# ImmunoFOMO: Are Language Models missing what oncologists see?

**Aman Sinha** **Bogdan-Valentin Popescu** **Xavier Coubez**
**Marianne Clausel** **Mathieu Constant**

Université de Lorraine, Nancy, France ICANS, Strasbourg, France
**Correspondence:** aman.sinha@univ-lorraine.fr

## Abstract

Language model (LM) capabilities have rapidly advanced over the past decade, driving their adoption across fields, including biomedical research. While domain-specific pretrained language models (PLMs) have long supported biomedical NLP, recent focus has shifted to medical large language models (LLMs) and their medical language understanding capabilities. In this paper, we investigate this capability by probing the medical conceptual alignment of various LMs against clinical expertise for identifying hallmarks of immunotherapy in breast cancer abstracts. Our results show a clear trade-off: LLMs achieve stronger performance and higher consistency on high-level concepts, whereas PLMs are more reliable for fine-grained, low-level concept identification. We find that hallucinations are the primary factor limiting LLM performance in these settings; once mitigated, LLMs surpass PLMs across all tiers of biomedical concepts. Despite this, PLMs remain competitive, particularly in closely aligning with medical expert judgments on detailed conceptual distinctions.

ICANS-Strasbourg/immuNLP

## 1 Introduction

Language models (LMs) are built for automating linguistic intelligence of humans. In practice, they are evaluated on various natural language processing (NLP) benchmarks including lexical tasks such as entity recognition or semantic tasks such as inference or document classification. Furthermore, LMs have also demonstrated strong performance in domains such as biomedical and clinical text (Lewis et al., 2020; Liu et al., 2024). However, beside performance it is crucial for biomedical LMs to capture the hierarchical structure of medical concepts (e.g. "pneumonia" being an instance of "respiratory infection", which is an instance of "disease"). Figure 1 shows an illustrative example from hierarchal structure in immunotherapy related concepts.

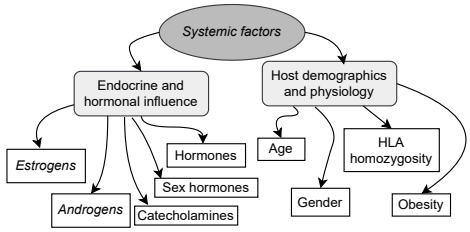



Figure 1: Illustration of conceptual hierarchical structure for a hallmark of immunotherapy (e.g., Systemic factors). denotes the TIER-I (high) level of hallmark; denotes the TIER-II level of hallmark; and denotes the TIER-III (low) level of hallmark.

In the context of oncology research, concepts such as Hallmarks of Cancer have been proposed to classify the most important characteristics of cancer development (Hanahan and Weinberg, 2000). For instance, Immunotherapy in breast cancer is a relatively new addition to the therapeutic arsenal (Loibl et al., 2024; Gennari et al., 2021). Better understanding of the underlying hallmarks of response and resistance to immune checkpoint inhibitors in breast cancer would provide valuable data for developing novel therapeutic strategies and treatment combinations. These hallmarks require nuanced understanding of unstructured clinical or scientific narratives. Despite this, little is known about whether existing LMs can reliably recognize and reason about such expert-level concepts.

Previous work in biomedical NLP has largely focused on fact-based question answering, named entity recognition, or relation extraction (Blake, 2010; Su et al., 2022). However, far fewer studies have examined the degree to which LMs align with expert interpretations in abstract, multi-faceted concept identification tasks — particularly those that demand conceptual inference rather than surface pattern matching (Navarro et al., 2025; Workum et al., 2025). To our knowledge, no existing evaluation benchmark tests model understanding of



*Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 11–21
November 6-10, 2026 ©2026 Association for Computational Linguistics

biomedical conceptual structure, beyond standard ontologies such as ICD-10 / SNOMED (Organization et al., 2004; Bos and Donnelly, 2006), as interpreted by clinicians.

In this work, we introduce a novel evaluation dataset, named ESMO HOI, to probe the biomedical concept alignment of LMs against expert-labeled multi-level concepts in clinical text. More precisely, our research question is : *Can LMs demonstrate hierarchical understanding of immunotherapy concepts?* by correctly mapping biomedical documents to specific biomedical mentions and their broader therapeutic mechanisms. We study a range of models including, pre-trained language models (cf. PLMs, e.g., BioBERT (Lee et al., 2020), SciBERT (Beltagy et al., 2019), BioClinicalBERT (Li et al., 2022), etc.) and large language models (e.g., LLaMA (AI@Meta, 2024), Qwen (Bai et al., 2023), Gemma (Team et al., 2024)) — to identify hallmarks of immunotherapy (HoI) in breast cancer research abstracts.

## 2 Related Work

**LMs in Biomedical Domain.** The success of PLMs in general NLP has driven the development of domain-specific biomedical models such as BIOBERT (Lee et al., 2020), SCIBERT (Beltagy et al., 2019), and PUBMEDBERT (Gu et al., 2021), which leverage large-scale biomedical corpora to improve medical language understanding. More recently, instruction-tuned medical LLMs like ALPACAMED and MED-PALM (Zhang et al., 2023; Singhal et al., 2022) aim to simulate diagnostic reasoning and clinical decision-making, though their ability to capture fine-grained concepts and alignment with expert reasoning remains an open challenge.

**Hallmark-Based Text Mining.** The *Hallmarks of Cancer* framework (Hanahan and Weinberg, 2000) provides a biologically grounded categorization of cancer mechanisms and is widely used in biomedical text mining to map literature to canonical processes (Hanahan and Weinberg, 2000; Baker et al., 2017; Baker and Korhonen, 2017). Automatic hallmark classification supports literature triage, evidence synthesis, and trend analysis, with early approaches using neural networks and biomedical LMs to capture semantic cues (Baker et al., 2017; Baker and Korhonen, 2017). LLMs can extend this by enabling deeper reasoning across abstraction levels from high-level mechanisms to fine-grained concepts making them suitable for evaluating conceptual alignment and hierarchical understanding in biomedicine.

