤度（Negative Log-Likelihood; NLL）を算出した。サンプル別 NLL を並べて $\ell_i$ を構成し、二重中心化により $\boldsymbol{q}_i$ を得た。

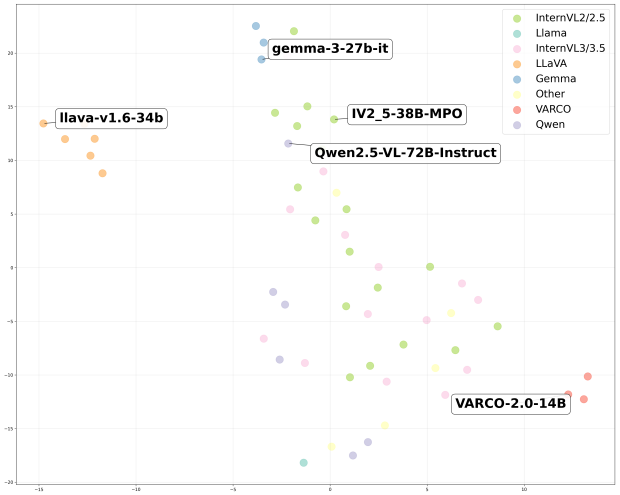


**図 1**　対数尤度ベクトルに基づく VLM モデルマップ

### 3.2 実験結果

モデルごとに得られた二重中心化ベクトル $\boldsymbol{q}_i$ による可視化結果を図 1 に示す。図 1 では可読性のため代表的なモデルのみラベルを付与し、全モデルのラベルを付与した図を付録の図 4 に示す。

### 3.3 モデルマップに見られる構造

図 1 に示す配置では、いくつかのモデル系列が近傍に集まり、モデルファミリーや世代差に対応するまとまりが形成されることを確認した。InternVL 系列では、世代・規模の近いモデルが比較的近くに配置され、Qwen ／ InternVL 近傍には連続的な広がりが見られる。LLaVA 系列は別領域にまとまりやすい傾向にある。英語と韓国語で学習され OCR を得意とするモデルである VARCO 系列 [10] が別領域としてまとまったのも特徴である。さらに、事後アラインメント手法の一つである MPO を適用したモデルが同系列の近傍に位置しつつ一定方向にシフトする例も見られ、事後アラインメントの差が配置に影響し得ることが示唆された。

ただし、図は t-SNE による 2 次元可視化であり、マップ上の大域的距離はそのまま定量解釈することはできない。本稿では、VLM に対しても対数尤度ベクトルに基づくマップが破綻せず、主要系列が同一空間上で分離して表現され得ることを確認した段階である。

