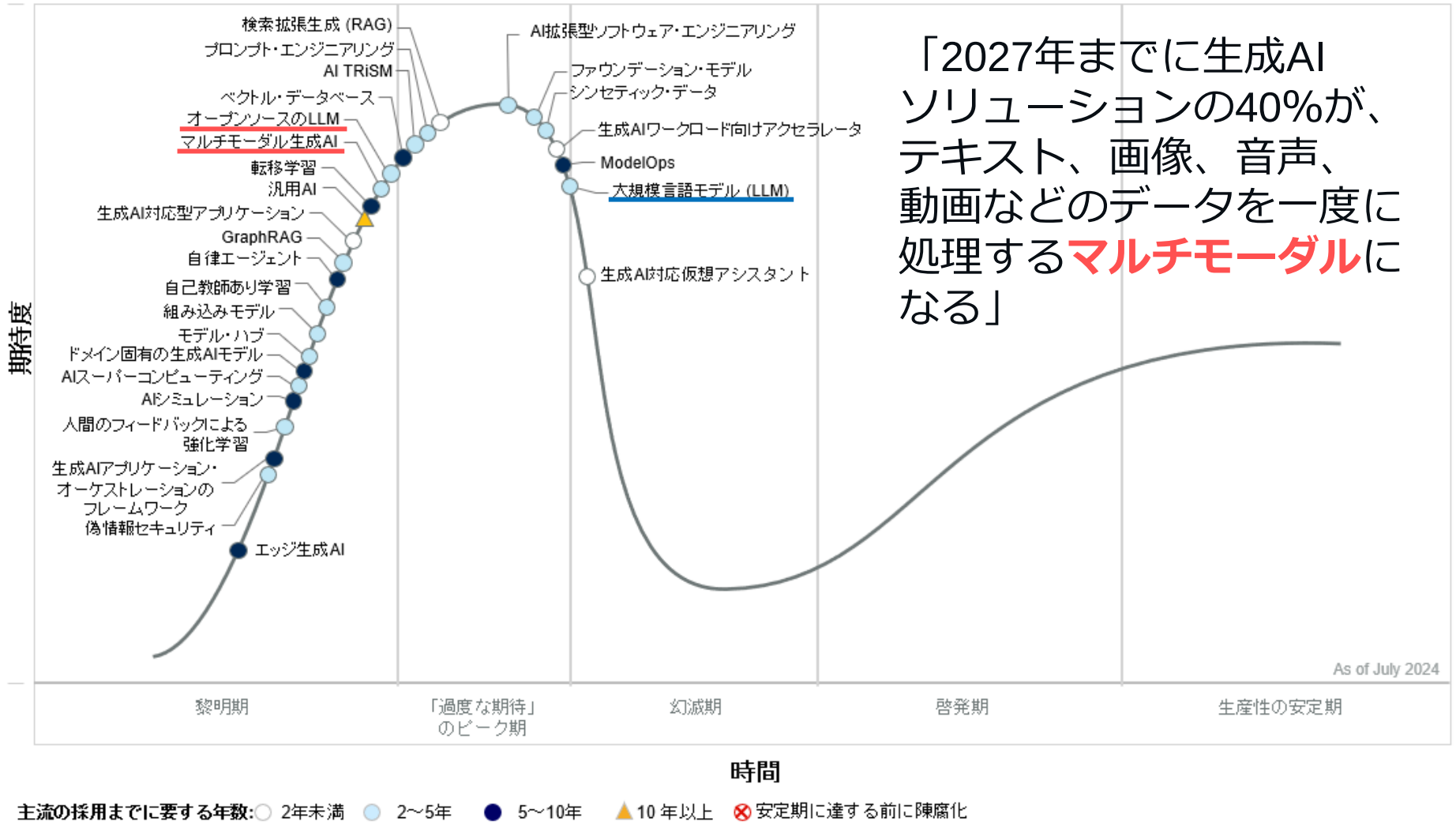
マルチモーダル基盤モデル

岡崎　直観

東京工業大学 情報理工学院情報工学系
国立情報学研究所 大規模言語モデル研究開発センター

大規模言語モデルは幻滅期に突入？



「2027年までに生成AIソリューションの40%が、テキスト、画像、音声、動画などのデータを一度に処理するマルチモーダルになる」

2024年9月10日. Gartner、「生成AIのハイプ・サイクル：2024年」を発表. https://www.gartner.co.jp/ja/newsroom/press-releases/pr-20240910-genai-hc