**Biomedical Conceptual Alignment.** Biomedical LMs perform well on domain-specific NLP benchmarks (Yan et al., 2024), but these mainly assess surface-level tasks like retrieval and classification rather than deep medical reasoning. This creates a gap between model predictions and clinical interpretation, as evaluations rarely test the hierarchical, compositional understanding central to expert reasoning (Fivez et al., 2021; Khatir and Reddy, 2024). To bridge this gap, we assess LMs on their ability to reflect expert interpretations of hierarchical biomedical concepts, focusing on conceptual hallmarks of immunotherapy across abstract (high-level) to specific (low-level) categories. Our framework builds on the psychological theory of conceptual hierarchies (Pedrotti et al., 2025), distinguishing *superordinate* (TIER-I), *basic* (TIER-II), and *subordinate* (TIER-III) levels. By comparing model outputs with expert annotations across tiers, we quantify conceptual alignment and probe whether LMs exhibit structured understanding akin to human medical reasoning.

## 3 Methodology

**Dataset.** We extracted free access abstracts from ESMO congresses, namely, ESMO (ES), ESMO-Immunooncology (IO), and ESMO Breast (BR) between 2020-2024 leading to a total of ∼10K abstracts[1]. We then filtered the abstracts in two steps: first, we filtered them to identify breast cancer related abstracts via term (`breast`) (Filter1). We then applied an additional filter (Filter2) to obtain immunotherapy-centric abstracts via the following four keywords: `immunotherapy` or `immune checkpoint` or `pembrolizumab` or `keynote`. These two steps resulted in a dataset comprising 239 abstracts. Finally, we also manually verified (MANVER) and discarded any abstract that was not related to breast cancer and/or immunotherapy that resulted into a final dataset of 188 abstracts and corresponding detailed statistics in Table 1. See Appendix D for more details.

**Hallmark of Immunotherapy.** An expert curated list of hallmarks of immunotherapy was constructed with the help of an oncologist and based

[1]While we provide the entire collection in our public repository, for current study we focus only on breast cancer and immunotherapy related abstracts.

| | ESMO-IO | ESMO-ES | ESMO-BR | TOTAL |
|---|---|---|---|---|
| All | 810 | 8232 | 967 | 10009 |
| +Filter1 | 62 | 1646 | 967 | 2675 |
| +Filter2 | 27 | 167 | 45 | 239 |
| +MANVER | 26 | 122 | 40 | 188 |

Table 1: ESMO HOI Dataset Description.

| Metric | Criteria | Description |
|---|---|---|
| Cohen's Kappa ($C_\kappa$) | inter- | Average inter-annotator agreement between each model and the expert clinician |
| Percentage agreement (A%) | inter- | Average exact match (accuracy) between each model and the expert clinician. |
| Percentage consensus (C%) | intra- | Percentage of samples for which all models predict the same label. |
| Jensen-Shannon Distance | inter- | Distance between model and expert class distribution |

Table 2: Description of metrics used for evaluating alignment between LMs against experts.

on available published literature (Morad et al., 2021; Karasarides et al., 2022; Cogdill et al., 2017). The hierarchical hallmark framework comprises three tiers: TIER-I includes 9 primary hallmarks; TIER-II contains 27 sub-categories derived from the first tier; and TIER-III consists of 177 keywords mapped to one of the TIER-II categories. The list for each tier of hallmarks is provided in Table 7.

**Expert Annotations.** Annotations were done by a clinical expert in oncology (notably breast cancer), who provided for each abstract the most appropriate two TIER-I hallmark labels. These labels then served as the gold reference for the alignment study of language models and comparison with the oncologist.

**Task Definition.** We evaluate biomedical conceptual alignment in LMs by formulating the task as a hallmark of immunotherapy based document classification task using unsupervised text classification techniques (Schopf et al., 2022). We explore two complementary approaches:

(a) *CLS Embedding-based approach.* For this approach, we extract the [CLS] token representation from pre-trained language models (PLMs) and use it to encode both clinical abstracts and hallmark category descriptions into a shared semantic space. For large language models (LLMs), which may not expose explicit [CLS] tokens, we aggregate token-level embeddings across the input text to produce document-level representations.[2] Hallmark labels are then assigned by computing similarity between the embedded abstract and each hallmark description.

(b) *LM Prompting-based approach.* Here, we adopt a zero-shot classification setup. Following the method proposed by Li et al. (2023), we prompt masked language models (e.g., BERT-based PLMs) using cloze-style templates to infer the most probable hallmark. For generative LLMs, we employ a direct prompting strategy (fig. 7), where the model is given a clinical abstract alongside a list of hallmark candidates and instructed to select the most appropriate one.

With the above two approaches we conduct this evaluation task across all three levels of the conceptual hierarchy TIER-I, TIER-II, and TIER-III and report performance with respect to expert-annotated TIER-I labels to measure alignment at a high level of abstraction.

**Models.** We used two pools of models: a pool of five pre-trained language models (PLMs), and a pool of five open sourced large language models (LLMs). The list of the different models used in the experiment is provided in Appendix A Table 6.

**Evaluation.** We evaluated the models against expert annotations for TIER-I labels[3] of HoI using F1-score (weighted avg). We use TIER-I labels for evaluation by converting predicted TIER-II and TIER-III to TIER-I level via aggregation (See mapping from other tiers to TIER-I in Table 7) . For eg. both "Psychosocial factor" (TIER-II) and "PTEN" (TIER-III) would yield "Oncogenic signaling" (TIER-I) label. Additionally, we evaluated the language models for inter-annotator agreement (IAA) using percentage agreement (A%) and Cohen's kappa ($\kappa$) against the expert and for intra-annotator agreement using percentage consensus ($\mathcal{C}_\delta$) where $\delta$ is a threshold. We also utilize Jensen–Shannon distance for LM-expert closeness evaluation. Description of the metrics can be found in Table 2.

## 4 Results & Discussion

**Performance on HoI classification task.** We present the task performance comparison of a pool

[2]We experimented with various aggregation strategies (e.g., mean pooling, max pooling), and found that *mean pooling* yielded the most consistent performance across LLMs.

