Supplementary Tables

72M with h/o Stage IIIA squamous cell NSCLC (pT4, pN1, cM0), diagnosed 11/2022. Prior therapy: 1. Neoadjuvant carbo/taxol/nivo x3 cycles (completed 2/14/23) 2. LUL lobectomy (4/17/23) 3. PORT radiotherapy (completed 7/31/23) 4. Carbo/abraxane/pembro x2 cycles (started 12/13/23) 5. Carbo/abraxane/durva/treme from cycle 3 (started 2/3/24) Tumor board question: Is there progression on scans? What are next line options for treatment?
• 70 year old woman who was in her usual state of health until mid-October of 2019 when she had the relatively abrupt onset of generalized abdominal discomfort, dark urine, acholic stool, low grade nausea and fatigue. • She went to her PMD who treated her presumptively for a UTI. However, 3-4 days later, she still felt unwell. She returned to her PMD who noted new scleral icterus. She was sent to Hospital for blood work and imaging. • An MRI was done on 05/19/XX showing a 2.8cm mass in the head of the pancreas, abutting the SMV. Chest imaging was negative. • The patient underwent EUS/ERCP with plastic stent placement. FNA was +adenocarcinoma. Her CA 19-9 was 334 at that time.(was plastic stent replaced by metal one)? • The patient started FOLFIRINOX on 06/20/XX and received a total of 9 cycles in the neoadjuvant setting. • Neo adjuvant chemo was interrupted by sepsis in 07/10/XX , cholangitis and an obstructed biliary stent, CT imaging on 07/11/XX suggested hepatic microabscesses.ERCP on 07/12/XX was unsuccessful with a biliary perforation. A percutaneous biliary drain was placed on 07/12/XX. • She resumed mFOLFIRINOX with cycle 3 on 8/14/XX and began cycle 9 on 11/6/XX. Treatment was held after cycle 9 in a good response due to increasing toxicity and a declining performance status. Her CA 19-9 nadir was 21. • Her imaging on 11/30/XX showed improvement in her primary pancreatic lesion and no new lesions. • On 12/26/XX, she underwent Whipple .Her tumor was 3cm, poorly differentiated, negative margins, 2/14+ nodes (T2N1; stage IIb). • IHC for MMR proteins was normal. NGS found KRAS and TP53 mutations. Interestingly, despite the decrease in CA 19-9 and size of the tumor on imaging, no treatment effect was observed on pathology. The patient went home to recover from her surgery. • On 2/25/XX, imaging and tumor markers were performed. CA 19-9 had risen from 38 post-operatively to 577. CT imaging showed small liver lesions-<1cm. • A PET-CT on 3/18/XX to further characterize the imaging findings showed hypermetabolic disease /recurrence at the tumor bed and in the liver. • The CA19.9 increased to 2351 on 3/21/XX. • Germline testing on 5/1/XX found an ATM mutation. • The patient began palliative Gem/Abr on 4/10/XX. Cycle 1 day 15 was held on 4/24/XX due to neutropenia • Gemzar and Abraxane cycle 3 day 15 was held on 06/12/XX due to neutropenia. • Her CA19.9 increased in 11/20XX. • Imaging in 12/20XX found disease progression. • She was found to have progressive disease in the liver and retroperitoneum on 07/08/XX; although notably this scan was compared to imaging done 1.5mo prior to starting Gem/Abr; CA 19-9 was up and down during this time. • She "qualified" for the CAR T-cell trial · • CT scans of the chest, abdomen and pelvis on 12/15//XX found the noted liver metastasis and minimal increase in the retroperitoneum. • Began FOLFIRI on 07/17/XX · Significant toxicity (fatigue , poor appetite, weight loss 3Ib) with cycle 1 requiring dose reduction with cycle 2 (initially Cycle 2 was held due to neutropenia 7/31/XX)· Prior to starting chemo - They were explained that the likelihood they would be ineffective with <6mo between final therapy and progressive disease. • Held FOLFIRI cycle 3 - 8/21/XX - due to neutropenia • Rising CA19.9 on 8/7/XX suggesting possible progression. She received a total of five cycles, last on 9/25/XX. CA 19-9 rose steadily during this time • The patient started palliative FOLFIRI on 07/17/XX. Dose reduce by 20% for cycle 2 due to symptoms of fatigue and nausea • Imaging on 10/6/XX showed progressive metastatic disease (increased liver metastasis and adenopathy. There is new intrahepatic ductal dilation). Patient reports that she was not doing well and tolerated chemo poorly. She was extremely fatigued, spending >70% of her waking hours in bed or chair. She was weak and tired, not leaving the house except for doctor appointments. She reported dizziness on standing; requiring IVF 3x/week. She was forcing herself to eat/drink. She suffered chronic nausea and early satiety, takes Compazine/Zofran around the clock. She had ongoing pain across her entire abdomen, she was taking 5-10mg oxycodone PRN and was requiring this around the clock, including at night. She noted that her most recent chemotherapy made her "miserable • She was under palliative care but was not ready to enroll in hospice, but was interested in Home palliative care. • Dr explained that that there were other treatments that could be tried (ie FOLFOX or trials) but that in her current physical condition, she was not a candidate for systemic therapy or clinical studies-CAR-T trial for pancreatic cancer because of her decreased performance status in the last few weeks. They were also explained that there are phase 1 trials, but these were not specific to pancreatic cancer and are basket trials for several different solid tumors. She and her daughter expressed understanding Past medical history:Left iliac deep vein thrombosis · Incidental finding on CT angio on 11/30/19 · Xarelto begun 11/30/XX, unspecified Anxiety disorder, Hypertension ,Fibromyalgia ,Diabetes ,Osteoarthritis. FH: Pancreas Cancer: Paternal Aunt Brain cancer: Paternal Cousin (dx less than 5 years old). Meds: Cetirizine Insulin Pancrelipase lpa Prochlorperazine