マルチモーダル基盤モデルとは

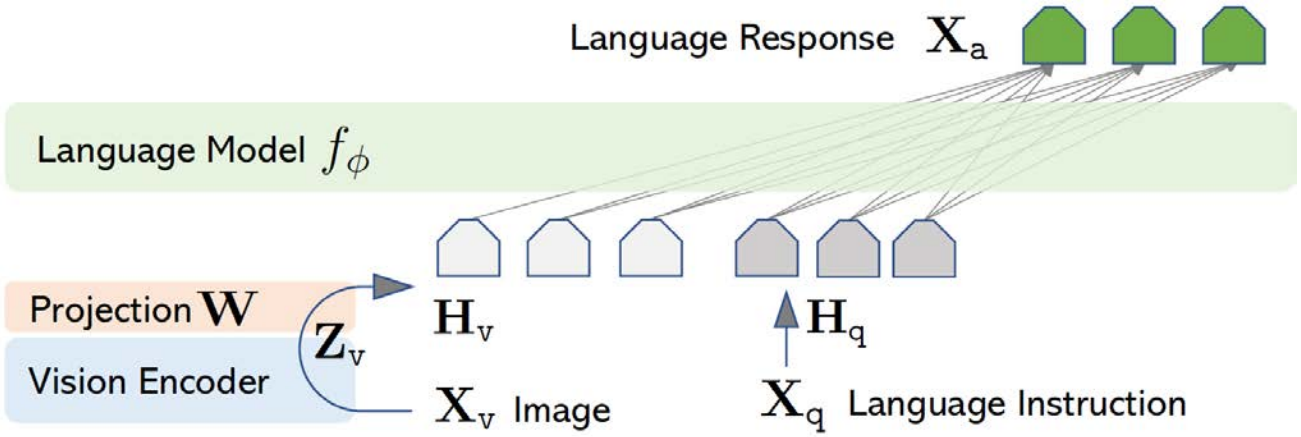
言語や画像、音声など、様々な伝達手段の情報を統合的に処理する基盤モデル

Art & Design	Business	Science
Question: Among the following harmonic intervals, which one is constructed incorrectly? **Options:** (A) Major third <*image 1*> (B) Diminished fifth <*image 2*> **(C) Minor seventh <*image 3*>** (D) Diminished sixth <*image 4*>	**Question:** ...The graph shown is compiled from data collected by Gallup <*image 1*>. Find the probability that the selected Emotional Health Index Score is between 80.5 and 82? **Options:** (A) 0 (B) 0.2142 **(C) 0.3571** (D) 0.5	**Question:** <*image 1*> The region bounded by the graph as shown above. Choose an integral expression that can be used to find the area of R. **Options:** **(A)** $\int_0^{1.5}[f(x)-g(x)]dx$ (B) $\int_0^{1.5}[g(x)-f(x)]dx$ (C) $\int_0^{2}[f(x)-g(x)]dx$ (D) $\int_0^{2}[g(x)-x(x)]dx$
Subject: Music; **Subfield:** Music; **Image Type:** Sheet Music; **Difficulty:** Medium	**Subject:** Marketing; **Subfield:** Market Research; **Image Type:** Plots and Charts; **Difficulty:** Medium	**Subject:** Math; **Subfield:** Calculus; **Image Type:** Mathematical Notations; **Difficulty:** Easy
Health & Medicine	**Humanities & Social Science**	**Tech & Engineering**
Question: You are shown subtraction <*image 1*>, T2 weighted <*image 2*> and T1 weighted axial <*image 3*> from a screening breast MRI. What is the etiology of the finding in the left breast? **Options:** (A) Susceptibility artifact (B) Hematoma **(C) Fat necrosis** (D) Silicone granuloma	**Question:** In the political cartoon, the United States is seen as fulfilling which of the following roles? <*image* 1> **Option**: (A) Oppressor (B) Imperialist **(C) Savior** (D) Isolationist	**Question:** Find the VCE for the circuit shown in <*image 1*>. Neglect VBE **Answer: 3.75** **Explanation**: ...IE = [(VEE) / (RE)] = [(5 V) / (4 k-ohm)] = 1.25 mA; VCE = VCC - IERL = 10 V - (1.25 mA) 5 k-ohm; VCE = 10 V - 6.25 V = 3.75 V
Subject: Clinical Medicine; **Subfield:** Clinical Radiology; **Image Type:** Body Scans: MRI, CT.; **Difficulty:** Hard	**Subject:** History; **Subfield:** Modern History; **Image Type:** Comics and Cartoons; **Difficulty:** Easy	**Subject:** Electronics; **Subfield:** Analog electronics; **Image Type:** Diagrams; **Difficulty:** Hard