[3]The curated hallmark list also contained low-level labels in TIER-II and III list (See table 7). So, we converted the lower-level labels (TIER-II or III) to TIER-I based on the aggregation criteria provided in the hallmark list.

| | CLS Embedding | | | LM Prompting | | |
|---|---|---|---|---|---|---|
| | TIER-I | TIER-II | TIER-III | TIER-I | TIER-II | TIER-III |
| Clinical-BigBIRD | 0.213 | 0.151 | 0.188 | 0.001 | 0.015 | 0.194 |
| PubmedBERT | **0.329** | **0.431** | **0.497** | 0.025 | 0.147 | 0.150 |
| BioBERT | 0.210 | 0.201 | 0.327 | 0.071 | 0.123 | 0.224 |
| BioClinicalBERT | 0.198 | 0.153 | 0.231 | 0.057 | 0.111 | 0.151 |
| SciBERT | 0.250 | 0.153 | 0.226 | 0.120 | 0.159 | 0.159 |
| Llama-8B | 0.248 | 0.291 | 0.271 | 0.378 | 0.381 | 0.269 |
| Gemma-9B | 0.111 | 0.236 | 0.416 | **0.445** | **0.472** | **0.381** |
| MedQwen-7B | 0.235 | 0.174 | 0.081 | 0.247 | 0.197 | 0.281 |
| DeepSeek-R1 | 0.221 | 0.384 | 0.462 | 0.421 | 0.388 | 0.247 |
| BioMistral-7B | 0.147 | 0.321 | 0.338 | 0.216 | 0.257 | 0.130 |

Table 3: Weighted F1 score comparison between PLMs and LLMs under two different approaches

| | TIER-I | | TIER-II | | TIER-III | |
|---|---|---|---|---|---|---|
| Inter-annotator (Against Expert) | | | | | | |
| | A% | $\kappa$ | A% | $\kappa$ | A% | $\kappa$ |
| PLMs | 31.70 | 0.11 | 32.45 | 0.062 | 32.66 | 0.101 |
| LLMs | 37.02 | 0.178 | 34.260 | 0.179 | 34.04 | 0.169 |
| Intra-annotator (Among Models) | | | | | | |
| | $\mathcal{C}_{\delta=1}$ | $\mathcal{C}_{\delta=0.8}$ | $\mathcal{C}_{\delta=0.6}$ | $\mathcal{C}_{\delta=1}$ | $\mathcal{C}_{\delta=0.8}$ | $\mathcal{C}_{\delta=0.6}$ | $\mathcal{C}_{\delta=1}$ | $\mathcal{C}_{\delta=0.8}$ | $\mathcal{C}_{\delta=0.6}$ |
| PLMs | 17.55 | 36.7 | 71.28 | 0 | 31.91 | 69.68 | 0 | 13.30 | 46.28 |
| LLMs | 10.11 | 29.79 | 62.23 | 3.19 | 17.02 | 52.66 | 2.66 | 18.62 | 67.55 |

Table 4: Alignment of LMs for HoI concepts.

of PLMs and LLMs against clinical expert annotations (See Table 3) for HoI-based classification task based on our two approaches. For CLS embedding based approach, PLMs surpasses LLMs in all the three tiers of concepts whereas for LM Prompting based approach, LLMs outperform PLMs across all tiers. Both sets of results are inline with the literature as PLMs have better embedding representation capacity whereas LLMs are auto-regressive next token predictors adapted for prompting based task. Furthermore, on comparing the best yielding approach from each pool of models and both the approaches, we notice that LLMs outperform PLMs for TIER-I and TIER-II however PLMs are better on TIER-III.

We observed an overall low performance (range between 0.1 - 49.7% Weighted F1; See Table 3) for PLMs which is intuitive as they are traditionally trained in a supervised/semi-supervised setting and then tested on downstream tasks. However, even with no supervised training, PubmedBERT and BioBERT performed competitively against LLMs especially for low-level hallmarks confirming their rich biomedical representation capacity. For LLMs, we notice some unexpected behavior as MedQwen-7B and BioMistral-7B were outperformed by general domain LLMs. Noticebly, LLMs did not demonstrate success on HoI classification task at TIER-III tier.

**LMs' alignment with oncologist.** Table 4 shows a comparison of inter annotator agreement between LMs against the clinical expert respectively via accuracy metric (A%). We observed a consistently low and decreasing IAA between PLMs and the clinical expert ($\kappa < 0.12$). For LLMs, we also see a low agreement yet more consistent overall agreement against expert. Among the two pools of models, we observe an opposite trend in average agreement (A%) from left to right. Additionally, we compare HoI predicted concept distributions from all LMs with medical expert preferences using Jensen–Shannon distance and observe that while LLMs are generally closer to experts, PubMedBERT is closest for Tier-II and Tier-III HoI predictions (See fig. 6, in Appendix).

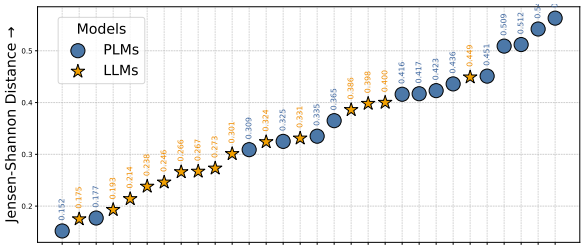



Figure 2: Jensen-Shannon distance between LMs and medical expert. Blue dots denotes PLMs and Star denotes LLMs. (Lower is better)

**LM's mutual alignment.** We report the overall intra- model consensus percentage for PLMs and LLMs in table 4. We observe that there is no consensus in PLMs for lower-level hallmark labels, and similarly LLMs struggle with the TIER-III labels. Both pools of models shows similar decreasing trend in consensus with towards lower-level hallmarks. Additionally, fig. 5 presents the agreement confusion matrix for every pair of models.

**Hallucinations in LLMs lead to lower precision but mitigation leads to improvement.** Table 5 reports the hallucination count and corresponding performance scores for LLMs under LM prompting, before and after the hallucination mitigation step. We observe that despite having better representational capacity for TIER-I concepts as seen in table 3, LLMs tend to produce unwanted or incorrect information due to having access to tremendous amount of information. The higher percentage of hallucinations for TIER-III can be attributed to the observation: despite providing an explicit list of classes, LLMs don't restrict to HoI label space. Noticeably, when LLMs are asked to identify low-level HoI, they hallucinate pro-

| | TIER-I | | TIER-II | | TIER-III | |
|---|---|---|---|---|---|---|
| | Orig | HM | Orig | HM | Orig | HM |
| Llama-8B | 0<br>0.378 | 0<br>0.378 | 1<br>0.381 | 1<br>0.381 | 107<br>0.269 | 9<br>0.332 |
| Gemma-9B | 2<br>0.445 | 1<br>0.450 | 3<br>0.472 | 2<br>0.477 | 92<br>0.381 | 26<br>0.544 |
| MedQwen-7B | 2<br>0.247 | 2<br>0.247 | 8<br>0.197 | 7<br>0.202 | 94<br>0.281 | 70<br>0.328 |
| DeepSeek-R1 | 0<br>0.421 | 0<br>0.421 | 9<br>0.388 | 8<br>0.387 | 70<br>0.247 | 30<br>0.335 |
| BioMistral-7B | 1<br>0.216 | 1<br>0.216 | 52<br>0.257 | 51<br>0.265 | 142<br>0.130 | 125<br>0.153 |

Table 5: Impact of hallucination mitigation on LLMs. Orig. denotes the original performance scores and HM denotes improved scores after hallucination mitigation. Green font/highlight denotes denotes improved hallucination count and performance.

ducing TIER-II or TIER-I hallmarks. We observe that when tasked with identifying low-level HoI (TIER-II and TIER-III), LLMs often generate higher-level (TIER-I/II) or overly generic terms (e.g., "immunotherapy", "breast cancer"). These hallucinations range from highly specific biomedical entities (e.g., "Hashimoto's thyroiditis", "MAP kinase-activated protein kinase 2 (MK2)") to broad, less informative concepts (e.g., "immunotherapy", "COVID-19"). This issue is most pronounced in the TIER-III setting, where the label space is substantially larger ($\approx \times 20$) than TIER-I.

