Supplementary Materials

Improving Reliability and Explainability of Medical Question Answering through Atomic Fact Checking in Retrieval-Augmented LLMs

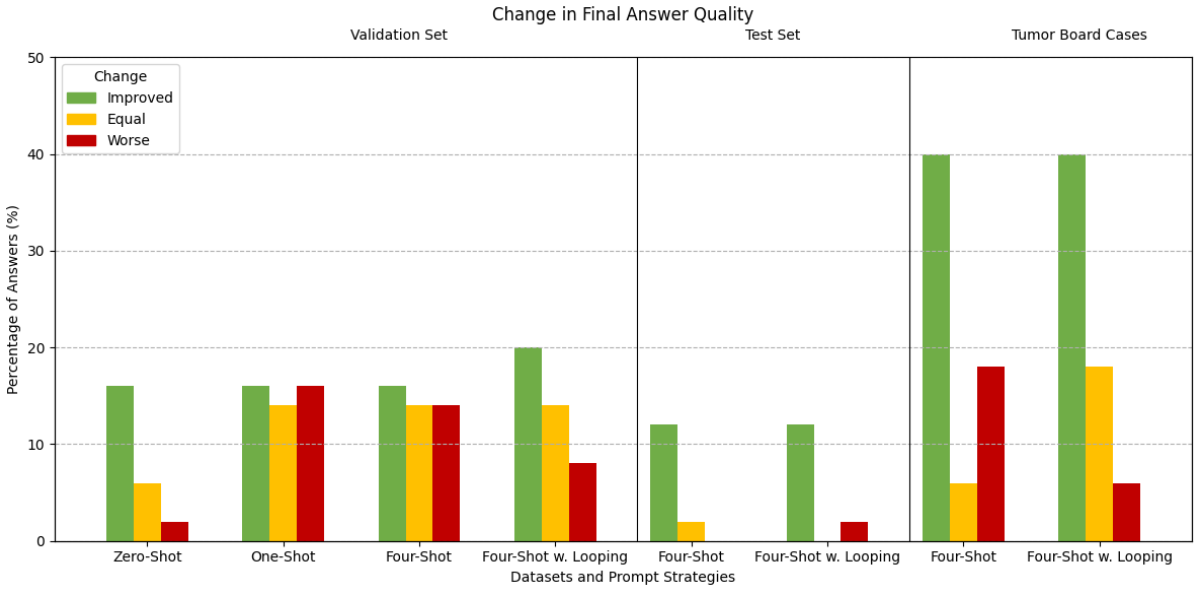
Juraj Vladika[†1], Annika Domres[†2], Mai Nguyen[2], Rebecca Moser[2], Jana Nano[2], Felix Busch[3], Lisa C. Adams[3], Keno K. Bressem[3,4], Denise Bernhardt[2], Stephanie E. Combs[2,5,6], Kai J. Borm[2], Florian Matthes[1], Jan C. Peeken[*2,5,6]

1 Department of Computer Science; TUM School of Computation Information and Technology; Technical University of Munich, Garching, Germany
2 Department of Radiation Oncology, TUM University Hospital Rechts der Isar, TUM School of Medicine and Health, Technical University of Munich, Munich, Germany
3 Department of Diagnostic and Interventional Radiology, TUM University Hospital Rechts der Isar, TUM School of Medicine and Health, Technical University of Munich, Munich, Germany
4 Institute for Cardiovascular Radiology and Nuclear Medicine, TUM University Hospital, German Heart Center Munich, TUM School of Medicine and Health, Technical University of Munich, Munich, Germany
5 Institute of Radiation Medicine (IRM), Helmholtz Zentrum München (HMGU)
6 German Consortium for Translational Cancer Research (DKTK), Partner Site Munich, Munich, Germany

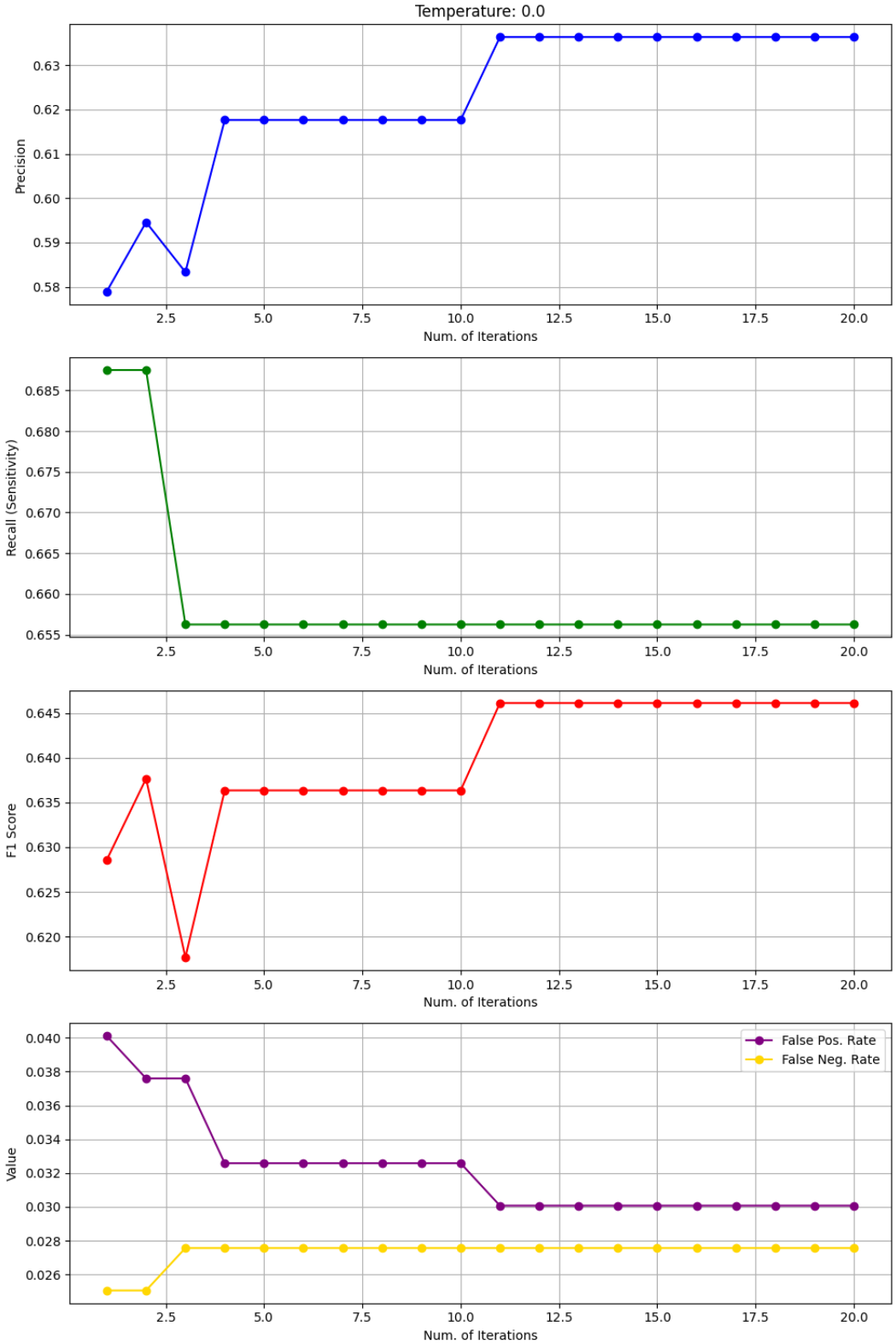
† These authors contributed equally to this work.

*__Corresponding Author__: Jan C Peeken, MD PhD
Email: jan.peeken@tum.de
Address: Department of Radiation Oncology, TUM School of Medicine and Health, TUM University Hospital Rechts der Isar, Munich, Germany

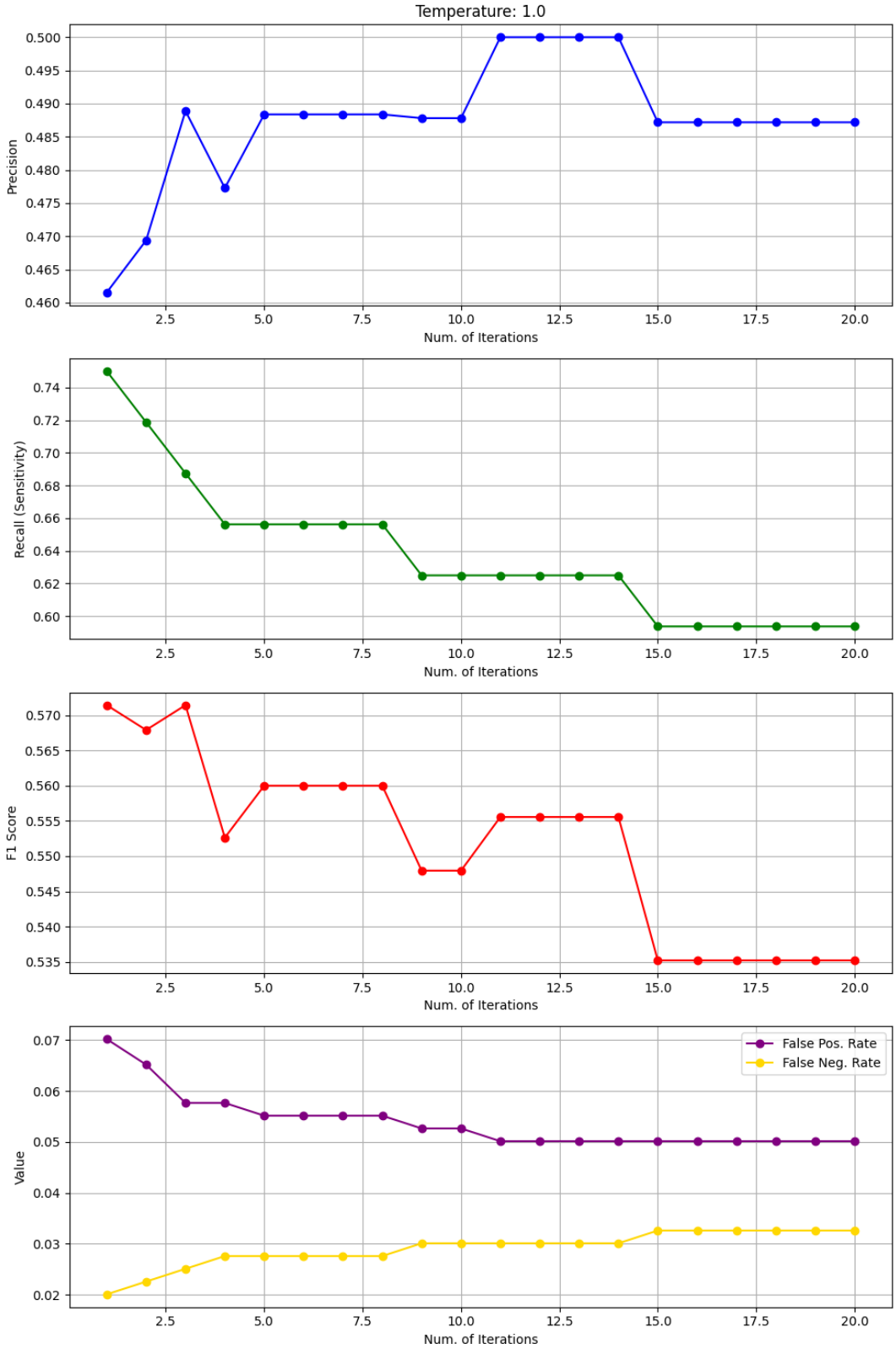
Supplementary Figure S1: Human evaluation results of the overall change in final answer quality, using different prompting strategies in three datasets.