Table 1. Tumor Board Case Presentation Examples From Various Organizations (A) - Textual Summaries. Top shows very brief bulleted text. Bottom shows very detailed prose-like summary.

The schema includes the following fact event type categories and its descriptions:

- demographic: demographic information including age, gender, smoking status, etc.
- family_history: anything related to family's condition
- stage: cancer stages including clinical or pathological (stage I, II, etc)
- diagnosis: clinical diagnoses (other relevant medical history can go here)
- performance_status: performance statuses
- symptom: clinical symptoms
- treatment_systemic: systemic treatments for cancer such as chemotherapy or immunotherapy
- treatment_surgery: surgical treatments for cancer
- treatment_radiation: radiation treatments for cancer
- treatment_other: other types of treatments

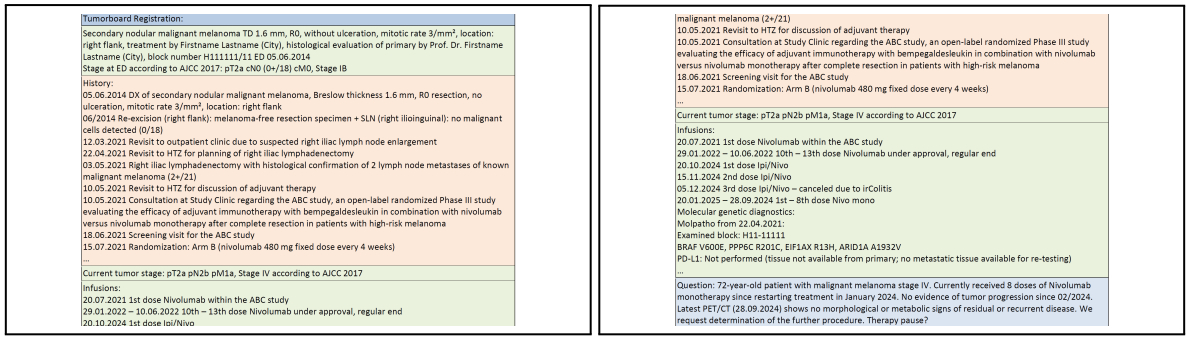

Figure 1. DS4 Structured Summaries showing a semi-structured presentation format.

- test_imaging: imaging tests such as MRI, CT, X-RAYS, PET scans, etc.
- testresult_labs: results of clinical laboratory testing (e.g. BUN, WBC)
- testresult_molecular: results of genetic/mutation testing (e.g. BRCA)
- testresult_pathology: results of pathology testings
- open_question: open questions
- other: other events not suitable in earlier classifications

rubric_criterion	event	event_argument	Fact Presence	Factual Accuracy	Importance Accurate Added-Fact	Harmfulness Inaccurate Added-Fact
1	**72M**		**not_present**			
2	Stage IIIA		**present**			
3	**squamous cell NSCLC** (pT4, pN1, cM0), diagnosed 11/2022					
4		Squamous cell NSCLC	**present**			
5		(pT4, pN1, cM0)	**not present**			
6		Nov-22	**present**			
7	**neoadjuvant carbo/taxol/nivo** x3 cycles (completed 2/14/23)					
8		neoadjuvant carbo/taxol/n	**present**			
9		X3 cycles	**not_present**			
10		completed 2/14/23	**present**			
11	LUL lobectomy (4/17/23)					
12		LUL lobectomy	**not_present**			
13		Completed 4/17/23	**not_present**			
	..	..	..			
	..	..	..			
..	Is there progression on scans?		**not_present**			
..	What are next line options for treatment?		**not_present**			
ADDED INFORMATION						
	74M			**not accurate**		**medium**
	What clinical trials is this patient still eligible for?			**accurate**	**noncritical**	
	squamous cell NSCLC, diagnosed 11/3/2022		**partially_present**			
		Diagnosed 11/3/2022		**accurate**	**critical**	