To acknowledge this problem in LLMs for fair evaluation, we performed two types of evaluation: firstly, we did strict evaluation by flagging any prediction the model predicted that does not exist in the provided reference list of hallmarks. Both for TIER-I and II, LLMs suffered relatively less hallucination across all the models with the exception of BioMistral-7B which contained the highest percentage of hallucinated output. TIER-III level was observed to be the most difficult for all LLMs with the range of hallucination in predictions to be between 37%-75%.

Secondly, we perform relaxed evaluation, for which we perform hallucinations mitigation by manual post-processing flagged predictions by mapping them to the closest valid concept from an expert-curated list. We provide systematic instruction for illustration of automatic hallucination mitigation module in fig. 3. This served two purposes: (i) to characterize the types of hallucinations produced by LLMs, and (ii) to mitigate their impact on predictions. While this approach substantially reduced hallucinations, it did not eliminate them entirely. Encouragingly, some hallucinated terms

You are an oncologist expert. An AI system has predicted the below category for TIER-II hallmarks:

«Myeloprotection and immune response»

Consider the TIER-II list and identify the closest match, and if not, return not found.

Response:
Immune regulation and dysregulation

Figure 3: Example of the procedure for manual hallucination mitigation. Blue denotes the TIER category for which correction is need; Red denotes the hallucinated prediction and Orange denotes closest correction.

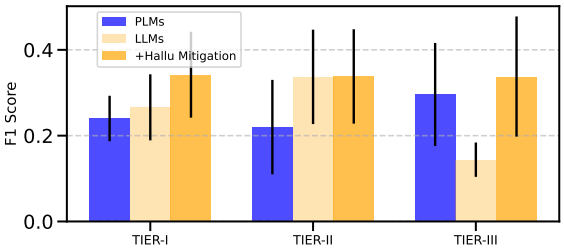



Figure 4: Comparison of PLMs and LLMs without and with Hallucination Mitigation. Y-axis denotes avg. F1.

are meaningful and could extend the TIER-III vocabulary. For instance, terms such as "miRNA", "CDK 4/6", "oncolytic viruses", and "PARP1" may enrich the concept set and highlight gaps in current LLM understanding. Overall, we observed that hallucination mitigation led LLMs to outperform PLMs on TIER-III level as shown in fig. 4.

## 5 Conclusion

Conceptual alignment is a crucial criterion for evaluating language understanding in language models. In this work, we examined PLMs and LLMs for classifying biomedical abstracts according to hallmarks of immunotherapy using our proposed ESMO-HoI dataset, and evaluated their alignment with a clinical expert. Our results show that LLMs perform better on high-level concepts and exhibit greater consistency, likely due to their broader knowledge access. In contrast, PLMs demonstrate stronger representation for low-level concept identification, where LLMs are more prone to hallucination. However, with effective hallucination mitigation, LLMs can outperform PLMs across different tiers of biomedical concepts. Overall, PLMs remain competitive in terms of alignment with medical expert judgments.

## Limitations

**Probing intermediate level concepts.** One of the limitation of this study is the lack of expert annotation at TIER-II and TIER-III. In the absence of such annotation, we investigated the level of agreement between models as a proxy for the alignment of language models reasoning with human expert. Such annotations would help better understand the logic behind the abstract TIER-I annotation by the human expert. However, the large number of labels for tier-III (low-level) hallmarks poses difficulties for the LMs to correctly annotate the TIER-I hallmarks starting from TIER-III.

**Filtering of abstracts.** Adding more filters as opposed to just four ("immunotherapy" ; "immune checkpoint" ; "pembrolizumab" ; "keynote") could increase the size of dataset and further guiding to robust analyses of potential research strategies for immunotherapy in breast cancer.

**Number of TIER-III keywords.** Another limitation is the fact that the TIER-III list can be further improved with keywords not previously presented. Adding TIER-III categories could potentially diminish hallucinations from LLMs as more specific data could be identified and interpreted correctly. We should also note that several keywords can also be reclassified as TIER-III for multiple TIER-I hallmarks. As previously mentioned, hallucinations could further be analyzed to enrich the TIER-III keywords and possibly improve performance

**Expert annotation.** The lack of agreement (see table 4) between models could be reflected in the disagreements between human annotators (Fu et al., 2024) and large label space. Therefore, having more annotators could help in explaining the model's disagreement.

## References


AI@Meta. 2024. Llama 3 model card.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, and 29 others. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Simon Baker, Imran Ali, Ilona Silins, Sampo Pyysalo, Yufan Guo, Johan Högberg, Ulla Stenius, and Anna Korhonen. 2017. Cancer hallmarks analytics tool (chat): a text mining approach to organize and evaluate scientific literature on cancer. *Bioinformatics*, 33(24):3973–3981.

Simon Baker and Anna Korhonen. 2017. Initializing neural networks for hierarchical multi-label text classification. In *Proceedings of the 16th BioNLP Workshop*, pages 307–315, Vancouver, Canada,. Association for Computational Linguistics.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: Pretrained language model for scientific text. In *EMNLP*.

Catherine Blake. 2010. Beyond genes, proteins, and abstracts: Identifying scientific claims from full-text biomedical articles. *Journal of Biomedical Informatics*, 43(2):173–189.

L Bos and K Donnelly. 2006. Snomed-ct: The advanced terminology and coding system for ehealth. *Stud Health Technol Inform*, 121:279–290.

Alexandria P Cogdill, Miles C Andrews, and Jennifer A Wargo. 2017. Hallmarks of response to immune checkpoint blockade. *British journal of cancer*, 117(1):1–7.