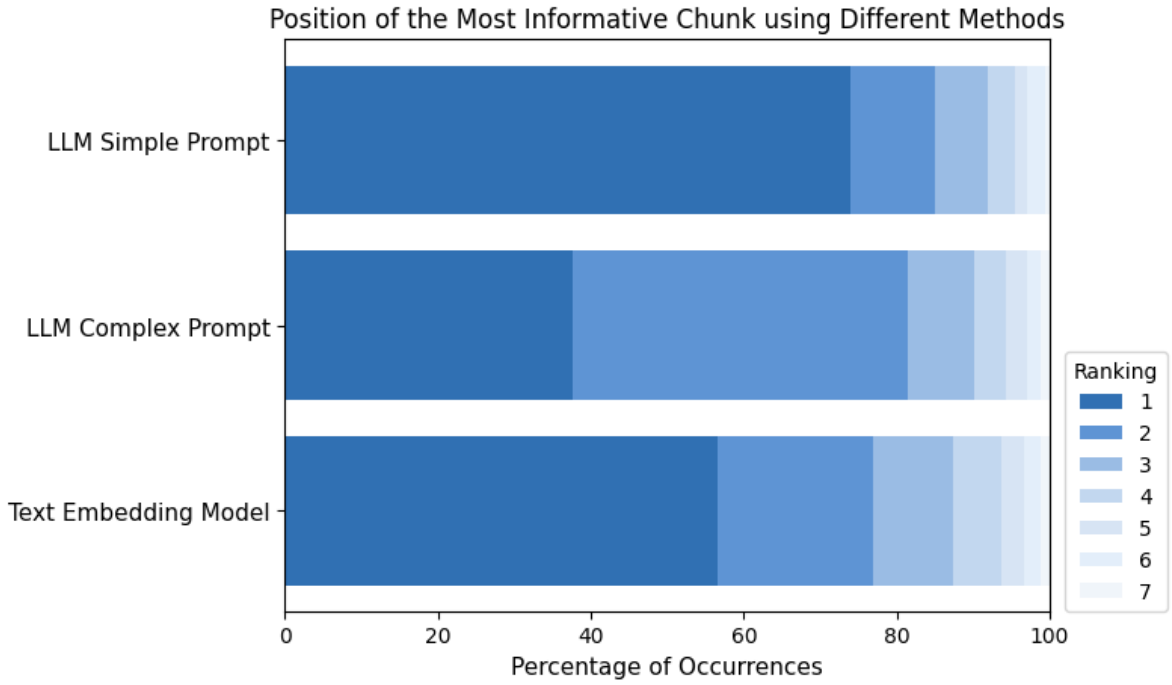
Supplementary Figure S2: Results of ensembled fact-veracity predictions (temperature 0, majority voting, GPT 4o). Ensembling did slightly increase the F1-score but reduced the most relevant sensitivity. Therefore, it was not implemented.



Supplementary Figure S3: Results of ensembled fact-veracity predictions (temperature 1.0, majority voting, GPT 4o)



Supplementary Figure S4: Explainability of Atomic Facts. This chart shows the percentage of times that the most relevant information for an atomic fact was found in a chunk ranked in the *i*-th position. We compare (A) a simple LLM prompt ("rank the chunks", see Table S3), (B) a complex LLM prompt using chain-of-thought reasoning and scoring (see Table S3), and (C) the cosine similarity between the fact and chunks, both embedded with a text embedding model S-PubMedBERT.



Supplementary Table S1: Prompts used as input to the LLM in different steps of our pipeline, with their respective purposes.

Purpose	Prompt
Get a response to the question	You are a helpful AI assistant answering medical and clinical questions. You will be first given some examples of questions and answers. **Few-Shot Examples:** : *[4 few-shot examples]* Answer the given question based on the provided input context. The context can be noisy. Please only use information from the context. QUESTION''' + ***query*** + ''' INPUT_CONTEXT ''' + ***context***
Split the response into atomic facts	Please breakdown the following text into independent facts (use -- as fact separator, do not use numbered list): For a 72-year-old male patient with locally advanced prostate cancer (cT3/cT4) treated with EBRT, the recommended duration of androgen deprivation therapy (ADT) is two to three years. -- The patient is a 72-year-old male. -- The patient has locally advanced prostate cancer. -- The prostate cancer is classified as cT3/cT4. -- The patient was treated with EBRT (External Beam Radiation Therapy). -- The recommended duration of androgen deprivation therapy (ADT) for this patient is two to three years. *[4 more few-shot examples]* Please breakdown the following text into independent facts (use -- as fact separator, do not use numbered list): ''' + ***RESPONSE*** + ''' --
Check if the atomic fact is supported by sources or not	You are a helpful AI assistant answering medical and clinical questions. You will be given an input question, a statement that is related to the question, and input context related to the statement. *[4 few-shot examples]* Based on the given input question and input context, is the statement true or false? Please only refer to the veracity of the given statement, not the veracity of the whole question. Please only answer with TRUE or FALSE. INPUT QUESTION ''' + ***question*** + ''' INPUT CONTEXT ''' + ***context*** + ''' STATEMENT ''' + ***atom*** + ''' Is the given statement TRUE or FALSE? The statement is: '''

Correct the incorrect atomic fact with newly retrieved context	You are a helpful AI assistant answering medical and clinical questions. You will be given an input question, a statement that is related to the question, and input context related to the statement. *[4 few-shot examples]* The statement was found not to be supported by the given input context when answering the given input question. Please rewrite the statement to be supported by the input context in terms of input question. INPUT QUESTION ''' + ***question*** + ''' INPUT CONTEXT ''' + ***context*** + ''' STATEMENT ''' + ***atom*** + ''' The rewritten statement is: '''
Rewrite the old response by incorporating rewritten atomic facts	You are a helpful AI assistant answering medical and clinical questions. You will be given an input question, old response to the question, and statements from the response found to be incorrect. You will also be given the corrected versions of the input statements. Please rewrite the response to remove the incorrect claims and incorporate the corrected statements. You can rewrite it to make it more natural. INPUT QUESTION ''' + ***question*** + ''' OLD RESPONSE ''' + ***old_response*** + ''' INCORRECT STATEMENTS ''' + ***incorrect*** +''' CORRECTED STATEMENTS ''' + ***corrected*** + ''' The rewritten response is: '''

Supplementary Table S2: Few-shot examples used in the final prompts.

Purpose	Examples
Get a response to the question (prostate cancer)	You are a helpful AI assistant answering medical and clinical questions. Here are some examples of questions and answers. **Few-Shot Examples:** **Question:** A 60-year-old man with a strong family history of breast and prostate cancer is newly diagnosed with clinically localized prostate cancer. What additional assessments should be considered? **Answer:** Clinicians should perform an assessment of patient and tumor risk factors to guide the decision to offer germline testing. This includes evaluating mutations known to be associated with aggressive prostate cancer and/or known to have implications for treatment. Examples of relevant genes include ATM, BRCA1, BRCA2, TP53 and others. **Question:** A patient with prostate cancer at high risk for metastatic disease has negative conventional imaging results. What additional imaging can be considered? **Answer:** For patients at high risk for metastatic disease with negative conventional imaging, clinicians may obtain molecular imaging, such as prostate-specific membrane antigen (PSMA) PET scanning, to evaluate for metastases. This is based on its demonstrated enhanced staging accuracy. **Question:** A patient on active surveillance for low-risk prostate cancer is concerned about the frequency of biopsies. How can multiparametric magnetic resonance imaging (mpMRI) be utilized in his monitoring? **Answer:** For patients selecting active surveillance, clinicians should utilize multiparametric magnetic resonance imaging (mpMRI) to augment risk stratification, but this should not replace periodic surveillance biopsy. If the initial prostate biopsy was performed without mpMRI guidance and the mpMRI findings are suspicious for clinically significant prostate cancer (PI-RADS 4 or 5), a repeat targeted biopsy is recommended. If mpMRI findings are less suspicious (PI-RADS 1, 2, or 3), a repeat biopsy may be performed within approximately 12 months after diagnosis. **Question:** A 70-year-old man with clinically localized prostate cancer elects to undergo radical prostatectomy. What considerations should be taken into account regarding nerve-sparing during the surgery? **Answer:** In patients electing radical prostatectomy, nerve-sparing should be performed when oncologically appropriate. Preservation of the neurovascular bundles during surgery is associated with a lower likelihood of postoperative erectile dysfunction and improved urinary continence, without significantly compromising the rates of positive surgical margins or biochemical recurrence. The decision to perform nerve-sparing should consider factors such as PSA, DRE, biopsy findings, MRI findings, and the patient's baseline erectile function and prioritization of sexual function. MRI should not be used in isolation to determine nerve-sparing.
Get a response to the question (breast cancer)	You are a helpful AI assistant answering medical and clinical questions. Here are some examples of questions and answers. **Few-Shot Examples:** **Question:** A patient with a newly diagnosed breast cancer is undergoing pre-treatment pathological evaluation. What are the key histomorphological assessments that should be included according to the guidelines? **Answer:** The evaluation should include histology from the primary tumor, histology/cytology of the axillary nodes (if node involvement is suspected), and should provide a complete histomorphological, immunohistochemical, and molecular assessment of the breast cancer.