Figure 2. Graded Rubric Example

DS2
She was presented at the tumor board today. She has been recommended neoadjuvant chemoimmunotherapy.
Presented to tumor board, tumor board recommended CT-guided biopsy.
After a case discussion the consensus of the team includes the following: Follow up with [Deid: Name] [Deid: Name] to discuss adjuvant therapy. Patient does not want chemotherapy. Will be treated with Tagrisso for 3 years. The final treatment plan will be left at the discretion of the patient and the treating physician.
After a case discussion the consensus of the team includes the following: Adjuvant chemotherapy. Not recommending radiation at this time. Patient is already scheduled for radiation oncology consult tomorrow, [DATE] , at 1:00 pm with [Deid: Name] at Radiation Department. The final treatment plan will be left at the discretion of the patient and the treating physician.
DS4
Offer of a therapy break. Guidelines-conform follow-up according to schema "Discontinuation of immunotherapy", HLA positive
Offer of adjuvant therapy within the context of Study A, alternatively investigational BRAF/MEKi combo; detailed discussion regarding incurred costs due to insurance status only available abroad. Alternative adjuvant therapy under approved standard of care with targeted kinase inhibitor combination or anti-PD-1 immunotherapy.
Clear progression with new lymph node metastases and liver metastases. Initiation of anti-TNFα agent 40 mg s.c. every two weeks in combination with steroids has already been started for recurrent irUveitis. Control of irUveitis for at least 8 weeks, further therapy pause for immunotherapy due to irUveitis, no alternative therapies available. After tapering immunosuppression, enrollment in Study B or assessment for investigational oncolytic virus therapy to be considered.
Excision of in-transit metastases and continuation of systemic therapy with anti-PD-1 immunotherapy.
DS5
Consensus was that the patient is not a surgical candidate for salvage radical prostatectomy due to age, morbid obesity, and prior cryotherapy. Recommended initiation of androgen deprivation therapy (ADT) with Lupron and Casodex. Plan for consultation with Radiation Oncology to evaluate for salvage radiation, potentially following 6 months of ADT and repeat imaging.
Consensus is that the patient has completed the planned HD-MTX chemotherapy for CNS lymphoma with interval treatment response on MRI. Next steps include follow-up with Neuro-Oncology in two weeks to discuss consolidation therapy, specifically pursuing low-dose whole brain radiation therapy (WBRT) as per the patient's preference.
Proceed with palliative external-beam radiation to the left inguinal mass and right iliacus (36 Gy in 12 fractions) and consider a targeted boost to the pancreatic head if symptomatic. After radiation, initiate salvage BR + polatuzumab if tolerable. Continue pain control, nutrition support, and refer to palliative-care services.
Complete the planned eighth high-dose methotrexate cycle, then consolidate with low-dose whole-brain radiotherapy ($\approx$23 Gy) given the excellent radiographic response and her age/comorbidities. If she declines radiation, discuss enrollment in a PCNSL clinical trial (e.g., BTK-inhibitor or CAR-T). Continue oral clindamycin for the PICC-site infection, maintain current insulin and supportive care (anti-emetics, leucovorin rescue, ocular lubricants, neuropathy management).

Table 2. Tumor Board Recommendations Prediction Examples from DS2, DS4 (English synthetic based on German data), DS5

section	axis	f1	f1_1diff	kendalltau	pearson	spearman	avgcorr
Description	completeness	0.592	1.000	0.359	0.473	0.370	0.401
	factual-accuracy	0.404	0.842	0.186	0.282	0.209	0.226
	relevance	0.354	0.792	0.197	0.399	0.212	0.269
	overall	0.396	0.979	0.151	0.416	0.167	0.245
Oncological History	completeness	0.325	0.721	0.132	0.456	0.153	0.247
	factual-accuracy	0.604	0.829	-0.007	0.165	0.004	0.054
	relevance	0.367	0.742	0.249	0.520	0.282	0.350
	overall	0.429	0.804	0.201	0.514	0.222	0.312
Supporting data	completeness	0.342	0.954	0.307	0.332	0.334	0.324
	factual-accuracy	0.496	0.792	0.233	0.356	0.256	0.282
	relevance	0.338	0.867	0.380	0.543	0.436	0.453
	overall	0.438	0.858	0.417	0.484	0.455	0.452
Considered Options	completeness	0.217	0.467	-0.101	-0.106	-0.114	-0.107
	factual-accuracy	0.450	0.512	0.371	0.435	0.376	0.394
	relevance	0.142	0.517	-0.140	-0.148	-0.163	-0.151
	overall	0.167	0.671	-0.306	-0.316	-0.333	-0.318

Table 3. DS1 Average Pairwise Agreement (Number of Experts=3)

You are a helpful medical assistant.