Pieter Fivez, Simon Suster, and Walter Daelemans. 2021. Conceptual grounding constraints for truly robust biomedical name representations. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2440–2450.

Yujuan Fu, Giridhar Kaushik Ramachandran, Nicholas J. Dobbins, Namu Park, Michael Leu, Abby R. Rosenberg, Kevin Lybarger, Fei Xia, Ozlem Uzuner, and Meliha Yetisgen. 2024. Extracting social determinants of health from pediatric patient notes using large language models: Novel corpus and methods. *ArXiv*, abs/2404.00826.

Alessandra Gennari, Fabrice André, CH Barrios, Javier Cortes, Evandro de Azambuja, Angela DeMichele, R Dent, Deborah Fenlon, Joseph Gligorov, SA Hurvitz, and 1 others. 2021. Esmo clinical practice guideline for the diagnosis, staging and treatment of patients with metastatic breast cancer. *Annals of oncology*, 32(12):1475–1495.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

Douglas Hanahan and Robert A Weinberg. 2000. The hallmarks of cancer. *cell*, 100(1):57–70.

Maria Karasarides, Alexandria P Cogdill, Paul B Robbins, Michaela Bowden, Elizabeth M Burton, Lisa H Butterfield, Alessandra Cesano, Christian Hammer, Cara L Haymaker, Christine E Horak, and 1 others. 2022. Hallmarks of resistance to immune-checkpoint

inhibitors. *Cancer immunology research*, 10(4):372–383.

Mehrdad Khatir and Chandan K Reddy. 2024. Concept formation and alignment in language models: Bridging statistical patterns in latent space to concept taxonomy. *arXiv preprint arXiv:2406.05315*.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Patrick Lewis, Myle Ott, Jingfei Du, and Veselin Stoyanov. 2020. Pretrained language models for biomedical and clinical tasks: Understanding and extending the state-of-the-art. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 146–157, Online. Association for Computational Linguistics.

Lei Li, Jingjing Xu, Qingxiu Dong, Ce Zheng, Xu Sun, Lingpeng Kong, and Qi Liu. 2023. Can language models understand physical concepts? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11843–11861, Singapore. Association for Computational Linguistics.

Yikuan Li, Ramsey M Wehbe, Faraz S Ahmad, Hanyin Wang, and Yuan Luo. 2022. Clinical-longformer and clinical-bigbird: Transformers for long clinical sequences. *arXiv preprint arXiv:2201.11838*.

Fenglin Liu, Zheng Li, Hongjian Zhou, Qingyu Yin, Jingfeng Yang, Xianfeng Tang, Chen Luo, Ming Zeng, Haoming Jiang, Yifan Gao, Priyanka Nigam, Sreyashi Nag, Bing Yin, Yining Hua, Xuan Zhou, Omid Rohanian, Anshul Thakur, Lei Clifton, and David A. Clifton. 2024. Large language models in the clinic: A comprehensive benchmark. *Preprint*, arXiv:2405.00716.

Sibylle Loibl, F André, T Bachelot, CH Barrios, J Bergh, HJ Burstein, MJ Cardoso, LA Carey, S Dawood, L Del Mastro, and 1 others. 2024. Early breast cancer: Esmo clinical practice guideline for diagnosis, treatment and follow-up. *Annals of Oncology*, 35(2):159–182.

Golnaz Morad, Beth A Helmink, Padmanee Sharma, and Jennifer A Wargo. 2021. Hallmarks of response, resistance, and toxicity to immune checkpoint blockade. *Cell*, 184(21):5309–5337.

David Fraile Navarro, Enrico Coiera, Thomas W. Hambly, Zoe Triplett, Nahyan Asif, Anindya Susanto, Anamika Chowdhury, Amaya Azcoaga Lorenzo, Mark Dras, and Shlomo Berkovsky. 2025. Expert evaluation of large language models for clinical dialogue summarization. *Scientific Reports*, 15.

World Health Organization and 1 others. 2004. *ICD-10: international statistical classification of diseases and related health problems: tenth revision*. World Health Organization.

Andrea Pedrotti, Giulia Rambelli, Caterina Villani, and Marianna Bolognesi. 2025. How humans and LLMs organize conceptual knowledge: Exploring subordinate categories in Italian. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4464–4482, Vienna, Austria. Association for Computational Linguistics.

Tim Schopf, Daniel Braun, and Florian Matthes. 2022. Evaluating unsupervised text classification: Zero-shot and similarity-based approaches. In *Proceedings of the 2022 6th International Conference on Natural Language Processing and Information Retrieval*, NLPIR 2022, page 6–15. ACM.

K. Singhal, Shekoofeh Azizi, Tao Tu, Said Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Kumar Tanwani, Heather J. Cole-Lewis, Stephen J. Pfohl, P A Payne, Martin G. Seneviratne, Paul Gamble, Chris Kelly, Nathaneal Scharli, Aakanksha Chowdhery, P. A. Mansfield, Blaise Agúera y Arcas, Dale R. Webster, and 11 others. 2022. Large language models encode clinical knowledge. *Nature*, 620:172 – 180.

Yansen Su, Minglu Wang, Pengpeng Wang, Chunhou Zheng, Yuansheng Liu, and Xiangxiang Zeng. 2022. Deep learning joint models for extracting entities and relations in biomedical: a survey and comparison. *Briefings in Bioinformatics*, 23(6):bbac342.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, and 1 others. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Jessica D. Workum, Bas W. S. Volkers, Davy van de Sande, Sumesh Arora, Marco Goeijenbier, Diederik A.M.P.J. Gommers, and Michel E. van Genderen. 2025. Comparative evaluation and performance of large language models on expert level critical care questions: a benchmark study. *Critical Care*, 29.

Lawrence K. Q. Yan, Qian Niu, Ming Li, Yichao Zhang, Caitlyn Heqi Yin, Cheng Fei, Benji Peng, Ziqian Bi, Pohsun Feng, Keyu Chen, Tianyang Wang, Yunze Wang, Silin Chen, Ming Liu, and Junyu Liu. 2024. Large language model benchmarks in medical tasks. *Preprint*, arXiv:2410.21348.

Xinlu Zhang, Chenxin Tian, Xianjun Yang, Lichang Chen, Zekun Li, and Linda Ruth Petzold. 2023. Alpacare:instruction-tuned large language models for medical application. *Preprint*, arXiv:2310.14558.

