

## Research Article

Huanshu Zhang, Lei Kang, Sawyer D. Campbell and Douglas H. Werner\*

# Chat to chip: large language model based design of arbitrarily shaped metasurfaces

<https://doi.org/10.1515/nanoph-2025-0343>

Received July 24, 2025; accepted September 28, 2025;

published online October 9, 2025

**Abstract:** Traditional metasurface design is limited by the computational cost of full-wave simulations, preventing thorough exploration of complex configurations. Data-driven approaches have emerged as a solution to this bottleneck, replacing costly simulations with rapid neural network evaluations and enabling near-instant design for meta-atoms. Despite advances, implementing a new optical function still requires building and training a task-specific network, along with exhaustive searches for suitable architectures and hyperparameters. Pre-trained large language models (LLMs), by contrast, sidestep this laborious process with a simple fine-tuning technique. However, applying LLMs to the design of nanophotonic devices, particularly for arbitrarily shaped metasurfaces, is still in its early stages; as such tasks often require graphical networks. Here, we show that an LLM, fed with descriptive inputs of arbitrarily shaped metasurface geometries, can learn the physical relationships needed for spectral prediction and inverse design. We further benchmarked a range of open-weight LLMs and identified relationships between accuracy and model size at the billion-parameter level. We demonstrated that 1-D token-wise LLMs provide a practical tool for designing 2-D arbitrarily shaped metasurfaces. Linking natural-language interaction to electromagnetic modelling, this “chat-to-chip” workflow represents a step toward more user-friendly data-driven nanophotonics.

**Keywords:** metasurfaces; large language model; deep learning

## 1 Introduction

Metasurfaces, which are defined as planar arrays of sub-wavelength scatterers that modulate the amplitude, phase, and polarization of light locally, have quickly become pivotal to nanophotonic devices [1], enabling applications from high-numerical-aperture meta-lenses [2] and holographic imagers [3] to augmented-reality displays [4]. Despite this progress, metasurface design remains constrained by the need for brute-force full-wave electromagnetic solvers such as the finite-difference time-domain (FDTD) [5] and finite-element methods (FEM) [6]. A single design iteration must traverse a high-dimensional parameter space, carry out numerous simulations, and finely adjust geometric features to satisfy spectral and/or wave-front requirements [7]. For practical devices targeting large apertures and multiple functionalities, the corresponding computational load may take days or weeks, even when executed on large clusters or supercomputers [8]. The resulting limitation discourages the exploration of unconventional materials, multi-layer stacks, and fully aperiodic layouts. To keep pace with the burgeoning applications for metasurfaces, new design paradigms that bypass repeated heavy-duty simulations are urgently required.

Recent breakthroughs in data-driven modelling offer a promising alternative route [9]. Once trained on curated pairs of optical or electromagnetic responses with corresponding metasurface geometry, deep neural networks (DNNs) can predict the optical response of previously unseen geometries within milliseconds, marking structure evaluation orders of magnitude faster than that based on full-wave solvers [10], [11]. Recent works have shown the potential of DNN-based approaches for metasurface design [12]–[14]. For instance, Malkiel et al. employed a DNN for H-shaped plasmonic nanostructure design [15]. An et al. developed a DNN to predict wideband amplitude and phase responses of quasi-freeform dielectric metasurfaces [16]. Chen et al. introduced a transformer-based model for both forward and inverse design of broadband solar metamaterial absorbers [17]. Moreover, Zhang et al. proposed a fixed-attention mechanism for the design of high-degree-of-freedom metamaterials [18].

\*Corresponding author: Douglas H. Werner, Department of Electrical Engineering, The Pennsylvania State University, University Park, PA, 16802, USA, E-mail: dhw@psu.edu

Huanshu Zhang, Lei Kang and Sawyer D. Campbell, Department of Electrical Engineering, The Pennsylvania State University, University Park, PA, 16802, USA. <https://orcid.org/0009-0009-8332-7298> (H. Zhang). <https://orcid.org/0000-0001-7718-7756> (L. Kang)

Although DNN-based models have demonstrated impressive accuracy and speed, integrating them into a metasurface design pipeline is still far from a turnkey off-the-shelf procedure [19]. Each new optical function typically requires a new training set, a custom network topology, and exhaustive hyper-parameter selections. This typically includes choosing the number of layers and neurons in each layer, which is an iterative, code-heavy process driven largely by heuristic intuition rather than first-principles guidance [20]. To this end, large language models (LLMs) present a qualitatively different proposition. LLMs are transformer-based neural networks that encapsulate billions of parameters in a single, frozen architecture pre-trained on vast amounts of natural-language text and code [21]. In this stage, the model is taught the simple objective of predicting the next word in a sequence; yet, by doing so at web-scale, it internalizes syntax, semantics, and a surprising amount of factual and mathematical structure [22]. Since the core model is fixed, researchers can simply train the LLM on task-specific datasets instead of re-designing and re-training a new network for every new task, thereby eliminating the laborious network-sizing and hand-tuning that DNNs demand. These characteristics make LLMs ideal candidates for enabling efficient design of metasurfaces with complex structures and layouts that possess various targeted functionalities.