Given a candidate text CANDIDATE TEXT and a reference text REFERENCE TEXT, rate the CANDIDATE text on a scale of 1-5 (1 is lowest, 5 is best) using the definitions below. The judgement should be based on the content of the REFERENCE TEXT as the ground truth.

Below are some guidelines for the RATING labels (choose the best option):
- rating_1: CANDIDATE TEXT and REFERENCE TEXT provide content with overlapping topics but with conflicting information; OR provide content with completely dissimilar topics; OR a mixture of conflicting information for same content topics and dissimilar topics
- rating_2: {CANDIDATE TEXT} and {REFERENCE TEXT} provide content with overlapping topics with some conflicting information, but with other content overlap
- rating_3: {CANDIDATE TEXT} and {REFERENCE TEXT} provide overlapping content and significant (but not conflicting) content differences
- rating_4: {CANDIDATE TEXT} and {REFERENCE TEXT} provide the same content but with related insignificant additional information from either text
- rating_5: {CANDIDATE TEXT} and {REFERENCE TEXT} provide the exact same content with minor language variations

Here are some simplified examples with a {CANDIDATE TEXT}, {REFERENCE TEXT} => {RATING}:
"today's blood pressure is 136/63", "today's blood pressure is 180/83" => rating_1
"today's blood pressure is 136/63", "blood sugars normal" => rating_1
"today's blood pressure is 136/63, weight is low", "today's blood pressure is 180/83, blood sugars normal" => rating_1
"today's blood pressure is 136/63, weight is low", "today's blood pressure is 136/83, blood sugars normal" => rating_2
"today's blood pressure is 136/63, weight is low", "today's blood pressure is 136/63, blood sugars normal" => rating_3
"patient reports nausea and difficulty swallowing", "pt has been losing due to nausea and swallowing problems" => rating_3
"patient reports nausea and difficulty swallowing", "pt has difficulty with eating due to nausea and swallowing problems" => rating_4
"patient reports nausea and difficulty swallowing", "pt symptoms include: nausea, swallowing difficulties" => rating_5

CANDIDATE TEXT: {}

REFERENCE TEXT: {}

RATING:

Table 4. LLM-as-Judge Generic Prompt

You are a helpful medical assistant.

Given a candidate text {CANDIDATE TEXT} and a reference text {REFERENCE TEXT}, rate the {CANDIDATE} text on a scale of 1-5 (1 is lowest, 5 is best) for the categories described below. The judgement should be based on the content of the {REFERENCE TEXT} as the ground truth. The final {OUTPUT} should be in the form of a json object with the following keys and categories defined below.

Categories:
1. completeness: (Completeness) Using the REFERENCE TEXT as the ground truth, is all necessary information required for the case summary present in the CANDIDATE TEXT? (1 - contains no relevant/helpful information, 3 - Some critical elements present but some missing, 5 - All necessary information present)
2. factual_accuracy: (Factual accuracy) Using the {REFERENCE TEXT} as the ground truth, is all content in the {CANDIDATE TEXT} medically accurate (regardless of if it is complete or relevant/helpful)? (1 - Incorrect or harmful information, 3 - Some accurate information and some inaccuracies (but may not cause harm), 5 - All information is accurate)
3. relevance: (Relevance/Helpfulness) Using the {REFERENCE TEXT} as the ground truth, is information presented in the {CANDIDATE TEXT} relevant? Is information helpful to your task (i.e., how helpful would it be to have this summary generated for you vs. having to create it yourself for a tumor board)? (1 - No information is relevant/helpful, 3 - There is both relevant/helpful and irrelevant/unhelpful information, 5 - All information is relevant/helpful)
4. overall: (Overall) Using the {REFERENCE TEXT} as the ground truth, how would you rate the overall answer from the {CANDIDATE TEXT}? (1 - Complete disagree, 2 - Would make major revisions, 3 - Agree but would make some critical revisions, 4 - Agree but would make some trivial revisions, 5 - Completely agree)