# A Reproducibility

## A.1 Models

Table 6 lists the huggingface model id of all the models used in the experimentation to enable better reproducibility.

| | | huggingface model_id |
|---|---|---|
| PLMs | ClinicalBigBird | yikuan8/Clinical-BigBird |
| | PubmedBERT | NeuML/pubmedbert-base-embeddings |
| | BioBERT | dmis-lab/biobert-v1.1 |
| | ClinicalBERT | emilyalsentzer/Bio_ClinicalBERT |
| | SciBERT | allenai/scibert_scivocab_uncased |
| LLMs | Llama-3-8B-It | meta-llama/Meta-Llama-3-8B-Instruct |
| | Gemma-2-9B | google/gemma-2-9b-it |
| | Med-Qwen-7B | Echelon-AI/Med-Qwen2-7B |
| | DeepSeek-R1 | deepseek-ai/DeepSeek-R1-Distill-Llama-8B |
| | BioMistral-7B | ZiweiChen/BioMistral-Clinical-7B |

Table 6: Model IDs used in the experimentation

### A.2 Alignment Metrics

Table 2 lists the different metrics we used to evaluate the medical concept alignment of the language models. We consider two criterion: (i) Inter model pool metrics, which evaluate the alignment of models against medical expert annotations. (ii) Intra model pool metrics, which evaluates the alignment of model pools themselves (eg. PLM/ LLM model pool.)

## B Additional Results

### B.1 Agreement Confusion matrix

Figure 5 presents model agreement in detail. The first row shows confusion matrices for the PLM pool across TIER-I to TIER-III (left to right), while the second row presents the corresponding results for the LLM pool.

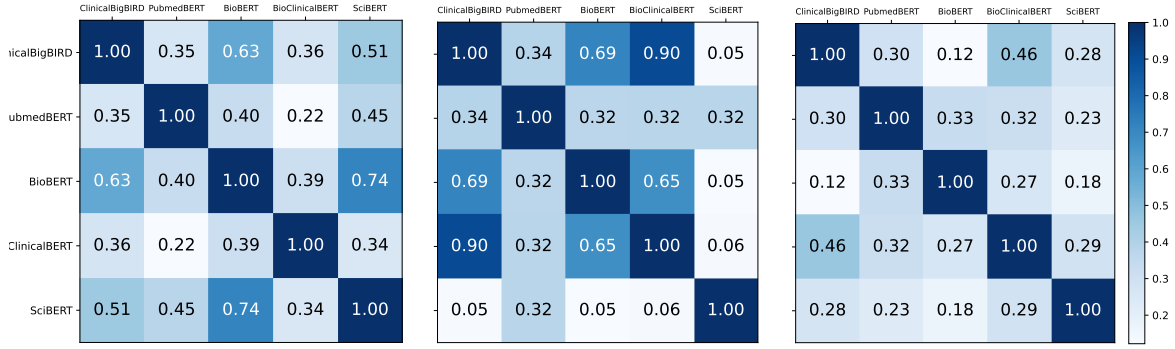
## C AI Usage.

We use AI tool like ChatGPT for improving the writing of the paper and kick-off coding setup. The authors take complete responsibility for the content of the paper.

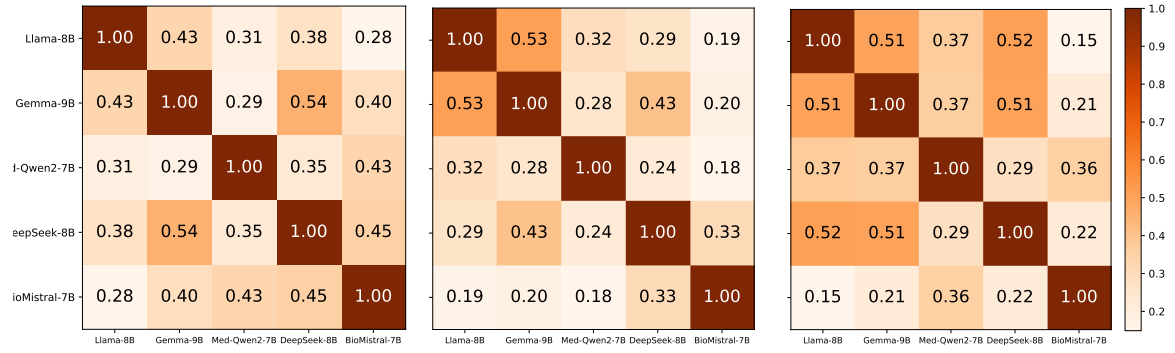
## D ESMO HOI Dataset Details

Our dataset is specifically focused on immunotherapy, a recent and significant advancement in cancer research, making it more relevant as a novel contribution. With guidance from a clinical expert, we curated a novel hierarchical label structure: for each HOI, we compiled a set of subcategories, and for each subcategory, we identified a set of associated tags. Figure 1 in the paper provides a visual overview of this structure via an example, and Table 1 presents its statistics in detail. This tree-like organization of labels constitutes a rich source of domain-specific information, tailored to immunotherapy-focused research. Our study aims to evaluate the conceptual understanding of language models, particularly in relation to their ability to understand hierarchical relationships between concepts. This objective necessitated access to a structured label hierarchy—a feature not available in any existing dataset, to the best of our knowledge.

We applied fuzzy string matching after concatenating all sections for each research abstract from the free access ESMO congress, including the Title, Year, Background, Methods, Results, Conclusions, and other relevant parts. Following this, we checked the presence of the keywords in the concatenated text, and based on the match we consider the subset of the 10k abstract for this study.




(a) PLM - TIER-I (b) PLM - TIER-II (c) PLM - TIER-III




(d) LLM - TIER-I (e) LLM - TIER-II (f) LLM - TIER-III

Figure 5: Detailed Intra-alignment (A%) between LMs over three tiers of HOI concepts (left to right).