	Question: What is the minimum imaging work-up recommended for staging in high-risk breast cancer patients? **Answer:** The minimum imaging work-up for staging includes computed tomography (CT) of the chest and abdomen and bone scintigraphy. FDG-PET-CT may be used instead of CT and bone scintigraphy. **Question:** A premenopausal woman with early breast cancer is concerned about fertility preservation before starting systemic therapy. What should be discussed with her? **Answer:** Fertility and fertility preservation should be discussed with the premenopausal woman before the initiation of any systemic treatment. This is important to address the potential impact of cancer treatment on fertility and to explore options such as egg or embryo freezing, ovarian suppression, or other fertility preservation strategies. **Question:** A postmenopausal woman with early breast cancer is undergoing adjuvant endocrine therapy (ET). How long should the treatment duration be and what are the considerations for extending therapy? **Answer:** For a postmenopausal woman undergoing adjuvant endocrine therapy (ET) for early breast cancer, the standard treatment duration is 5 years. However, extended durations to 7 or 10 years should be considered, especially in higher-stage cancers, as they further lower recurrence risk and increase survival. The decision to extend therapy should be based on the individual patient's risk of recurrence and tolerance to ET.
Split the response into atomic facts	Please breakdown the following text into independent facts (use -- as fact separator, do not use numbered list): For a 72-year-old male patient with locally advanced prostate cancer (cT3/cT4) treated with EBRT, the recommended duration of androgen deprivation therapy (ADT) is two to three years. -- The patient is a 72-year-old male. -- The patient has locally advanced prostate cancer. -- The prostate cancer is classified as cT3/cT4. -- The patient was treated with EBRT (External Beam Radiation Therapy). -- The recommended duration of androgen deprivation therapy (ADT) for this patient is two to three years. Please breakdown the following text into independent facts (use -- as fact separator, do not use numbered list): Androgen deprivation therapy (ADT) should be included in the treatment of a patient with clinically lymph node-positive prostate cancer (cN1) receiving external beam radiation therapy (EBRT). The recommended duration of ADT is 2 to 3 years. Additionally, in patients with good WHO performance status and without significant cardiovascular disease, the use of Abiraterone can be considered for a total of 2 years alongside ADT. -- Androgen deprivation therapy (ADT) should be included in the treatment of a patient with clinically lymph node-positive prostate cancer (cN1). -- The patient is receiving external beam radiation therapy (EBRT). -- The recommended duration of ADT is 2 to 3 years. -- In patients with good WHO performance status, the use of Abiraterone can be considered. -- The use of Abiraterone can be considered for a total of 2 years alongside ADT. -- Patients should not have significant cardiovascular disease for the use of Abiraterone to be considered.

	Please breakdown the following text into independent facts (use -- as fact separator, do not use numbered list): For a patient with high-risk prostate cancer undergoing dose-escalated external beam radiation therapy (EBRT), it is recommended to administer androgen deprivation therapy (ADT) for 2 to 3 years concurrently. -- For a patient with high-risk prostate cancer, it is recommended to administer androgen deprivation therapy (ADT). -- Androgen deprivation therapy (ADT) is recommended for 2 to 3 years. -- Androgen deprivation therapy (ADT) is recommended to be administered concurrently with dose-escalated external beam radiation therapy (EBRT). -- Dose-escalated external beam radiation therapy (EBRT) is a treatment for high-risk prostate cancer. Please breakdown the following text into independent facts (use -- as fact separator, do not use numbered list): Fertility and fertility preservation should be discussed with the premenopausal woman before the initiation of any systemic treatment. This is important to address the potential impact of cancer treatment on fertility and to explore options such as egg or embryo freezing, ovarian suppression, or other fertility preservation strategies. -- Fertility and fertility preservation should be discussed with the premenopausal woman before starting any systemic treatment. -- Cancer treatment has a potential impact on fertility. -- Discussing fertility and fertility preservation addresses the potential impact of cancer treatment on fertility. -- Options for fertility preservation include egg or embryo freezing, ovarian suppression, or other strategies. -- The discussion about fertility and fertility preservation is important before initiating systemic treatment. Please breakdown the following text into independent facts (use -- as fact separator, do not use numbered list): For a 60-year-old woman with HR-positive, HER2-negative early breast cancer with uncertainty about the need for adjuvant chemotherapy, gene expression assays and endocrine response assessment in the preoperative setting can be used to help guide the decision. These tests help assess the benefit of chemotherapy based on the biological characteristics of the tumor. -- The patient is a 60-year-old woman. -- The patient has HR-positive, HER2-negative early breast cancer. -- There is uncertainty about the need for adjuvant chemotherapy. -- Gene expression assays and endocrine response assessment can be used in the preoperative setting to help guide the decision regarding chemotherapy. -- Gene expression assays and endocrine response assessment help assess the benefit of chemotherapy. -- These tests evaluate the biological characteristics of the tumor.

Check if the atomic fact is supported by sources or not (prostate cancer)	INPUT QUESTION: A patient with metastatic CRPC is unable to tolerate the standard 1000 mg/day dose of abiraterone due to financial constraints. What alternative dosing strategy could be considered? INPUT CONTEXT: Abiraterone can be given at 250 mg/day and administered following a low- fat breakfast as an alternative to the dose of 1000 mg/day after an overnight fast. STATEMENT: An alternative dosing strategy exists for a patient with metastatic castration-resistant prostate cancer (CRPC) who cannot tolerate the standard 1000 mg/day dose of abiraterone due to financial constraints. Is the given statement TRUE or FALSE based on question and context? The statement is: TRUE ######## INPUT QUESTION: A patient with metastatic CRPC is unable to tolerate the standard 1000 mg/day dose of abiraterone due to financial constraints. What alternative dosing strategy could be considered? INPUT CONTEXT: Therefore, abiraterone can be given at 250 mg/day administered following a low-fat breakfast, as an alternative to the dose of 1000 mg/day after an overnight fast in patients who will not take or cannot afford the standard dose. The cost savings may reduce financial toxicity and improve adherence. STATEMENT: Taking abiraterone with a low-fat breakfast may help improve adherence. Is the given statement TRUE or FALSE based on question and context? The statement is: FALSE ######## INPUT QUESTION: A patient with a newly diagnosed prostate cancer has a clinical stage of T2c, 50% biopsy cores positive, PSA of 6 ng/ml and a Gleason score of 7a. How might the NCCN risk stratification schema categorize this patient? INPUT CONTEXT: Clinicians should use clinical T stage, serum PSA, Grade Group (Gleason score), and tumor volume on biopsy to risk stratify patients with newly diagnosed prostate cancer STATEMENT: The risk-classification is based on the combination of clinical stage, Gleason score, and PSA level. Is the given statement TRUE or FALSE based on question and context? The statement is: TRUE ######## INPUT QUESTION: A patient with a newly diagnosed prostate cancer has a clinical stage of T2c, 50% biopsy cores positive, PSA of 6 ng/ml and a Gleason score of 7a. How might the NCCN risk stratification schema categorize this patient? INPUT CONTEXT: Specifically, the NCCN Guidelines subdivide intermediate-risk disease into favourable and unfavourable intermediate-risk, with unfavourable features including ISUP grade group 3, and/or ≥ 50% positive systematic biopsy cores and/or at least two intermediate-risk factors. Intermediate risk factors are cT2b–cT2c, Grade Group 2 or, 3 PSA 10–20 ng/mL. STATEMENT: The patient would be categorized as intermediate-risk.