Example OUTPUT:
{ "completeness": 4, "factual_accuracy": 3, "relevance": 2, "overall": 4 }
These categories may be related but not directly computable from each other. For example, your CANDIDATE TEXT can be complete but not factually accurate. On the other hand, the CANDIDATE TEXT can have complete information but be factually inaccurate. Similary, the CANDIDATE TEXT can be factually accurate and complete but also include a lot of irrelevant and unhelpful information requiring a lot of effort to correct the text. The overall score should take all these items into account holistically, factoring need for physician to edit as well as potential harm in the incomplete/inaccurate information that can lead to erroneous medical judgements.
Here are some simplified examples with a {CANDIDATE TEXT}, {REFERENCE TEXT}, and {OUTPUT}:
Example 1 CANDIDATE TEXT:
Patient is a 68-year-old male diagnosed on 01.10.2025 with a glioblastoma in the right temporal lobe. Patient completed chemotherapy cycle for their multicentric lesion on 03.17.2020. Patient is currently receiving treatment with gemcitabine and wishes to determine alternative treatment options towards treating their multifocal disease.

Example 1 REFERENCE TEXT:
Patient is a 65-year-old male diagnosed on 01.09.2020 with an MGMT methylated and IDH wildtype glioblastoma in the left temporal lobe. Patient completed their surgery and subtotal resection of their multicentric lesion on 01.09.2020 and their chemoradiation on 03.17.2020. Patient is currently receiving treatment with pembrolizumab and bevacizumab and wishes to determine alternative treatment options towards treating their multifocal disease.

Example 1 OUTPUT:
{ "completeness": 3, "factual_accuracy": 1, "relevance": 1, "overall": 1 }
Patient information, test and treatments in the {CANDIDATE TEXT} are all factually inaccurate as compared to the {REFERENCE TEXT}; including incorrect dates, incorrect anatomic laterality, and incorrect treatments. However, completeness is 3 as the text itself does describe some history and provides the curent status and point of decision-making, though misses critical biomarker information. This led to overall lower scores for factual_accuracy, relevance, and overall.

CANDIDATE TEXT:{}

REFERENCE TEXT:{}

OUTPUT:

Table 5. LLM-as-Judge-4xes Prompt

Summary

This is a 55-year-old gentleman with an MGMT-methylated, IDH-wildtype glioblastoma of the left occipital lobe. He underwent gross total resection on 12/30/2020, followed by concurrent chemoradiation with temozolomide beginning 01/21/2021. Upon completing chemoradiation on 03/09/2021, he enrolled in clinical trial NCT02287428 (NeoVax personalized neoantigen vaccine plus pembrolizumab). He began adjuvant temozolomide on 04/12/2021, initiated NeoVax anti-cancer vaccine on 04/27/2021, and commenced pembrolizumab on 05/24/2021. He remains neurologically and radiographically stable on trial. The patient follows a ketogenic diet and is interested in identifying additional studies or alternative treatment options in the event of future disease recurrence.

Extracted Facts


```
[
  { "text": "55-year-old gentleman", "type": "demographic", "mention": "55-year-old gentleman", "attributes": [] },
  { "text": "MGMT-methylated", "type": "testresult_molecular", "mention": "MGMT-methylated", "attributes": [] },
  { "text": "IDH-wildtype", "type": "testresult_pathology", "mention": "IDH-wildtype", "attributes": [] },
  { "text": "glioblastoma of the left occipital lobe", "type": "diagnosis", "mention": "glioblastoma", "attributes":
  ["left occipital lobe"] },
  { "text": "gross total resection on 12/30/2020", "type": "treatment_surgery", "mention": "gross total resection",
  "attributes": ["12/30/2020"] },
  { "text": "concurrent chemoradiation with temozolomide beginning 01/21/2021", "type": "treatment_systemic",
  "mention": "chemoradiation with temozolomide", "attributes": ["concurrent", "01/21/2021"] },
  { "text": "enrolled in clinical trial NCT02287428 (NeoVax personalized neoantigen vaccine plus pembrolizumab)",
  "type": "treatment_other", "mention": "clinical trial NCT02287428", "attributes": ["NeoVax personalized neoantigen
  vaccine plus pembrolizumab"] },
  { "text": "adjuvant temozolomide on 04/12/2021", "type": "treatment_systemic", "mention": "adjuvant temozolomide",
  "attributes": ["04/12/2021"] },
  { "text": "initiated NeoVax anti-cancer vaccine on 04/27/2021", "type": "treatment_systemic", "mention": "NeoVax
  anti-cancer vaccine", "attributes": ["initiated", "04/27/2021"] },
  { "text": "commenced pembrolizumab on 05/24/2021", "type": "treatment_systemic", "mention": "pembrolizumab",
  "attributes": ["commenced", "05/24/2021"] },
  { "text": "remains neurologically and radiographically stable on trial", "type": "other", "mention":
  "neurologically and radiographically stable", "attributes": ["on trial"] },
  { "text": "The patient follows a ketogenic diet", "type": "other", "mention": "ketogenic diet", "attributes":
  ["follows"] },
  { "text": "interested in identifying additional studies or alternative treatment options in the event of future
  disease recurrence", "type": "open_question", "mention": "additional studies or alternative treatment options",
  "attributes": ["in the event of future disease recurrence", "interested in identifying"] }
]
```