MMMUの問題の例 (Liu+ 2024)

R Liu *et al*. 2024. MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI. *CVPR*.

画像と言語のマルチモーダル基盤モデル: LLaVA (Liu+ 2023)



LLaVAのアーキテクチャ (Liu+ 2023)

入力画像の例[1]

学習手順

1. 言語モデルf_ϕと画像エンコーダを事前に学習しておく
2. 画像の特徴量$\boldsymbol{Z}_v$を言語の空間に変換する行列$\boldsymbol{W}$を学習する

$$\boldsymbol{H}_v = \boldsymbol{W}\boldsymbol{Z}_v, \qquad \boldsymbol{Z}_v = g(\boldsymbol{X}_v)$$

※画像$\boldsymbol{X}_v$、質問$\boldsymbol{X}_q$、答え$\boldsymbol{X}_a$が組になった学習データを用いる

3. 変換行列$\boldsymbol{W}$と言語モデルf_ϕを同時に学習する（画像エンコーダのパラメータは固定しておく）

User: What is unusual about this image?
LLaVA: The unusual aspect of this image is a man ironing clothes on the back of a minivan or van. This is not a typical place to perform this activity, as one would usually iron clothes in a more stationary and safe location, such as a home, using a regular ironing board. The scene depicted in the image is peculiar as it involves a makeshift ironing setup on a vehicle, which can be both unsafe and unconventional. Additionally, it is not clear how the man is able to maintain balance and stability while ironing clothes in such an unstable environment.

H Liu, C Li, Q Wu, Y J Lee. 2023. Visual Instruction Tuning. *NeurIPS*.
[1] https://www.barnorama.com/wp-content/uploads/2016/12/03-Confusing-Pictures.jpg

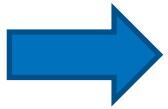
マルチモーダルWGでの取り組み

目的

- 言語や画像、音声など、様々なモダリティの情報を統合的に扱える基盤モデルを開発する
- 大規模言語モデルの言語に関する知能を外界と接続していく（グラウンディング）
- マルチモーダル基盤モデルの有用な応用、およびその評価方法を研究・確立する

今年度の取り組み

- VLMの学習データの整備・構築
- VLMの試作（学習ノウハウの蓄積）
  - LLMに画像のモダリティを追加
- VLMの評価基盤の検討・確立
- 医学文献に対するVLMの応用



来年度以降の計画

- 日本の文化や風景に対応した画像特徴量抽出器（エンコーダ）の採用・開発
- 透明性の高いマルチモーダル基盤モデルのフルスクラッチでの構築
- 動画などの新たなモダリティへの対応
- マルチモーダル基盤モデルの応用の開拓

今年度の取り組み (1/4): VLMの学習データの整備・構築

日本語版MMC4

- 画像とテキストのインターリーブ・データ
- Common Crawlの日本語ウェブページから画像をダウンロード・フィルタリングして構築

42) 口コミを見ると、土・日はいつも満席、比較的平日の方が座れるとのことでした。¥n¥n
43) ## ペットOKは外のテラス席のみ¥n¥n私達が選んだ席は、広瀬川に面したオープンスタイルのテラス席です。 [類似度: 0.264]
44) カフェに着いた時は、少し小雨が降ってる様子でしたが、幸いなことに食事の途中で小雨はやんでくれました。¥n¥n [類似度: 0.255]
45) デッキに屋根はありませんので小雨の影響でデッキは濡れていたため、持参したこまめ用のタオルを敷いています。¥n¥n [類似度: 0.209]
46) カウンターに注文をしにいくご主人をじっとみつめるこまめ。¥n¥n
47) デッキの席は全部で10席程度ありました。
48) デッキに行くまではペットも店内を通ります。 [類似度: 0.263]
49) ちなみに店内での飲食はペットNGです。¥n¥n