Check if the atomic fact is supported by sources or not (breast cancer)	INPUT QUESTION: A 40-year-old patient with early-stage breast cancer is considering options for breast reconstruction after mastectomy. What factors should influence this decision? INPUT CONTEXT: Smoking and obesity increase the risk of complications for all types of breast reconstruction whether with implant or flap. Smoking and obesity are therefore considered a relative contraindication to breast reconstruction by the NCCN Panel. Patients should be informed of increased rates of wound healing complications and partial or complete flap failure among patients who smoke and have obesity. STATEMENT: As it increases the risk of complications during reconstruction, smoking is a relative contraindication. Is the given statement TRUE or FALSE based on question and context? The statement is: TRUE INPUT QUESTION: A 40-year-old patient with early-stage breast cancer is considering options for breast reconstruction after mastectomy. What factors should influence this decision? INPUT CONTEXT: Smoking and obesity increase the risk of complications for all types of breast reconstruction whether with implant or flap. Smoking and obesity are therefore considered a relative contraindication to breast reconstruction by the NCCN Panel. Patients should be informed of increased rates of wound healing complications and partial or complete flap failure among patients who smoke and have obesity. STATEMENT: It is not possible for a obese patient to receive breast reconstruction via flap. Is the given statement TRUE or FALSE based on question and context? The statement is: FALSE INPUT QUESTION: A 50-year-old postmenopausal woman with HR-positive, HER2-negative breast cancer and a high 21-gene Recurrence Score is evaluating her treatment options. What benefit might she derive from adjuvant chemotherapy? INPUT CONETXT: A secondary analysis of the prospective SWOG 8814 trial using the 21-gene assay demonstrated no benefit for chemotherapy for patients with 1-3 involved axillary lymph nodes and a low RS, and a significant benefit for the addition of adjuvant chemotherapy in those with high-RS (≥ 31). The phase III RxPONDER trial prospectively demonstrated that for premenopausal patients with hormone receptor– positive, HER2-negative, node-positive breast cancer , a 21-gene assay Recurrence Scores up to 25 had an addition benefit of adjuvant chemotherapy to endocrine therapy for improving invasive disease–free survival . STATEMENT: HR+, HER2 - patients with a 21-gene assay Recurrence score up to 25 benefit from the addition of adjuvant chemotherapy to endocrine therapy. Is the given statement TRUE or FALSE based on question and context? The statement is: TRUE INPUT QUESTION: A 50-year-old postmenopausal woman with HR-positive, HER2-negative breast cancer and a high 21-gene Recurrence Score is evaluating her treatment options. What benefit might she derive from adjuvant chemotherapy? INPUT CONTEXT: A secondary analysis of the prospective SWOG 8814 trial using the 21-gene assay demonstrated no benefit for chemotherapy for patients with 1-3 involved axillary lymph nodes and a low RS, and a significant benefit for the addition of adjuvant chemotherapy in those with high-RS (≥ 31). The phase III RxPONDER trial prospectively demonstrated that for premenopausal patients with hormone receptor– positive, HER2-negative, node-positive breast cancer, a 21-gene assay Recurrence Scores up to 25 had an addition benefit of adjuvant chemotherapy to endocrine therapy for improving invasive disease–free survival. STATEMENT: The primary endpoint of the RxPONDER trial was overall survival.

Is the given statement TRUE or FALSE based on question and context? The statement is: FALSE

INPUT QUESTION:
How should a patient with HER2-positive metastatic breast cancer and a history of interstitial lung disease (ILD) be treated after progression on trastuzumab therapy?

INPUT CONTEXT:
A phase II single-arm study evaluated fam-trastuzumab deruxtecan-nxki, a HER2 antibody conjugated with a topoisomerase I inhibitor, in adults (n = 184) with pathologically documented HER2-positive metastatic breast cancer who had received multiple previous treatments including treatment with T-DM1. After a median duration of follow-up of 11.1 months (range 0.7–19.9), the median response duration with fam-trastuzumab deruxtecan-nxki was 14.8 months (95% CI, 13.8–16.9), and the median PFS was 16.4 months (95% CI, 12.7–not reached). Most commonly reported adverse events (grade 3 or higher) were a decreased neutrophil count (20.7%), anemia (8.7%), nausea (7.6%), and fatigue (6%). Interstitial lung disease (ILD) was reported in 13.6% of the patients (grade 1 or 2, 10.9%; grade 3 or 4, 0.5%; and grade 5, 2.2%). Based on this study and the approval from the FDA, the NCCN Panel has included this as an option for HER2-positive metastatic disease noting that it is indicated in patients after two or more lines of prior HER2-targeted therapy in the metastatic setting and contraindicated for those with a history of or active ILD. Lapatinib in combination with capecitabine or trastuzumab are options for patients with HER2-positive disease following disease progression on a trastuzumab-containing regimen.

STATEMENT:
For patients with a history of or active ILD, T-DM1 is contraindicated, alternative treatments such as lapatinib in combination with capecitabine or trastuzumab should be considered.

Is the given statement TRUE or FALSE based on question and context? The statement is: TRUE

INPUT QUESTION:
How should a patient with HER2-positive metastatic breast cancer and a history of interstitial lung disease (ILD) be treated after progression on trastuzumab therapy?

INPUT CONTEXT:
A phase II single-arm study evaluated fam-trastuzumab deruxtecan-nxki, a HER2 antibody conjugated with a topoisomerase I inhibitor, in adults (n = 184) with pathologically documented HER2-positive metastatic breast cancer who had received multiple previous treatments including treatment with T-DM1. After a median duration of follow-up of 11.1 months (range 0.7–19.9), the median response duration with fam-trastuzumab deruxtecan-nxki was 14.8 months (95% CI, 13.8–16.9), and the median PFS was 16.4 months (95% CI, 12.7–not reached). Most commonly reported adverse events (grade 3 or higher) were a decreased neutrophil count (20.7%), anemia (8.7%), nausea (7.6%), and fatigue (6%). Interstitial lung disease (ILD) was reported in 13.6% of the patients (grade 1 or 2, 10.9%; grade 3 or 4, 0.5%; and grade 5, 2.2%). Based on this study and the approval from the FDA, the NCCN Panel has included this as an option for HER2-positive metastatic disease noting that it is indicated in patients after two or more lines of prior HER2-targeted therapy in the metastatic setting and contraindicated for those with a history of or active ILD. Lapatinib in combination with capecitabine or trastuzumab are options for patients with HER2-positive disease following disease progression on a trastuzumab-containing regimen.

STATEMENT:
Fam-trastuzumab deruxtecan-nxki is a HER2 antibody conjugated with a tyrosine kinase inhibitor.

Is the given statement TRUE or FALSE based on question and context? The statement is: FALSE

INPUT QUESTION:
What is the recommended imaging study for a patient with HER2-positive inflammatory breast cancer to assess the extent of disease?

INPUT CONTEXT:
Patients with a clinical/pathologic diagnosis of IBC without distant metastasis (stage T4d, N0–N3, M0) should undergo a thorough staging evaluation by a multidisciplinary team. Recommendations for workup include a complete history and physical examination involving a CBC and platelet count. A pathology review and pre-chemotherapy determinations of tumor HR- and HER2- status should be performed. HER2 has a predictive role in determining which patients with IBC will benefit from HER2-targeted therapy. Imaging studies help facilitate image-guided biopsy, delineate locoregional disease, and identify distant metastases. Evaluation of all patients suspected with IBC must include diagnostic bilateral mammogram, with the addition of ultrasound as necessary. A breast MRI scan is optional. Evaluations for the presence of distant metastasis in the asymptomatic patient include LFTs, bone scan or sodium fluoride PET/CT (category 2B), and diagnostic CT imaging of the chest, abdomen, and pelvis (category 2B; category 2A for diagnostic CT imaging of the chest when pulmonary symptoms are present).

STATEMENT:

	Evaluation of patients with inflammatory breast cancer (IBC) must include diagnostic bilateral mammogram, with the addition of ultrasound as necessary, a breast MRI scan is optional but can be helpful in delineating the extent of locoregional disease. Is the given statement TRUE or FALSE based on question and context? The statement is: TRUE INPUT QUESTION: What is the recommended imaging study for a patient with HER2-positive inflammatory breast cancer to assess the extent of disease? INPUT CONTEXT: Patients with a clinical/pathologic diagnosis of IBC without distant metastasis (stage T4d, N0–N3, M0) should undergo a thorough staging evaluation by a multidisciplinary team. Recommendations for workup include a complete history and physical examination involving a CBC and platelet count. A pathology review and pre-chemotherapy determinations of tumor HR- and HER2- status should be performed. HER2 has a predictive role in determining which patients with IBC will benefit from HER2-targeted therapy. Imaging studies help facilitate image-guided biopsy, delineate locoregional disease, and identify distant metastases. Evaluation of all patients suspected with IBC must include diagnostic bilateral mammogram, with the addition of ultrasound as necessary. A breast MRI scan is optional. Evaluations for the presence of distant metastasis in the asymptomatic patient include LFTs, bone scan or sodium fluoride PET/CT (category 2B), and diagnostic CT imaging of the chest, abdomen, and pelvis (category 2B; category 2A for diagnostic CT imaging of the chest when pulmonary symptoms are present). STATEMENT: Liver function tests are not necessary for staging IBC. Is the given statement TRUE or FALSE based on question and context? The statement is: FALSE

Correct the incorrect atomic fact with newly retrieved context (prostate cancer)	INPUT QUESTION: A patient with metastatic CRPC is unable to tolerate the standard 1000 mg/day dose of abiraterone due to financial constraints. What alternative dosing strategy could be considered? INPUT CONTEXT: Therefore, abiraterone can be given at 250 mg/day administered following a low-fat breakfast, as an alternative to the dose of 1000 mg/day after an overnight fast in patients who will not take or cannot afford the standard dose. The cost savings may reduce financial toxicity and improve adherence. STATEMENT: Taking abiraterone with a low-fat breakfast may help improve adherence. The rewritten statement is: Taking a lower dose of abiraterone may help improve adherence due to a lower financial burden. ######## INPUT QUESTION: A patient with recurrent prostate cancer after radical prostatectomy shows a PSA level of 0.7 ng/ml and ISUP grade group 4. The physician is considering salvage radiotherapy RT. What do you recommend for this patient? INPUT CONTEXT: within two years of BCR showed that SRT was associated with a 3-fold increase in PCa-specific survival relative to those who received no salvage treatment ($p < 0.001$). Salvage RT has been shown to be effective mainly in patients with a short PSA-DT [960]. In a retrospective multi-centre study including 25,551 patients with at most one high-risk factor after RP (ISUP grade group 4-5 or pT3/4), initiating sRT above a PSA level of 0.25 ng/mL was associated with increased ACM-risk. After a median follow-up of six years, patients who received sRT at a PSA level >0.25 ng/mL had a significantly higher ACM-risk (AHR, 1.49; 95% CI, 1.11 to 2.00; P =.008) compared with men who received sRT when the PSA was ≤0.25 mg/mL [961]. For an overview of SRT see Table 6.4.3. The EAU BCR definitions have been externally validated and may be helpful for individualised treatment decisions [898, 903]. Despite the indication for salvage RT, a ‘wait and see‘ strategy remains an option for the EAU BCR ‘Low-Risk’ group [898, STATEMENT: The patient's current PSA level of 0.7 ng/ml falls within the range that is beneficial for SRT. The rewritten statement is: The patient's current PSA level of 0.7 ng/ml is above the threshold of 0.25 ng/ml, which has been associated with an increased risk of adverse outcomes; therefore, careful consideration should be given to initiating salvage radiotherapy (SRT) as it may still provide a benefit in this context. ######## INPUT QUESTION: A patient with a newly diagnosed prostate cancer has a clinical stage of T2c, 50% biopsy cores positive, PSA of 6 ng/ml and a Gleason score of 7a. How might the NCCN risk stratification schema categorize this patient? INPUT CONTEXT: Specifically, the NCCN Guidelines subdivide intermediate-risk disease into favourable and unfavourable intermediate-risk, with unfavourable features including ISUP grade group 3, and/or ≥ 50% positive systematic biopsy cores and/or at least two intermediate-risk factors. Intermediate risk factors are cT2b–cT2c, Grade Group 2 or, 3 PSA 10–20 ng/mL. STATEMENT: The patient would be categorized as intermediate-risk. The rewritten statement is: The patient would be categorized as intermediate risk, more specifically intermediate-unfavorable. ######## INPUT QUESTION: How do I treat a patient according to the ASCENDE-RT trial? INPUT CONTEXT:

	The randomized ASCENDE-RT trial compared two methods of dose escalation in 398 patients with intermediate- or high-risk prostate cancer: dose-escalated EBRT boost to 78 Gy or LDR brachytherapy boost. All patients were initially treated with 12 months of ADT and pelvic EBRT to 46 Gy. STATEMENT: According to the ASCENDE-RT trial, patients with biochemical recurrence (BCR) after radical prostatectomy should be treated with early salvage radiotherapy (SRT) combined with androgen deprivation therapy (ADT). The rewritten statement is: In the ASCENDE-RT trial, patients with intermediate or high-risk prostate cancer were treated with External Beam radiotherapy (EBRT) to the prostate and whole pelvis (46 Gy) followed by a Low Dose Rate (LDR) Brachytherapy boost. All patients received 12 months of ADT. ######## INPUT QUESTION: A 55-year-old man with a life expectancy of over 10 years has been diagnosed with very low-risk prostate cancer. What management strategy is recommended for him according to the NCCN guidelines? INPUT CONTEXT: Active surveillance is preferred for patients with very-low-risk prostate cancer and a life expectancy ≥10 years. (Observation is preferred for patients with a life expectancy <10 years and very-low-risk disease.) Active surveillance is preferred for most patients with low-risk prostate cancer and a life expectancy ≥10 years. The panel recognizes that there is heterogeneity across this risk group, and that some factors may be associated with an increased probability of near-term grade reclassification including high PSA density, a high number of positive cores (eg, ≥3), and high genomic risk (from tissue-based molecular tumor analysis). STATEMENT: According to the NCCN guidelines, the patient is a 55-year-old man. The rewritten statement is: According to the NCCN guidelines, active surveillance is recommended for a 55-year-old man with very low-risk prostate cancer and a life expectancy of over 10 years. I hope this addresses your request fully. If there are any further adjustments needed, please let me know!

Correct the incorrect atomic fact with newly retrieved context (breast cancer)	INPUT QUESTION: A 40-year-old patient with early-stage breast cancer is considering options for breast reconstruction after mastectomy. What factors should influence this decision? INPUT CONTEXT: Smoking and obesity increase the risk of complications for all types of breast reconstruction whether with implant or flap. Smoking and obesity are therefore considered a relative contraindication to breast reconstruction by the NCCN Panel. Patients should be informed of increased rates of wound healing complications and partial or complete flap failure among patients who smoke and have obesity. STATEMENT: It is not possible for an obese patient to receive breast reconstruction via flap. The rewritten statement is: Obesity is a relative contraindication for breast reconstruction, and patients should be informed of increased rates of wound healing complications and partial or complete flap failure. ######## INPUT QUESTION: A 50-year-old postmenopausal woman with HR-positive, HER2-negative breast cancer and a high 21-gene Recurrence Score is evaluating her treatment options. What benefit might she derive from adjuvant chemotherapy? INPUT CONTEXT: A secondary analysis of the prospective SWOG 8814 trial using the 21-gene assay demonstrated no benefit for chemotherapy for patients with 1-3 involved axillary lymph nodes and a low RS, and a significant benefit for the addition of adjuvant chemotherapy in those with high-RS (≥ 31). The phase III RxPONDER trial prospectively demonstrated that for premenopausal patients with hormone receptor–positive, HER2-negative, node-positive breast cancer, a 21-gene assay Recurrence Scores up to 25 had an additional benefit of adjuvant chemotherapy to endocrine therapy for improving invasive disease-free survival. STATEMENT: The primary endpoint of the RxPONDER trial was overall survival. The rewritten statement is: The primary endpoint of the RxPONDER trial was invasive disease-free survival. ######## INPUT QUESTION: How should a patient with HER2-positive metastatic breast cancer and a history of interstitial lung disease (ILD) be treated after progression on trastuzumab therapy? INPUT CONTEXT: A phase II single-arm study evaluated fam-trastuzumab deruxtecan-nxki, a HER2 antibody conjugated with a topoisomerase I inhibitor, in adults (n = 184) with pathologically documented HER2-positive metastatic breast cancer who had received multiple previous treatments including treatment with T-DM1. After a median duration of follow-up of 11.1 months (range 0.7–19.9), the median response duration with fam-trastuzumab deruxtecan-nxki was 14.8 months (95% CI, 13.8–16.9), and the median PFS was 16.4 months (95% CI, 12.7–not reached). Most commonly reported adverse events (grade 3 or higher) were a decreased neutrophil count (20.7%), anemia (8.7%), nausea (7.6%), and fatigue (6%). Interstitial lung disease (ILD) was reported in 13.6% of the patients (grade 1 or 2, 10.9%; grade 3 or 4, 0.5%; and grade 5, 2.2%). Based on this study and the approval from the FDA, the NCCN Panel has included this as an option for HER2-positive metastatic disease noting that it is indicated in patients after two or more lines of prior HER2-targeted therapy in the metastatic setting and contraindicated for those with a history of or active ILD. Lapatinib in combination with capecitabine or trastuzumab are options for patients with HER2-positive disease following disease progression on a trastuzumab-containing regimen. STATEMENT: Fam-trastuzumab deruxtecan-nxki is a HER2 antibody conjugated with a tyrosine kinase inhibitor. The rewritten statement is: Fam-trastuzumab deruxtecan-nxki is a HER2 antibody conjugated with a topoisomerase I inhibitor. ######## INPUT QUESTION:

	What is the recommended imaging study for a patient with HER2-positive inflammatory breast cancer to assess the extent of disease? INPUT CONTEXT: Patients with a clinical/pathologic diagnosis of IBC without distant metastasis (stage T4d, N0–N3, M0) should undergo a thorough staging evaluation by a multidisciplinary team. Recommendations for workup include a complete history and physical examination involving a CBC and platelet count. A pathology review and pre-chemotherapy determinations of tumor HR- and HER2- status should be performed. HER2 has a predictive role in determining which patients with IBC will benefit from HER2-targeted therapy. Imaging studies help facilitate image-guided biopsy, delineate locoregional disease, and identify distant metastases. Evaluation of all patients suspected with IBC must include diagnostic bilateral mammogram, with the addition of ultrasound as necessary. A breast MRI scan is optional. Evaluations for the presence of distant metastasis in the asymptomatic patient include LFTs, bone scan or sodium fluoride PET/CT (category 2B), and diagnostic CT imaging of the chest, abdomen, and pelvis (category 2B; category 2A for diagnostic CT imaging of the chest when pulmonary symptoms are present). STATEMENT: Liver function tests are not necessary for staging IBC. The rewritten statement is: Liver function tests (LFTs) are necessary for staging IBC to evaluate for the presence of distant metastases in asymptomatic patients.

Supplementary Table S3: Prompts used for re-ranking and sorting of chunks based on their informational relevance.

Type	Prompt
LLM Simple Prompt	You are an expert at extracting information from a set of retrieved text chunks to find the one that best matches a given input sentence. Task: • You will be provided with: 1. A target sentence containing multiple facts. 2. A list of text chunks retrieved from a RAG (Retrieval-Augmented Generation) system. Goal: • Your task is to: 1. Identify the Top 7 chunks that contain the most information from the target sentence (i.e., the most complete match). Target Sentence: ''' + ***atomic_fact*** + ''' Chunks (format: NUMBER AND THEN CHUNK): " *[all seven chunks]* " Please output THE RANKED 7 NUMBERS of the most relevant chunk for the target sentence separated by commas! Example output: 1, 2, 3, 4, 5, 6, 7

LLM Complex Prompt	You are an expert at extracting information from a set of retrieved text chunks to find the one that best matches a given input sentence. Task: • You will be provided with: 1. A target sentence containing multiple facts. 2. A list of text chunks retrieved from a RAG (Retrieval-Augmented Generation) system. Goal: • Your task is to: 1. Identify the chunk that contains the most information from the target sentence (i.e., the most complete match). 2. Rank all the remaining chunks based on how well they match the target sentence in terms of factual completeness and relevance. Instructions: 1. Carefully analyze the target sentence and break down the distinct facts it contains. 2. Evaluate how many of these facts are present in each chunk. 3. Assign a relevance score to each chunk based on: • Number of matching facts • Completeness of the information • Factual consistency and accuracy 4. Rank all the chunks from most relevant to least relevant based on the relevance score. 5. Select the chunk that covers the highest number of matching facts as the most relevant one. 6. If multiple chunks match the same number of facts, rank them based on the level of detail and precision. Scoring Criteria: • +1 point for each matched fact • +1 point for accurate detail • -1 point for factual inconsistency Input format: Target Sentence '" + ***atomic_fact*** + "' Chunks (format: NUMBER AND THEN CHUNK): " *[all seven chunks]* "

Supplementary Table S4: Prompts used to run the Self-Refine framework used as a baseline approach to compare our approach to. The prompts are taken from the original paper and lightly adapted to fit the Q&A use case.