Facts with Entailment Classification


```
[
    {
        "mention": "55-year-old gentleman",
        "attributes": [],
        "mention_classification": "ENTAILED_BY",
        "mention_classification_reason": "Context text describes 'a 55-year-old left-handed male', which logically infers '55-year-old gentleman'.",
        "attribute_classifications": [],
        "attribute_classifications_reasons": []
    },
    {
        "mention": "MGMT-methylated",
        "attributes": [],
        "mention_classification": "ENTAILED_BY",
        "mention_classification_reason": "MGMT-methylated is mentioned directly in the context text, confirming the molecular feature.",
        "attribute_classifications": [],
        "attribute_classifications_reasons": []
    },
    {
        "mention": "IDH-wildtype",
        "attributes": [],
        "mention_classification": "ENTAILED_BY",
        "mention_classification_reason": "IDH-wildtype is mentioned exactly in the context text.",
        "attribute_classifications": [],
        "attribute_classifications_reasons": []
    },
    {
        "mention": "glioblastoma",
        "attributes": [
            "left occipital lobe"
        ],
        "mention_classification": "ENTAILED_BY",
        "mention_classification_reason": "Glioblastoma diagnosis is clearly mentioned in the context text. Laterality and lobe (occipital) are not
        directly specified.",
        "attribute_classifications": [
            "OTHER_NOT_SUPPORTED"
        ],
        "attribute_classifications_reasons": [
            "The context text does not specify 'left occipital lobe'; thus, this attribute cannot be logically inferred."
        ]
    },
    {
        "mention": "gross total resection",
        "attributes": [
            "12/30/2020"
        ],
        "mention_classification": "ENTAILED_BY",
        "mention_classification_reason": "Gross total resection is stated in the context text, but the exact date is only referenced as 'December
        2020'.",
        "attribute_classifications": [
            "RELEVANT_BUT_NOT_SUPPORTED"
        ],
        "attribute_classifications_reasons": [
            "The context mentions 'December 2020', but does not give the precise date '12/30/2020'; thus, it is relevant but not supported."
        ]
    },
```


Table 6. Example Fact Extraction and Entailment

Dataset	Task	Baseline Systems	LLM-as-Judge
DS1	Case Summ/Opt Gen	DeepSeek-r1 Gemini-2.5-pro GPT4.1 Qwen-plus	 GPT4.1 Qwen-plus
DS2	Case Summ/Opt Gen	DeepSeek-r1 Gemini-2.5-pro GPT4.1 Qwen-plus	 GPT4.1 Qwen-plus
	TB Rec	DeepSeek-r1 Gemini-2.5-pro GPT4.1 Qwen-plus	 GPT4.1 Qwen-plus
DS3	Case Summ	GPT4.1	GPT4.1
DS4	Case Summ	DeepSeek-R1-0528-AWQ medgemma-27b-it gpt-oss-120b Qwen3-Next-80B-A3B-Instruct	 medgemma-27b-it gpt-oss-120b Qwen3-Next-80B-A3B-Instruct
	TB Rec	DeepSeek-R1-0528-AWQ medgemma-27b-it gpt-oss-120b Qwen3-Next-80B-A3B-Instruct	 gpt-oss-120b Qwen3-Next-80B-A3B-Instruct
DS5	TB Rec	Gemma-4-31B gpt-oss-120b	Gemma-4-31B gpt-oss-120b

Table 7. LLM models used for each institution. Best LLM-as-judge per dataset-task are bolded.

Agreements

We measure inter-rater variability in terms of F1 exact, F1 1-difference (1 point off is still considered a match), pearson, spearman, and average correlations; and average over all pairwise measurements for DS4 and DS2 in Table 8 and 9, respectively. In general, across both datasets, exact F1 agreement is low, however given a tolerance 1/5, we see high agreements of >0.8 for case summarization subtasks. For both datasets, the agreement for Options Generation drops to 0.5 F1 and 0.7 F1 for DS1 and DS2.

section	F1	F1_1diff	avg-corr	n pairs
Description	0.269	0.742	-0.087	10
Oncological History	0.338	0.808	0.201	10
Supporting data	0.348	0.809	0.055	10
Options Generation	0.235	0.611	-0.046	10

Table 8. DS4 average pairwise inter-rater agreement on the OVERALL axis, averaged across all 10 rater pairs (Number of Experts=5). F1=exact match rate; F1_1diff=match within ± 1 point; avg-corr=mean of Kendall's τ, Pearson r, Spearman ρ across pairs.