43


44


45


48

https://kurashiba.com/sendai-cafe-mozart

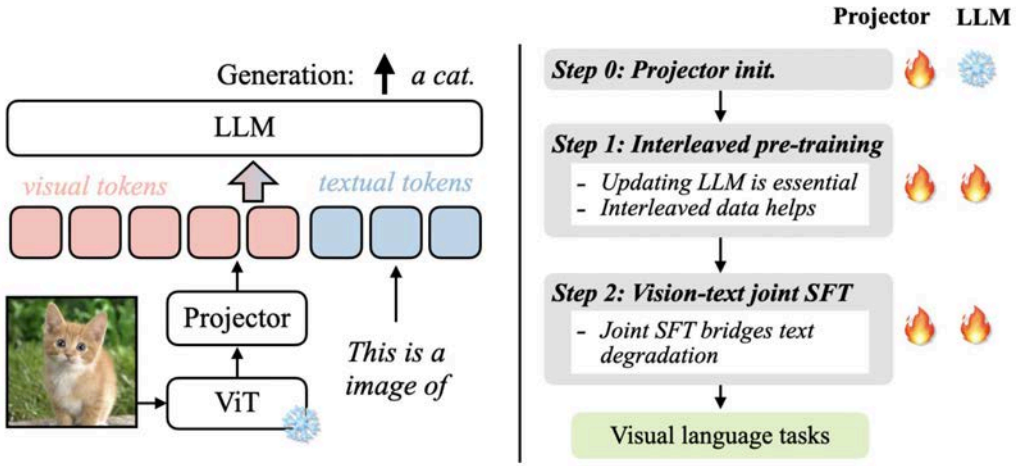
指示チューニングデータの自動生成

- 日本で撮影された画像11,810枚に対してGPT-4oを用いて質問・応答のペアを生成

https://huggingface.co/datasets/ThePioneer/japanese-photos

Q: この画像にはどんな建物が写っていますか？
A: 画像には「LaLaport」と書かれた建物が写っています。
Q: この建物のファサードには何がプロジェクションされていますか？
A: 建物のファサードにはアニメのキャラクターがプロジェクションされています。
Q: この画像の中で目立つ大きなフィギュアは何ですか？
A: 目立つ大きなフィギュアはガンダムのロボットです。
……

今年度の取り組み (2/4): 日本語LLMのVLM化



採用したアーキテクチャ (VILA: Lin+ 2024)
ViTはOpen AI CLIP (336px)、Projectorは2層のFFN、
LLMはLlama 2 7Bを用いた

Q: 鳥の羽の色は何色でしょうか？
A: 鳥の羽の色は青です。
Q: 映像に写っている少年は何をしていますか？
A: 画像の少年は、ハサミを使って紙を切っています。

Q: あなたは車を運転しており、交差点で右折しようと思っています。このまま交差点に向かってよいでしょうか？
A: いいえ

Heron Bench (Inoue+ 2024) での生成例

学習段階	学習に用いたデータ	学習部分	学習時間 (7Bモデル)
Step 0	LLaVAのPretrainデータ (595K)	Projector	8xA100, 3.5h
Step 1	mmc4-core (25M), coyo (25M)	LLMとProjector	16x8xA100, 30h
Step 2	LLaVA1.5にinstruction tuningデータに他のデータを混ぜたもの	LLMとProjector	8xA100, 25h

J Lin *et al*. 2024. VILA: On Pre-training for Visual Language Models. *CVPR*.
Y Inoue *et al*. 2024. Heron-Bench: A Benchmark for Evaluating Vision Language Models in Japanese.*arXiv*:2404.07824.