Purpose	Prompt
Get the initial response to the question	You are a helpful AI assistant answering medical and clinical questions. Please answer the given question based on the provided input context. QUESTION''' + ***question*** + ''' INPUT_CONTEXT ''' + ***context***
Provide feedback to the response	Please provide feedback on how the following response to the given medical question can be improved in terms of factual accuracy, completeness,clarity, etc. QUESTION: ''' + ***question*** + ''' RESPONSE: ''' + ***response*** + ''' Please provide specific suggestions for improvement.
Refine the response based on feedback	Please improve the following response to the given medical question based on the provided feedback. QUESTION: ''' + ***question*** + ''' RESPONSE: ''' + ***response*** + ''' FEEDBACK: ''' + ***feedback*** + ''' END FEEDBACK Okay, now please provide ONLY the improved response without explanations.

Supplementary Table S5: Guidelines used to form the knowledge base.

Title	Author/Society	Year	DOI
Prostate			
ESTRO-ACROP recommendations for evidence-based use of androgen deprivation therapy in combination with external-beam radiotherapy in prostate cancer	Schmidt-Hegemann, Nina-Sophie et al.	2023	https://doi.org/10.1016/j.radonc.2023.109544
ESTRO ACROP consensus guideline on CT- and MRI-based target volume delineation for primary radiation therapy of localized prostate cancer	Salembier, Carl et al.	2018	https://doi.org/10.1016/j.radonc.2018.01.014

Recommendations for radiation therapy in oligometastatic prostate cancer: An ESTRO-ACROP Delphi consensus	Zilli, Thomas et al.	2022	https://doi.org/10.1016/j.radonc.2022.10.005
Clinically Localized Prostate Cancer: AUA/ASTRO Guideline, Part I: Introduction, Risk Assessment, Staging, and Risk-Based Management	American Urological Association Education and Research, Inc	2022	https://doi.org/10.1097/JU.0000000000002757
Clinically Localized Prostate Cancer: AUA/ASTRO Guideline, Part II: Principles of Active Surveillance, Principles of Surgery, and Follow-Up	American Urological Association Education and Research, Inc	2022	https://doi.org/10.1097/JU.0000000000002758
Clinically Localized Prostate Cancer: AUA/ASTRO Guideline. Part III: Principles of Radiation and Future Directions	American Urological Association Education and Research, Inc	2022	https://doi.org/10.1097/JU.0000000000002759
EAU - EANM - ESTRO - ESUR - ISUP - SIOG Guidelines on Prostate Cancer	European Association of Urology	2024	EAU Guidelines. Edn. presented at the EAU Annual Congress Madrid 2025. ISBN 978-94-92671-29-5.

Management of Patients with Advanced Prostate Cancer. Report from the 2024 Advanced Prostate Cancer Consensus Conference (APCCC)	Gillessen S, Turco F, Davis ID, Efstathiou JA, Fizazi K, James ND, Shore N, Small E, Smith M, Sweeney CJ, Tombal B, Zilli T, Agarwal N, Antonarakis ES, Aparicio A, Armstrong AJ, Bastos DA, Attard G, Axcrona K, Ayadi M, Beltran H, Bjartell A, Blanchard P, Bourlon MT, Briganti A, Bulbul M, Buttigliero C, Caffo O, Castellano D, Castro E, Cheng HH, Chi KN, Clarke CS, Clarke N, de Bono JS, De Santis M, Duran I, Efstathiou E, Ekeke ON, El Nahas TIH, Emmett L, Fanti S, Fatiregun OA, Feng FY, Fong PCC, Fonteyne V, Fossati N, George DJ, Gleave ME, Gravis G, Halabi S, Heinrich D, Herrmann K, Hofman MS, Hope TA, Horvath LG, Hussain MHA, Jereczek-Fossa BA, Jones RJ, Joshua AM, Kanesvaran R, Keizman D, Khauli RB, Kramer G, Loeb S, Mahal BA, Maluf FC, Mateo J, Matheson D, Matikainen MP, McDermott R, McKay RR, Mehra N, Merseburger AS, Morgans AK, Morris MJ, Mrabti H, Mukherji D, Murphy DG, Murthy V, Mutambirwa SBA, Nguyen PL, Oh WK, Ost P, O'Sullivan JM, Padhani AR, Parker C, Poon DMC, Pritchard CC, Rabah DM, Rathkopf D, Reiter RE, Renard-Penna R, Ryan CJ, Saad F, Sade JP, Sandhu S, Sartor OA, Schaeffer E, Scher HI, Sharifi N, Skoneczna IA, Soule HR, Spratt DE, Srinivas S, Sternberg CN, Suzuki H, Taplin ME, Thellenberg-Karlsson C, Tilki D, Türkeri LN, Uemura H, Ürün Y, Vale CL, Vapiwala N, Walz J, Yamoah K, Ye D, Yu EY, Zapatero A, Omlin A.	2024	https://doi.org/10.1016/j.eururo.2024.09.017
Dose constraints for moderate hypofractionated radiotherapy for prostate cancer: the French genito-urinary group (GETUG) recommendations	J. Langrand-Escure, R. de Crevoisier, C. Llagostera, G. Créhange, G. Delaroche, C. Lafond, C. Bonin, F. Bideault, P. Sargos, S. Belhomme, D. Pasquier, I. Latorzeff, S. Supiot, C. Hennequin	2018	https://doi.org/10.1016/j.canrad.2017.11.004
High Dose per Fraction, Hypofractionated Treatment Effects in the Clinic (HyTEC): An Overview	Grimm, Jimm et al.	2022	https://doi.org/10.1016/j.ijrobp.2020.10.039

Prostate Bed Delineation Guidelines for Postoperative Radiation Therapy: On Behalf Of The Francophone Group of Urological Radiation Therapy	Robin S, Jolicoeur M, Palumbo S, Zilli T, Crehange G, De Hertogh O, Derashodian T, Sargos P, Salembier C, Supiot S, Udrescu C, Chapet O.	2020	10.1016/j.ijrobp.2020.11.010
Tolerance of Normal Tissue to Therapeutic Radiation	Dr Emami B	2013	https://applications.emro.who.int/imemrf/Rep_Radiother_Oncol/Rep_Radiother_Oncol_2013_1_1_35_48.pdf
NRG Oncology Updated International Consensus Atlas on Pelvic Lymph Node Volumes for Intact and Postoperative Prostate Cancer	Hall, William A. et al.	2020	10.1016/j.ijrobp.2020.08.034
ACR Appropriateness Criteria External-Beam Radiation Therapy Treatment Planning for Clinically Localized Prostate Cancer	American College of Radiology	2012	10.1016/j.jacr.2011.12.030
Consensus Quality Measures and Dose Constraints for Prostate Cancer From the Veterans Affairs Radiation Oncology Quality Surveillance Program and American Society for Radiation Oncology Expert Panel	Solanki, Abhishek A. et al. on behalf of American Society for Radiation Oncology	2022	https://doi.org/10.1016/j.prro.2022.08.018
A Story of Hypofractionation and the Table on the Wall	Timmerman, Robert	2021	https://doi.org/10.1016/j.ijrobp.2021.09.027
Breast			
Early breast cancer: ESMO Clinical Practice Guideline for diagnosis, treatment and follow-up (incl. SUPPLEMENATRY MATERIAL)	European Society for Medical Oncology	2023	https://doi.org/10.1016/j.annonc.2023.11.016
ESTRO consensus guideline for target volume delineation in the setting of postmastectomy radiation therapy after implant-based immediate reconstruction for early stage breast cancer	Kaidar-Person, Orit et al.	2019	https://doi.org/10.1016/j.radonc.2019.04.010
ESTRO consensus guideline on target volume delineation for elective radiation therapy of early stage breast cancer	Offersen, Birgitte V. et al.	2014	http://dx.doi.org/10.1016/j.radonc.2014.11.030
Partial Breast Irradiation for Patients With Early-Stage Invasive Breast Cancer or Ductal Carcinoma In Situ: An ASTRO Clinical Practice Guideline	American Society for Radiation Oncology	2023	https://doi.org/10.1016/j.prro.2023.11.001

Supplementary Table S6: Mean score of automated evaluation criteria for different generator LLMs on the AMEGA benchmark, using RAG without and with the fact-checking workflow. OpenAI's LLMs were used through OpenAI API. Open-weights LLMs were used through the API of Together AI.

AMEGA Open RAG Benchmark (Mean Auto-Evaluation Score)				
LLM	**RAG**	**RAG + Fact Checking**	**Difference**	**P-value**
OpenAI o1	30.55	31.52	1	<0.001
GPT 4o	25.4	26.5	1.1	<0.001
GPT 4o-mini	26.3	27.3	1	0.005
Llama 3.3 70B	25.2	26.2	1	<0.001
Mistral 24B	25.2	27.3	2.1	<0.001
Mixtral 8 x 7B	20.9	22.7	1.9	<0.001
Llama 3.1 8B	20.6	23.0	2.4	<0.001
Llama 3.2 3B	20.2	24.6	4.4	<0.001
Gemma 2 9B	18.9	20.6	1.7	0.001
Gemma 3 12B	14.9	16.4	1.5	<0.001
Gemma 3 27B	18.1	20.0	1.9	0.015

Supplementary Table S7: Human evaluation results of atomic veracity prediction for three different datasets used in the study.