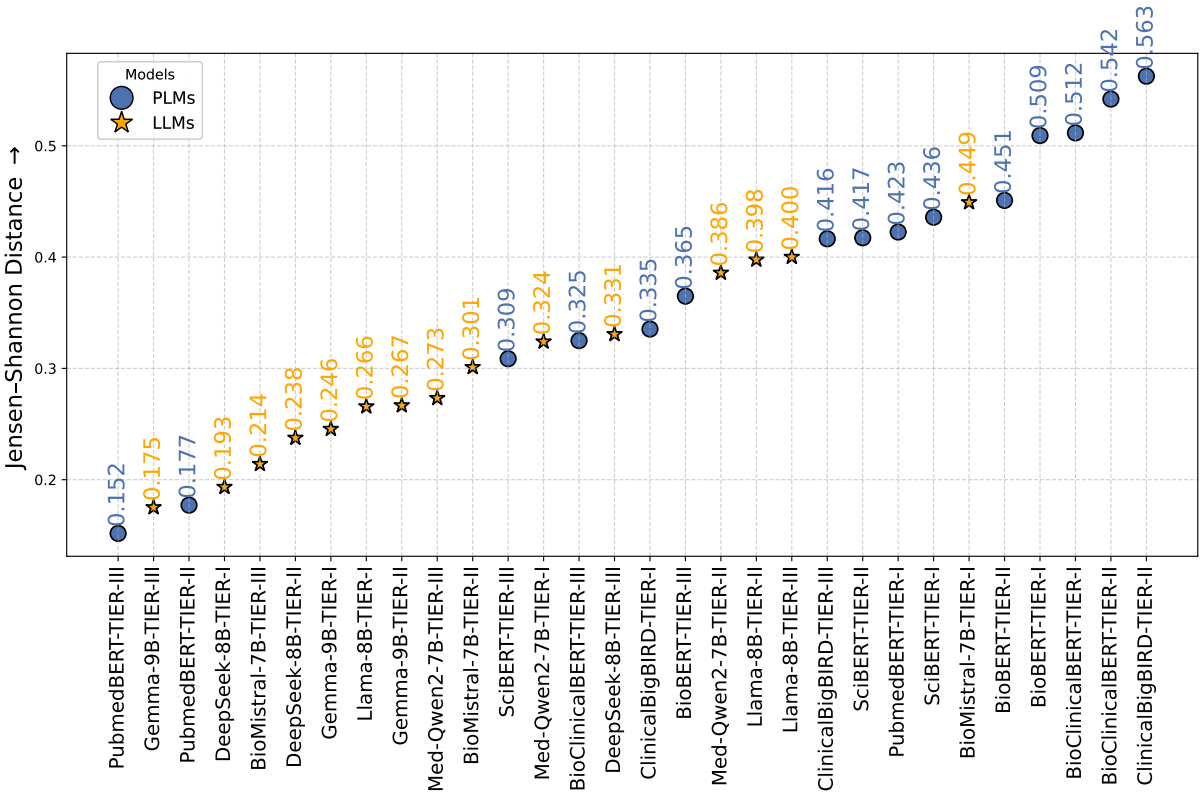



Figure 6: Jensen-Shannon distance between language models and medical expert (detailed).

Your task is to classify the given abstract based on the following 9 categories :
[ "Tumor genome and epigenome",
"Tumour microenvironment",
"Systemic immunity",
"Systemic factors",
"Microbiome",
"Oncogenic signaling",
"Tumor metabolism",
"Environmental",
"Immune checkpoint inhibitor toxicity"
]
you should return the top three suited categories (in top to bottom order) accurately as a python list. Abstract:

Durvalumab could be effective in combination with anti-HER2 agents in HER2-low breast cancer (ID 181)
The clinical challenge for treating HER2 (human epidermal growth factor receptor 2)-low breast cancer is the paucity of actionable drug targets. However, the discovery of immune checkpoint inhibitors has made immunotherapy an emerging new treatment modality for breast cancer. Moreover, several chemotherapeutic agents are known to induce immunogenic cell death by activating the immune system. Therefore, we hypothesized that modulating the tumour microenvironment using trastuzumab and or trastuzumab deruxtecan (T-Dxd) in breast organoids co-cultured with T-cells might enhance the response to immunotherapy. We established a panel of HER2-low breast cancer patient-derived organoids (PDOs), recapitulating the derived tumour. These PDOs were cocultured with immune cells (T- cells and Natural killer cells (NK cells)) and treated with T-Dxd and or trastuzumab in combination with durvalumab. Levels of cytotoxic markers were assessed using flow cytometry and cytokine assays. Our findings revealed synergistic effects in HER2-low BC patient-derived organoids when treated with T-Dxd and or trastuzumab in combination with durvalumab. We also observed antibody-dependent cellular cytotoxicity (ADCC) response with trastuzumab in combination with durvalumab. These results highlight the need to develop a combination treatment of PD-1/PD-L1 inhibitors with targeted therapies, and other immunotherapies to maximize clinical efficacy. Altogether, despite preliminary, these findings support the rationale for combining anti-HER2 therapies with immunotherapy in HER2-low BC patients.

Figure 7: Zero-shot prompting template used for LLM based HOI identification.

Text :
Durvalumab could be effective in combination with anti-HER2 agents in HER2-low breast cancer (ID 181)
The clinical challenge for treating HER2 (human epidermal growth factor receptor 2)-low breast cancer is the paucity of actionable drug targets. However, the discovery of immune checkpoint inhibitors has made immunotherapy an emerging new treatment modality for breast cancer. Moreover, several chemotherapeutic agents are known to induce immunogenic cell death by activating the immune system. Therefore, we hypothesized that modulating the tumour microenvironment using trastuzumab and or trastuzumab deruxtecan (T-Dxd) in breast organoids co-cultured with T-cells might enhance the response to immunotherapy. We established a panel of HER2-low breast cancer patient-derived organoids (PDOs), recapitulating the derived tumour. These PDOs were cocultured with immune cells (T- cells and Natural killer cells (NK cells)) and treated with T-Dxd and or trastuzumab in combination with durvalumab. Levels of cytotoxic markers were assessed using flow cytometry and cytokine assays. Our findings revealed synergistic effects in HER2-low BC patient-derived organoids when treated with T-Dxd and or trastuzumab in combination with durvalumab. We also observed antibody-dependent cellular cytotoxicity (ADCC) response with trastuzumab in combination with durvalumab. These results highlight the need to develop a combination treatment of PD-1/PD-L1 inhibitors with targeted therapies, and other immunotherapies to maximize clinical efficacy. Altogether, despite preliminary, these findings support the rationale for combining anti-HER2 therapies with immunotherapy in HER2-low BC patients.

is this about «Tumor genome and epigenome»? [MASK]

Figure 8: Masked Language prompting template used for PLM based HOI identification.