今年度の取り組み (3/4): **VLM評価基盤の構築**

日本語指示追従性評価

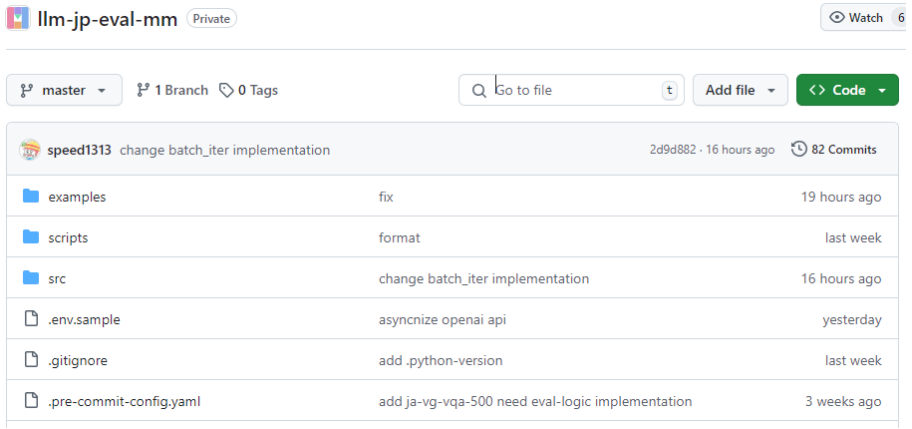
- turing-motors/Japanese-Heron-Bench
- SakanaAI/JA-VLM-Bench-In-the-Wild