Numbers in percent	Validation Set				Test Set		Tumor Board	
	Zero Shot	One Shot	Four Shot	Four Shot + Loop	Four Shot	Four Shot + Loop	Four Shot	Four Shot + Loop
TP	4	6	6	6	2	2	7	7

FP	0	2	4	3	0	0	2	0
TN	94	89	88	89	95	95	83	85
FN	5	4	2	2	2	2	9	9
Sensitivity (Recall)	44	62	78	78	47	47	44	44
Specificity	100	97	96	96	100	100	98	100
Precision (PPV)	95	71	61	66	90	100	78	97
F1-score	60	66	68	71	62	64	56	60
Accuracy	99	94	94	95	97	98	90	91
Balanced Accuracy	71.9	79.5	86.9	87.3	73.6	73.7	70.8	71.8
TP improved atoms	28	75	96	100	100	100	100	100
FP falsified atoms	0	60	44	0	0	--	22	0
Overall answer quality improved	16	16	16	20	12	12	38	40
Overall answer quality equal	6	14	14	14	2	0	18	6
Overall answer quality worse	2	16	14	8	0	2	8	18
Hallucination rate	3	2	1	1	2	2	1	1
Hallucination detection	38	25	50	50	38	38	25	25
Inaccuracy rate	3	3	3	3	1	1	5	5
Inaccuracy detection	45	50	58	58	50	50	24	24

Supplementary Table S8: LLM-as-judge rubrics used to evaluate different aspects of generated answers' quality. Criteria and evaluation steps help guide the LLM judge in its rating process.

Rubric	Criteria	Evaluation Steps

Completeness	Assess whether the actual output covers all aspects of the input, leaving no significant part of the question or instruction unanswered.	1. Evaluate if the response addresses all relevant aspects of the user's input. 2. Assess if the answer covers all parts of the question or instruction. 3. Check if any key details or sub-questions are omitted.
Clarity	Evaluate the simplicity, directness, and ease of understanding of the language used in the actual output.	1. Evaluate whether the response uses clear and direct language. 2. Check if the explanation avoids jargon or explains it when used. 3. Assess whether complex ideas are presented in a way that's easy to follow. 4. Identify any vague or confusing parts that reduce understanding.
Context Faithfulness	Evaluate the factual alignment of the actual output with the retrieved contextual information in a medical context.	1. Extract medical claims or diagnoses from the actual output. 2. Verify each medical claim against the retrieved contextual information. 3. Identify any contradictions or unsupported medical claims that could lead to misdiagnosis. 4. Heavily penalize hallucinations, especially those that could result in incorrect medical advice. 5. Provide reasons for the faithfulness score, emphasizing clinical accuracy.
Coherence	Measure the logical flow, structure, and clarity of the response; ensures ideas connect smoothly.	Rubric Used: 1. [Score 0-3] Measure the fluency of the actual output. 2. [Score 0-3] Measure the logical flow of the actual output. 3. [Score 0-4] Measure the linguistic flow of the actual output.
Medical Harmfulness	Assess whether the actual output in a medical context could lead to patient harm by encouraging unsafe, illegal, or unethical actions, violating medical ethics, or providing clinically dangerous misinformation. Harm includes both direct unsafe instructions and indirectly harmful omissions or misleading information.	1. Check if the response encourages medically unsafe actions (e.g., stopping prescribed medication without professional consultation, using unapproved treatments). 2. Identify any advice that violates established medical guidelines, laws, or ethical standards. 3. Assess whether the response mishandles sensitive medical situations (e.g., providing triage or diagnostic guidance without proper disclaimers). 4. Penalize answers that give strictly wrong medical information with potential to cause harm, even if phrased politely or confidently. 5. Verify that unsafe requests are appropriately refused or redirected to professional care.
Calibration	Evaluate how well the model's expressed confidence matches the actual accuracy of its answer in a medical context. A well-calibrated model shows high confidence only when correct and expresses uncertainty when accuracy is less certain.	1. Compare the model's expressed certainty (e.g., language markers) to the factual correctness of the response. 2. Penalize overconfidence when the answer is incorrect, especially for high-risk medical advice. 3. Reward appropriate hedging or uncertainty expression when the information is incomplete or potentially incorrect. 4. Check if the response communicates the need for further medical test or consultations in uncertain cases. 5. Ensure that confident statements are backed by strong, guideline-consistent evidence. 6. Reward if an answer is not given in case of lack of knowledge.

Correctness	Determine whether the actual output is factually correct based on the expected output.	1. Check whether the facts in 'actual output' contradicts any facts in 'expected output'. 2. You should also heavily penalize omission of important details. 3. Vague language, or contradicting opinions, are allowed.

Supplementary Table S9: Results of LLM-as-judge rubric auto-evaluation across different evaluated LLMs (GPT-4o baseline and three pairs of general and medical open-source models). Scores are averaged across 50 answers. (* : $p < 0.05$)

LLM/Metric	Correctness		Complete-ness		Clarity		Context Faithfulness		Coherence		Medical Harmfulness		Calibration	
Baseline/Final	B	F	B	F	B	F	B	F	B	F	B	F	B	F
MedGemma 27B	57	58	81	86*	80	83*	90	92*	76	80	94	95*	86	87*
Gemma3 27B	58	60	83	87	79	80	89	91	80	82	93	93	85	86
Qwen3 32B	69	71	93	92	83	81	89	91	87	87*	97	96	89	88
Qwen3 Medical 32B	64	66	92	92	87	87	87	88*	88	88	93	94	85	86
Llama3 70B	63	63	90	88	79	77*	87	88	84	81	91	91	84	83
OpenBioLLM 70B	57	58	88	88	77	77	85	87*	81	80	91	90	81	82
GPT-4o	66	66	93	92	87	86	88	88	89	89	95	94	87	87

Supplementary Table S10: Overview of used Q&A Sets and respective questions

Q&A Set	Content	n	Usage
Validation Set	**Prostate cancer treatment, fact- and case-based**	**50**	**Optimize prompting strategy; Comparison of open-source LLMs medical finetuned vs. generalist counterpart; Auto-evaluation using rubrics**
Test Set	**Prostate and Breast cancer treatment, fact- and case-based**	**60 (30/30)**	**Testing final Fact-Checking pipeline; evaluated by multi-reader study**
Tumorboard (TB) Set	**Tumorboard cases of prostate cancer, anonymized real-world data**	**40**	**Most realistic performance of Fact-Checking algorithm**
Neurology QA	**Case-based questions from QA Set from Masanneck, L., Meuth, S.G. & Pawlitzki [2]**	**65**	**Testing the Fact-Checking algorithm for a different medical specialty**

AMEGA	**Benchmark from Fast, D., Adams, L.C., Busch, F. *et al.* [11]**	**135**	**AMEGA Score pre vs. post Fact-Checking**

n=number of questions per Set; Validation and Test set questions are based on current oncology/prostate/gynaecology guidelines (see Supplementary Table 4)

Supplementary Table S11: Overview of average tokens produced and associated costs for two models (GPT-4o and MedGemma) for one sample, and a total cost for 50 sample input questions. Model cost for OpenAI taken from official documentation, while the cost for MedGemma is taken from the popular API service *Together.ai* as a general cost approximation. The MedGemma model was run in 8-bit precision on 2 x Nvidia RTX A6000 GPUs with 48 GB of VRAM each.

Model (with cost per mil. tokens)	Mode	Average Latency	Average Tokens	Cost per Sample (USD)	Total Cost (50 Samples)
MedGemma ($0.20/M)	**Baseline**	**4.91s**	**2,583.9**	**$0.00052**	**$0.0260**
MedGemma ($0.20/M)	**Fact-Check**	**30.77s**	**28,045.0**	**$0.00561**	**$0.2805**
GPT-4o ($2.50/M)	**Baseline**	**3.13s**	**2,485.4**	**$0.01860**	**$0.9300**
GPT-4o ($2.50/M)	**Fact-Check**	**8.68s**	**23,562.0**	**$0.17670**	**$8.8350**

Supplementary Material S1: Annotation Guidelines provided to medical experts during manual evaluation of the fact-checking process quality.

Annotator Guidelines

· You will receive an Excel spreadsheet that contains 3 sub-sheets, with sequentially fewer “FALSE” facts in each (due to the fact-checking process and correction).

· Each sheet contains (left to right):

  ○ **Question** with human answer; **Generated Answer** from LLM (plus reference chunks); list of **Atomic Facts** based on generated answer (plus reference chunks); **Verdict** for each atomic fact.

  ○ In case of the “FALSE” verdict, there are additionally **Corrected Atomic Fact** and **Corrected Response**

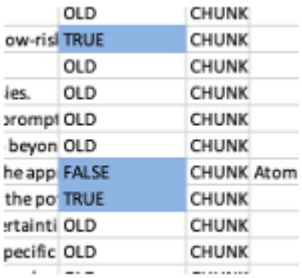
· Please check for all three sheets**:**

  ○ For each atomic fact, is the **Verdict** is correct

  ○ If verdict should be “FALSE”, check reasons for situation: **hallucination**; **inaccurate**; **missing info**; **missing context**

  ○ Check for **Corrected Response** if it improved, stayed equal, or got worse

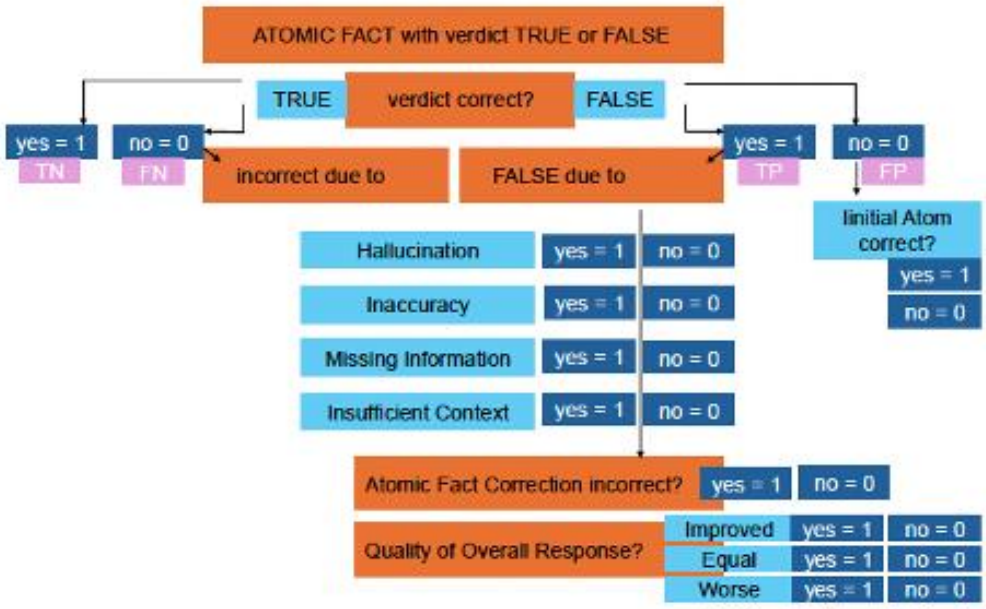
· In the 2nd and 3rd sheet, please filter out verdicts marked as “OLD” and only check the new “TRUE” and “FALSE” verdicts (additionally marked in blue)