### 3.4 NLL とベンチマークの相関分析

LLM においては、対数尤度ベクトルはベンチマークスコアを一定程度説明・予測できることも報告さ

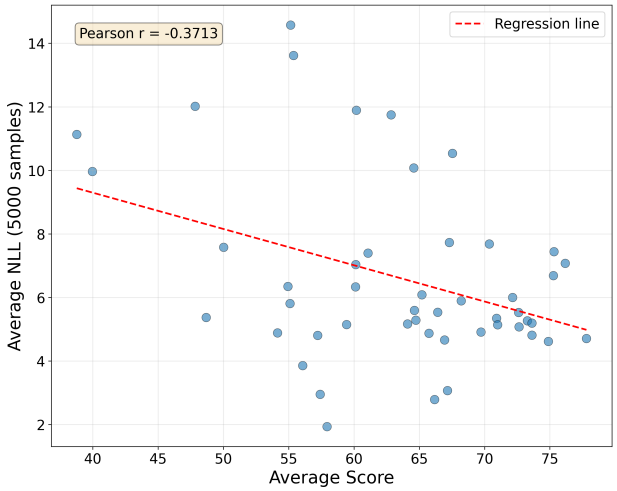


**図 2**　平均ベンチマークスコアと平均 NLL の関係

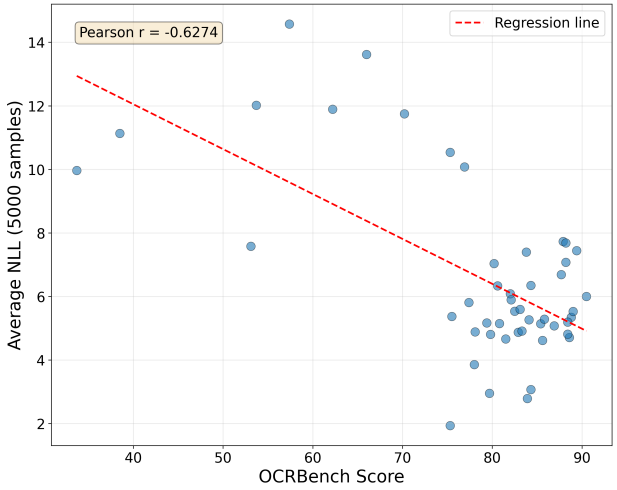


**図 3**　OCRBench スコアと平均 NLL の関係

れている。本稿の主題は対数尤度ベクトルに基づく LLM マップを VLM に対応させ、その成立を確認することであるが、画像キャプションペアデータ上の平均 NLL が公開ベンチマークのスコアとどの程度整合するかについても調査した。OpenVLM Leaderboard で多くのモデルに対して共通に報告されている 8 指標 MMBench_V11[11]、MMStar[12]、MMMU_VAL[13]、MathVista[14]、OCRBench[15]、AI2D[16]、HallusionBench[17]、MMVet[18] とのピアソン相関係数を算出した。また、8 指標の平均スコアとの相関も併せて確認した。

8 指標の平均スコアと平均 NLL の関係を図 2 に、OCRBench と平均 NLL の関係を図 3 に示す。なお、NLL は小さいほど良いため、スコアとの相関は負になりやすい。

### 3.5　相関結果に基づく性能特性の考察

画像キャプションペアデータ上の平均 NLL と公開ベンチマークスコアの関係を確認したところ、8 指標の平均スコアとの相関は弱く（図 2）、OCRBench のみ中程度の相関が観察された（図 3）。

この結果は、MS-COCO 上の平均 NLL が主として自然画像に対する短い記述生成に関する振る舞いを要約しており、複合的な推論や安全性、長文対話といった能力軸を直接反映しにくいことを示唆する。一方、OCRBench は文字情報を含む画像を扱うため、画像キャプションペアデータに含まれる文字領域に関する記述の違いが平均 NLL に現れ、相関が相対的に高くなった可能性がある。なお、本分析は平均 NLL という 1 次元の要約に基づくため、ベクトル表現を用いた場合にはベンチマークとの関係が異なる可能性がある。

## 4　おわりに

本稿では、LLM に対して提案された対数尤度ベクトルを用いたモデルマップ手法を VLM へ拡張し、画像キャプションに対する条件付き対数尤度から VLM のモデル座標を構成する枠組みを示した。モデルの出力を画像で条件付けられた生成分布として捉え、共通の画像集合に対する条件付き対数尤度からモデル間距離を定めることで、LLM の議論と整合する形で VLM を同一空間に配置できることを示した。公開 VLM を用いた初期実験では、モデルファミリーや世代差に対応するまとまりがマップ上に現れることを確認した。一方、ベンチマークとの関係を調べた結果、MS-COCO 上の平均 NLL と 8 指標平均の相関は弱く、OCRBench のみ中程度の相関が観察された。キャプショニングデータだけでは OCR 以外の指標の予測が難しいことを明らかにした。

今後は、MS-COCO とは性質の異なるデータ集合（VQA、図表理解、ハルシネーション検出など）に対しても対数尤度ベクトルを構成し、データ集合の選択によってモデルマップ上の配置やクラスタ構造がどのように変化するかを比較する必要がある。あわせて、サンプル数・モデル数・埋め込み手法（t-SNE/UMAP 等）の違いに対するマップの安定性を評価し、距離や近傍関係の解釈可能性を高めることが望ましい。これらの検討を通じて、対数尤度ベクトルに基づく VLM マップを、モデル選定や系統理解に資する実用的な評価手法へ発展させたい。

# 参考文献


[1] Momose Oyama, Hiroaki Yamagiwa, Yusuke Takase, and Hidetoshi Shimodaira. Mapping 1,000+ language models via the log-likelihood vector. **arXiv preprint arXiv:2502.16173**, 2025.

[2] Open vlm leaderboard - a hugging face space by opencompass. [Online; accessed 2025-12-22].

[3] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In **European conference on computer vision**, pp. 740–755. Springer, 2014.

[4] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. **Journal of machine learning research**, Vol. 9, No. Nov, pp. 2579–2605, 2008.

[5] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. **arXiv preprint arXiv:1802.03426**, 2018.

[6] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. **arXiv preprint arXiv:2508.18265**, 2025.

[7] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. **arXiv preprint arXiv:2502.13923**, 2025.

[8] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. **Advances in neural information processing systems**, Vol. 36, pp. 34892–34916, 2023.

[9] Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. **arXiv preprint arXiv:2503.19786**, 2025.

[10] Young-rok Cha, Jeongho Ju, SunYoung Park, Jong-Hyeon Lee, Younghyun Yu, and Youngjune Kim. Varco-vision-2.0 technical report. **arXiv preprint arXiv:2509.10105**, 2025.

[11] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In **European conference on computer vision**, pp. 216–233. Springer, 2024.

[12] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? **Advances in Neural Information Processing Systems**, Vol. 37, pp. 27056–27087, 2024.

[13] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition**, pp. 9556–9567, 2024.