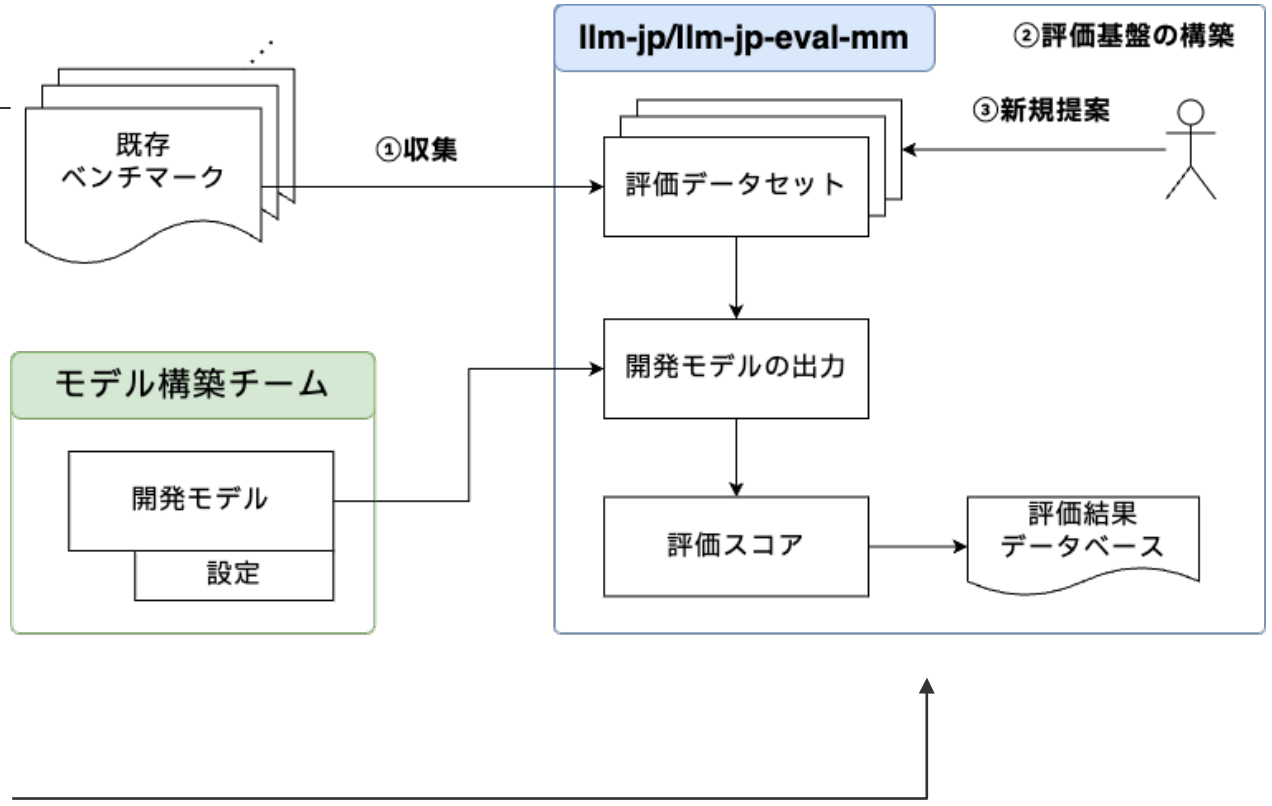
日本語VQA

※ 淡色は検討中

- JDocQA (Onami+ 2024)
- GazeVQA (Inadumi+ 2024)
- SakanaAI/JA-VG-VQA-500

VLMベンチマーク基盤の開発





E Onami, S Kurita, T Miyanishi, T Watanabe. 2024. JDocQA: Japanese Document Question Answering Dataset for Generative Language Models. *LREC-COLING*.
S Inadumi, S Kawano, A Yuguchi, Y Kawanishi, K Yoshino. 2024. A Gaze-grounded Visual Question Answering Dataset for Clarifying Ambiguous Japanese Questions. *LREC-COLING*.

今年度の取り組み (4/4): 医学文献に対するVLMの応用

J-Stageの医学文献からのコーパス構築

- PDF論文からテキストを検出・抽出し、レイアウトの分析、および読み順を解析する
- PDFの処理にSurya[1]を用いる

レイアウト解析の例[2]

マルチモーダル試験コーパスの構築

- 医師国家試験問題から問題を抽出し、別冊として収録されている画像と対応付ける

24　52 歳の女性。2 週間前に受けた人間ドックの腹部超音波検査で胆嚢の異常を指摘されたため精査目的で来院した。自覚症状はない。既往歴に特記すべきことはない。身長 158 cm、体重 64 kg。BMI 25.6。体温 36.2 ℃。腹部は平坦、軟で、圧痛を認めない。血液所見：赤血球 458 万、Hb 13.7 g/dL、Ht 41 %、白血球 7,300。血液生化学所見：総ビリルビン 0.9 mg/dL、AST 20 U/L、ALT 18 U/L、LD 148 U/L(基準 124〜222)、ALP 86 U/L(基準 38〜113)、γ-GT 28 U/L(基準 9 〜32)、CEA 1.1 ng/mL(基準 5 以下)、CA19-9 14 U/mL(基準 37 以下)。CRP 0.1 mg/dL。腹部超音波像(**別冊No. 8**)を別に示す。

この患者の治療方針で適切なのはどれか。

a　経過観察
b　抗菌薬投与
c　腹腔鏡下胆嚢摘出術
d　薬物による抗癌治療
e　経皮経肝胆嚢ドレナージ

別　冊
No. 8

第１１８回医師国家試験問題 A問題

[1] https://github.com/VikParuchuri/surya
[2] 小林 俊諒, 豊嶋 弘一, 山田 英嗣, 南 博仁, 藤枝 敦史, 玉木 茂久. Jarisch-Herxheimer反応が診断の端緒となったWeil病の1例, 日本内科学会雑誌, 112巻, 6号, pp. 1005-1011, 2023.

まとめ

- 目的
  - 言語や画像、音声など、様々なモダリティの情報を統合的に扱える基盤モデルを開発する
- 今年度の取り組み
  - VLMの学習データの整備・構築、VLMの試作、VLM評価基盤の検討・確立
  - 医学文献に対するVLMの応用に向けて
- 主なメンバー
  - 岡崎、河原、栗田、小田、相澤、金澤、笹川、前田、Zhishen、杉浦、Yin、村岡
  - 定例ミーティングに参加して頂いている方々
- 活動場所
  - 定例ミーティング: 隔週月曜日 9:00-10:00
  - #マルチモーダル Slackチャネル
  - マルチモーダルWGの活動に興味のある方はぜひお声がけください！