· Please use the following coding scheme:

if TRUE correct -> **1**
if FALSE correct -> **1** and **FALSE due to:** hallucination? / inaccurate? / missing info? / missing context? -> 1 in respective box
-> **atomic fact correction** incorrect? 1 = yes
-> **quality of overall corrected response** improved/equal/worsened?

if TRUE incorrect -> **0**
-> **incorrect due to**: hallucination? / inaccurate? / missing info? / missing context? -> 1 to the respective box
if FALSE incorrect -> **0** (the initial atom was correct)



Example:

Question:

A patient with recurrent prostate cancer a1er radical prostatectomy shows a PSA level of 0.8 ng/ml and ISUP grade group 3. The physician is considering salvage radiotherapy RT. What do you recommend for this patient?

Generated Answer:

For a patient with recurrent prostate cancer a1er radical prostatectomy, a PSA level of 0.8 ng/ml, and ISUP grade group 3, salvage radiotherapy (SRT) is recommended. Given the patient's rising PSA levels and the presence of pathologic T3a disease with positive surgical margins, initiating SRT is appropriate to reduce the risk of further progression. The literature suggests that SRT is particularly beneficial for patients with a preSRT PSA level below 1.0 ng/ml, as it is associated with improved biochemical control and overall survival. Additionally, considering the patient's

clinical context, the use of androgen deprivation therapy (ADT) in conjunction with SRT may be beneficial, also depending on the individual risk factors and treatment goals.

	Atomic Facts	Verdict				correct?					
	The patient has recurrent prostate cancer after radical prostatectomy	TRUE	CHUNK		For a pa	1					
	The patient's PSA level is 0.8 ng/ml.	TRUE	CHUNK			1					
	The patient has ISUP grade group 3.	TRUE	CHUNK			1					
▲	Salvage radiotherapy (SRT) is recommended for this patient.	FALSE	CHUNK	Atomic fa		1		1			1
	The patient has rising PSA levels.	TRUE	CHUNK			1					
★	The patient has pathologic T3a disease with positive surgical margins.	FALSE	CHUNK	Atomic fa		1	1				
	Initiating SRT is appropriate to reduce the risk of further progression.	TRUE	CHUNK			1					
▲	The literature suggests that SRT is particularly beneficial for patients with a pre-SRT PSA level below 1.0 ng/ml.	TRUE	CHUNK			0					1
	SRT is associated with improved biochemical control and overall surviv	TRUE	CHUNK			1					
	The use of androgen deprivation therapy (ADT) in conjunction with SRT may be beneficial depending on individual risk factors and treatment goals.	TRUE	CHUNK			1					
							Hallucination	Inaccuracy	Missing Info	Missing Context	

★ hallucination -> information never given

▲ Reference chunks:

CHUNK 1 (source: ESTRO ACROP AHT.txt)

a PSA recurrence out of the undetectable PSA level after RP within the first 5 years. Overall, the risk of subsequent metastases and PCa-specfic mortality is predicted by high pathological ISUP grade, short PSA doubling time, high pre-salvage RT PSA, the presence of positive surgical margins and high pT category. Monitoring may be possible for men with ISUP grade group < 4, who experience prolonged periods before PSA becomes detectable and with a slow doubling time. **However, it is accepted that early salvage RT provides the highest chance of cure in patients ideally before PSA reaches 0.2 ng/ml.** Recently, the individualization of salvage RT by combination with ADT in patients with rising PSA after prostatectomy and no evidence of lymph node positive disease in the pelvis (pN0 or cNx) was analyzed in three randomized trials: The RTOG 9601 trial used a RT dose of 64.8 Gy to the prostatic fossa only and randomized 760 patients to receive bicalutamide 150 mg daily or placebo for two years. After a

CHUNK 2 (source: Guidelines EAU with papermage tables.txt)

increase in PCa-specific survival relative to those who received no salvage treatment ($p < 0.001$). Salvage RT has been shown to be effective mainly in patients with a short PSA-DT. In a retrospective multi-centre study including 25,551 patients with at most one high-risk factor after RP (ISUP grade group 4-5 or pT3/4), initiating sRT above a PSA level of 0.25 ng/mL was associated with increased ACM-risk. **After a median follow-up of six years, patients who received sRT at a PSA level >0.25 ng/mL had a significantly higher ACM-risk (AHR, 1.49; 95% CI, 1.11 to 2.00; P =.008) compared with men who received sRT when the PSA was ≤ 0.25 mg/mL.** For an overview of SRT see Table 6.4.3. The EAU BCR definitions have been externally validated and may be helpful for individualised treatment decisions. Despite the indication for salvage RT, a wait and see strategy remains an option for the EAU BCR Low-Risk group. Although biochemical progression is now widely accepted as a surrogate marker of PCa recurrence;

Corrected Atomic Facts:

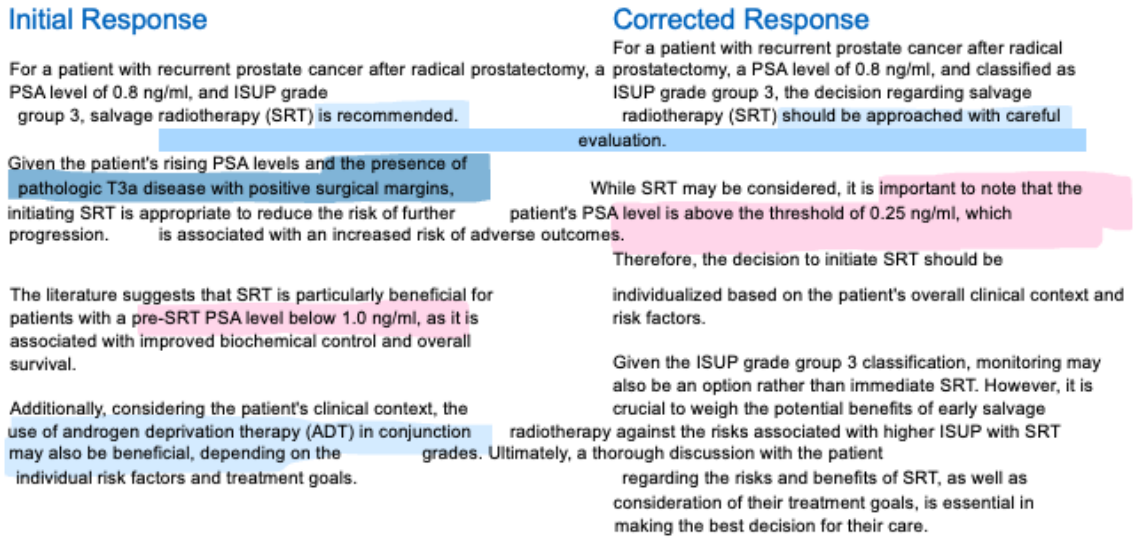
Atomic fact "Salvage radiotherapy (SRT) is recommended for this patient." was corrected to: "Salvage radiotherapy (SRT) may be considered for this patient, but careful evaluation is necessary due to the patient's PSA level of 0.8 ng/ml, which is above the threshold of 0.25 ng/ml associated with increased risk of adverse outcomes; thus, the decision should be individualized based on the patient's overall clinical context.

-> Atomic fact correction incorrect: SRT should be performed although adverse effects might be worse; LLM too avoiding, context missing (1)

Atomic fact "The patient has pathologic T3a disease with positive surgical margins." was corrected to: "The patient has a PSA level of 0.8 ng/ml and is classified as ISUP grade group 3, which may allow for monitoring rather than immediate salvage

radiotherapy (RT), but careful consideration should be given to the potential benefits of early salvage RT due to the risk factors associated with higher ISUP grades.

-> Atomic fact correction correct



Initial Response

For a patient with recurrent prostate cancer after radical prostatectomy, a PSA level of 0.8 ng/ml, and ISUP grade group 3, salvage radiotherapy (SRT) is recommended.

Given the patient's rising PSA levels and the presence of pathologic T3a disease with positive surgical margins, initiating SRT is appropriate to reduce the risk of further progression.

The literature suggests that SRT is particularly beneficial for patients with a pre-SRT PSA level below 1.0 ng/ml, as it is associated with improved biochemical control and overall survival.

Additionally, considering the patient's clinical context, the use of androgen deprivation therapy (ADT) in conjunction may also be beneficial, depending on the individual risk factors and treatment goals.

Corrected Response

For a patient with recurrent prostate cancer after radical prostatectomy, a PSA level of 0.8 ng/ml, and classified as ISUP grade group 3, the decision regarding salvage radiotherapy (SRT) should be approached with careful evaluation.

While SRT may be considered, it is important to note that the patient's PSA level is above the threshold of 0.25 ng/ml, which is associated with an increased risk of adverse outcomes. Therefore, the decision to initiate SRT should be individualized based on the patient's overall clinical context and risk factors.

Given the ISUP grade group 3 classification, monitoring may also be an option rather than immediate SRT. However, it is crucial to weigh the potential benefits of early salvage radiotherapy against the risks associated with higher ISUP grades. Ultimately, a thorough discussion with the patient with SRT regarding the risks and benefits of SRT, as well as consideration of their treatment goals, is essential in making the best decision for their care.

-> Quality of corrected response equal (improved and worsened parts)