| TIER-I | TIER-II |
|---|---|
| "Tumor genome and epigenome" | "Genetic alterations and immune visibility", "Epigenetic dysregulation", "Antigen presentation and immune evasion" |
| "Tumour microenvironment" | "Immune cell infiltrates", "Microbiome and extracellular factors", "Checkpoint molecules and inhibitory receptors", "Cytokines and soluble immune modulators", "Immune landscape and spatial architecture" |
| "Systemic immunity" | "Immune cell function and states", "Immune regulation and dysregulation", "Soluble mediators and modifiers" |
| "Systemic factors" | "Endocrine and hormonal influence", "Host demographics and physiology" |
| "Microbiome" | "Microbial sites", "Microbial imbalances and signatures", "Microbial species or families" |
| "Oncogenic signaling" | "Key pathways", "Genetic drivers and regulators" |
| "Tumor metabolism" | "Energy metabolism", "Metabolic reprogramming" |
| "Environmental" | "Exposures", "Psychosocial factors", "Lifestyle", "Microbial factors" |
| "Immune checkpoint inhibitor toxicity" | "Clinical manifestations", "Mechanisms", "Management" |
| **TIER-I** | **TIER-III** |
| "Tumor genome and epigenome" | 'CTLA-4 polymorphism', 'Human leukocyte antigen (HLA)', 'HLA heterozygosity', 'Beta-2 microglobulin', 'Tumour mutational burden (TMB)', 'Mutational load', 'Mutations/megabase', 'Mutations', 'Neoantigens', 'Genetic alterations', 'DNA mismatch repair', 'dMMR', 'Microsatellite instability (MSI)', 'Histone methyltransferases', 'Dysregulated DUX4 expression', 'Epigenetic alterations', 'DNA methyltransferase inhibitors', 'Histone deacetylase inhibitors (HDAC)', 'Epigenetic chromatin modifications', 'Histone methyltransferases', 'Epigenetic alterations', 'DNA methyltransferase inhibitors', 'Histone deacetylase inhibitors (HDAC)', 'Epigenetic chromatin modifications', 'PALB2', 'RAD51', 'Tumor mutational burden' |
| "Tumor microenvironment" | 'CD8 T cells', 'CD4 T cells', 'Tumor-infiltrating lymphocyte (TILs)', 'Teff', 'Effector T cells', 'NK cells', 'B cells', 'Tumor-infiltrating B cell', 'Tertiary lymphoid structures (TLS)', 'Treg', 'Tumor-infiltrating Tregs', 'Myeloid-Derived Suppressor Cells (MDSCs)', 'Tumour-associated macrophages (TAMs)', 'M1/M2 macrophages', 'Neutrophils', 'Cancer associated fibroblasts (CAF)', 'Fibroblast activation protein (FAP)', 'Tissue-specific stromal', 'Tumor endothelium', 'Combined Positive Score (CPS)', 'Intratumoral microbes', 'Intratumoral microbiome signature', 'Tumor microbiota', 'Tumor-associated microbes', 'Extracellular vesicles (EVs)', 'Exosomal PD-L1', 'Extracellular matrix', 'PD-1', 'PD-L1 (B7-H1)', 'PD-L2 (B7-DC)', 'CD273', 'CTLA-4', 'CD125', 'LAG-3 (CD223)', 'TIM-3', 'TIGIT', 'VISTA / VSIG', 'B7-H3', 'BTLA (CD272)', 'Siglec-15', 'VEGF', 'IL-2', 'IL-6', 'IL-10', 'IL-12', 'IL-17', 'IL-35', 'TGF-$\beta$', 'Indoleamine-2,3-dioxygenase', 'CSF1R', 'Immune excluded', 'Immune depleted', 'Angiogenesis', 'Cellular components promoting and inhibiting immunity' |
| "Systemic immunity" | 'Peripheral immune cells', 'CD8+ T-cell immune memory', 'Effector T cells', 'T-cell activation', 'T-cell exhaustion', 'Antigenic stimulation', 'Antigenic tolerance', 'Immunosuppressive phenotypes', 'Dysfunctional IFN-$\gamma$ signaling', 'Costimulation', 'Priming of immune response', 'Peripheral antigen presenting cells (APCs)', 'Host polymorphisms', 'Chronic inflammation', 'Cytokine markers', 'Soluble factors', 'Stress hormones', 'Glucocorticoids' |
| "Systemic factors" | 'Sex hormones', 'Estrogens', 'Androgens', 'Hormones', 'Catecholamines', 'Age', 'Gender', 'Obesity', 'HLA homozygosity' |
| "Microbiome" | 'Gut microbiota', 'Oral microbiota', 'Circulating microbiota', 'Dysbiosis', 'Microbiome signature', 'Bacterial 'signatures', 'Fecal microbiota transplantation', 'Bacteroidetes', 'B. fragilis', 'Bacteroides', 'Burkholderiaceae' |
| "Oncogenic signaling" | 'Interferon (IFN)', 'IFN$\gamma$ signaling', 'JAK-STAT', 'PI3K', 'MAPK', 'Wnt/$\beta$-catenin pathway', 'Hedgehog pathway', 'BRAF', 'FGFR3', 'PTEN', 'PPAR-$\gamma$', 'KRAS', 'Myc', 'BRCA', 'HER2', 'Hormone-receptor signaling', 'Estrogen-receptor (ER) signaling', 'Progesteron-receptor (PR) signaling' |
| "Tumor metabolism" | 'Glucose', 'Lactate', 'ATP', 'NAD+', 'Aerobic glycolysis', 'Oxidative phosphorylation', 'Deregulated tumor immunometabolism', 'Hypoxia', 'Oxidative stress', 'Dysregulated mitochondrial biogenesis' |
| "Environmental" | 'Food and water contamination', 'Chemical and industrial exposure', 'Household exposures', 'Air pollution', 'UV radiation', 'Pets', 'Racial injustice', 'Depression/anxiety', 'Psychological and mental stress', 'Socioeconomic inequalities', 'Sexual discrimination', 'Smoking', 'Drug use', 'Medication', 'Surgeries', 'Physical activity', 'Alcohol consumption', 'Diet and dietary supplements', 'Viral antigens', 'Pathogenic microbes', 'Household microbial contamination', 'Environmental non pathogenic microbes' |
| "Immune checkpoint inhibitor toxicity" | 'Immune toxicity', 'irAE (immune-related adverse effects)', 'Autoimmune disease', 'Aberrant T cell activity', 'Macrophage infiltrate', 'Underlying autoimmune disease', 'Use of systemic corticotherapy', 'Corticotherapy' |

Table 7: Curated list of hallmarks of immunotherapy. The top half presents a mapping between TIER-I to TIER-II concepts of hallmarks; and the bottom half presents a mapping between TIER-I to TIER-III concepts of hallmarks.