For DS4, exact-match pairwise agreement on the OVERALL axis is low (F1 between 0.24 and 0.35), while within-1-point agreement reaches 0.74-0.81 on the three case-summarization sections (Description, Oncological History, Supporting data). The recommendation section drops to F1=0.24 and F1_1diff=0.61, mirroring the trend observed for options generation on DS1/DS2: the more open-ended the task, the lower the agreement. Per-pair Pearson/Spearman/Kendall correlations average near zero across sections because DS4 ratings concentrate in a narrow 3-4 band; the ordinal-weighted Gwet AC2 over all five raters jointly is 0.60 (95% CI [0.58, 0.62]), which we report as the IAA figure.

section	axis	F1	F1_1diff	kendalltau	pearson	spearman	avgcorr
Description	completeness	0.490	0.889	0.035	0.046	0.039	0.040
	factual-accuracy	0.540	0.857	0.164	0.212	0.172	0.183
	relevance	0.546	0.886	-0.142	-0.138	-0.139	-0.140
	overall	0.390	0.887	0.069	0.085	0.071	0.075
Oncological History	completeness	0.559	0.915	0.095	0.091	0.097	0.094
	factual-accuracy	0.605	0.904	0.125	0.138	0.129	0.131
	relevance	0.394	0.797	0.245	0.265	0.273	0.261
	overall	0.419	0.893	0.285	0.294	0.302	0.294
Supporting data	completeness	0.440	0.915	0.271	0.311	0.287	0.290
	factual-accuracy	0.644	0.923	0.055	0.068	0.059	0.060
	relevance	0.554	0.906	0.246	0.257	0.260	0.254
	overall	0.485	0.908	0.183	0.192	0.201	0.192
Options Generation	completeness	0.385	0.843	-0.012	0.008	-0.009	-0.004
	factual-accuracy	0.307	0.704	0.096	0.101	0.102	0.100
	relevance	0.306	0.684	0.095	0.110	0.109	0.104
	overall	0.322	0.820	0.100	0.104	0.112	0.106

Table 9. DS2 Average Pairwise Agreement (Number of Experts=13)

TBFact Error Analysis

To assess the metric reliability, we calculate the fidelity of entailment classifications. Specifically, given a set of extracted facts, GPT4.1 and Qwen-plus conserved classification structures (same number of facts and fact-attributes) on average 99.7% and 72.3%, across DS1 case summarization, DS2 case summarization, and DS2 tumor board recommendations tasks. Indidcating that fact entailment classification is relatively stable for GPT but not for Qwen and is model-dependent. To assess the accuracy of the entailment classification, we randomly sampled two sets of encounters per totaling to 159, 255, and 42 entailment classifications. Each GPT4.1 and QWEN-plus models were assessed using a binary classification (ENTAILED_BY or not). For DS1, GPT4.1 had 30 errors out of 132 classifications while qwen-plus failed to output a full list of facts for one encounter but otherwise had 0 errors for the other encounter 26 classifications. For DS2 case summarization, GPT4.1 committed 8 errors out of 130 ; QWEN 23 errors out of 125. Finally, for DS2 tumor board recommendation, GPT commited 0 out of 23; qwen 2 out of 19. Qualitatively, we observed several occurring errors: (a) entailment classification finds it challenging when disambiguating multiply events of the same type, especially when one text condenses and the other provide each detail. For example if the same chemotherapy is started and stopped multiple times – represented in detail for one text but not the while other does not. (b) There are often many mistakes related to test such as CT's and PET scans - this is attributable to the fact that many results may not always mentioned the exact diagnostic test, though some are common (e.g. CT); thus it sometimes is a source of hallucinated inferred knowledge. (c) Related information may be classified as entailed although not mentioned, for example common dosages, or different clinical trials that use the same treatment.