[14] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. **arXiv preprint arXiv:2310.02255**, 2023.

[15] Ling Fu, Zhebin Kuang, Jiajun Song, Mingxin Huang, Biao Yang, Yuzhe Li, Linghao Zhu, Qidi Luo, Xinyu Wang, Hao Lu, et al. Ocrbench v2: An improved benchmark for evaluating large multimodal models on visual text localization and reasoning. **arXiv preprint arXiv:2501.00321**, 2024.

[16] Min Joon Seo, Hannaneh Hajishirzi, Ali Farhadi, and Oren Etzioni. Diagram understanding in geometry questions. In **Proceedings of the AAAI Conference on Artificial Intelligence**, Vol. 28, 2014.

[17] Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition**, pp. 14375–14385, 2024.

[18] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. **arXiv preprint arXiv:2308.02490**, 2023.

# A　付録

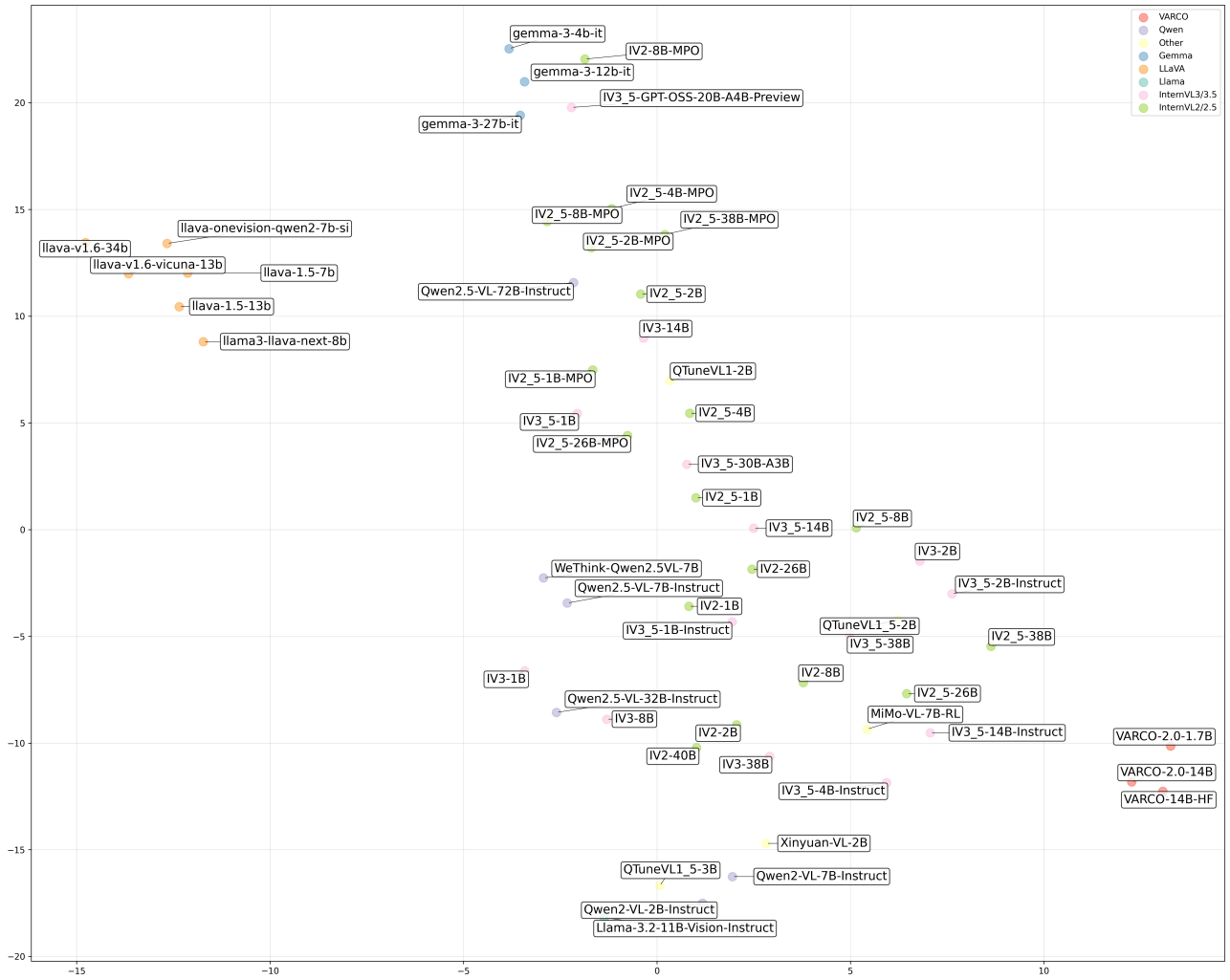


**図 4**　対数尤度ベクトルに基づく VLM モデルマップの例