Results Discussion

system	comments
descript	Didn't include PET scan. It says that molecular testing is pending, but that is wrong - the pathology report says it needs to be requested. Probably worth mentioning the size of the nodules no HAS mutation. can mention about upstaging after surgery, and on MTX Highly accurate and concise but lacks important clinical detail e.g., tumor burden, nodal disease, contralateral lesion. It correctly states the stage, biomarkers, surgery, and unresected 10R node. However, it leaves out the patient's broader comorbidity burden, MRI brain result, and postoperative course, which matter for adjuvant treatment decisions.
onchx	No need to list each treatment day Left lower back pathology not needed Worth mentioning PFTs as being poor Good - but this case can easily be summarized by an entry for CT chest, EBUS biopsy, and PET alone. A bit wordy.
supp data	Does not mention PET. Also, I think for the pathology, it is important to clarify that it is inconclusive and lymphoma and atypical carcinoid not excluded The chest xray is not helpful. Should mention CT chest. Very concise and accurate, focusing strictly on core oncologic data. However, too minimal Covers key events correctly, including staging, surgery, and pathology. However, omits postoperative course details and lacks explicit tumor board framing
options	It incorrectly states that surgery is only for atypical carcinoid. Does not mention SABR radiation. Clinical trial has wrong link. Misses thoracic surgery consult. Some irrelevant entries. Should mention liquid biopsy. The options for systemic chemoimmunotherapy is not relevant to stage I. For stage IA, brain MRI not in NCCN. inappropriate option eg adj pembo/Chemo +IO

Table 10. DS1/DS2 Expert Comments for Best System

DS4 Supplementary Tables

	DS4			
system	DEEPSEEK	GEMMA/GEMINI	GPT	QWEN
eval	GPT	GPT	GPT	GPT
BLEU	0.020	0.017	0.014	0.020
ROUGE	0.151	0.101	0.093	0.145
BERTScore	0.616	0.594	0.550	0.609
LLM-as-judge	2.950	2.233	2.662	2.736
LLM-4xes	2.275	1.638	1.931	2.182
Human score	3.674	2.820	3.390	3.260

Table 11. Performance for Recommendation Prediction

DS5 Supplementary Tables

	DS5			
system	GEMMA	GPT	GEMMA	GPT
eval	GPT	GPT	GEMMA	GEMMA
BLEU	1.310	0.893	1.310	0.893
ROUGE	0.092	0.071	0.092	0.071
BERTScore	0.835	0.814	0.835	0.814
LLM-as-judge	2.325	2.337	1.928	1.952
LLM-4xes	1.952	1.964	2.012	1.771
TBFact	0.071	0.050	0.110	0.081
TBFact-precision	0.057	0.037	0.099	0.057
TBFact-recall	0.246	0.303	0.323	0.409
Human score	2.145	2.108	2.145	2.108

Table 12. Performance for Recommendation Prediction

Source	Subtype	Included Types
ORCA note type	Oncology and radiation care	Hematology/Oncology - Inpt Record; Hematology/Oncology - Outpt Record; SCCA - Outpt Record; SCCA - Inpt Record; SCCA EH - Outpt Record; SCCA NWH - Outpt Record; SCCA Treatment History; Previous Oncology Treatments; Radiation Oncology - Outpt Record; Radiation Oncology - Inpt Record
	Imaging, pathology, and molecular testing	Medical Imaging - Outpt; Radiology Note; Molecular Profiling
	Surgery and procedures	Surgery - Inpt Record; Operative Report; Post-Op Surgical Note; Surgery Admit/Initial Consult Note; Procedure Report
	Consultations	Consultation - Inpt; Consultation - Outpt Record; Neuro-Oncology - Inpt Record; Neurosurgery - Inpt Record; GI - Inpatient Record; Pulmonary Medicine - Inpt Record
	Hospital and emergency care	History & Physical; Admit Note; Discharge Summary; Interim Summary; ED Note
EPIC note type	Oncology and radiation care	Weekly On Treatment Visit; End Of Treatment Summary; IMRT Medical Necessity; Historic SCCA Note
	Imaging, pathology, and molecular testing	Result QuickNote
	Surgery and procedures	Op Note; Brief Op Note; Brief Procedure Note; Procedures
	Consultations	Consults
	Hospital and emergency care	Progress Notes; H&P; Interval H&P Note; Hospital Course; Discharge Summary; ED Provider Notes; ED Notes; Addendum Note
Specialty type	Hematology/Oncology	Medical Oncology; Hematology/Oncology; Oncology Medicine; Hematology; Surgical Oncology
	Radiation oncology	Radiation Oncology; Radiation Oncology Med
	Surgery	General Surgery; Thoracic Surgery; Neurosurgery; Urology; Colorectal Surgery; Vascular Surgery; Plastic and Reconstructive Surgery; Transplant Surgery Medicine; Cardiac Surgery; Maxillofacial Surgery
	Radiology	Diagnostic Radiology; Radiology Med; Interventional Radiology
	Pathology	Pathology

Table 13. Included Note and Specialty Types for Longitudinal Clinical Context Construction