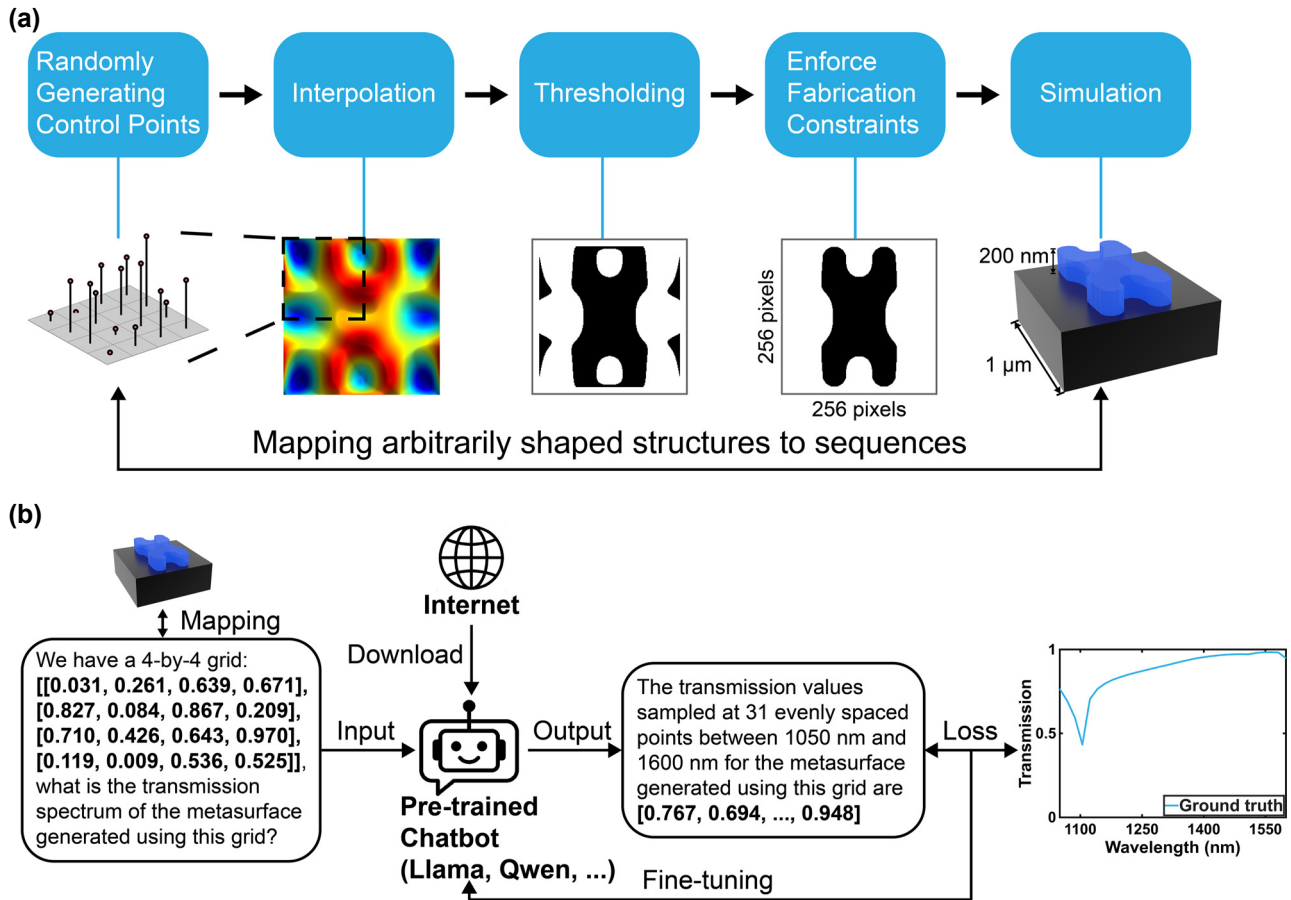
When domain precision is required, the same model can be “further trained” in an efficient manner (*i.e.*, fine-tuned) on a relatively small, task-specific dataset, such as predicting the transmission spectrum of a metasurface [23], [24]. In practice, this dataset pairs sequence-based descriptions of each unit cell (geometric parameters, material indices, lattice spacing, and so on) with descriptions of its simulated optical response (spectral magnitude, phase, near-field maps, *etc.*). This pairing mirrors the input-output structure of conventional DNNs but represents both geometry and responses in a language-like format amenable to LLMs. Because LLMs accept byte streams, neither architectural redesign nor feature engineering is necessary: the model simply learns the mapping of geometries to responses in passes. After fine-tuning on a new dataset, which takes slightly longer than the time taken to train one custom DNN, the LLMs can predict spectra within seconds, providing near-real-time feedback during design loops while removing the code-heavy scaffolding and exhaustive hyper-parameter sweeps that traditional DNN-based methods demand. Thus, LLMs promise a “chat-to-chip” route for modelling metasurfaces. For example, in their pioneering study Kim et al. fine-tuned Llama [25] for both forward

prediction and inverse design of all-dielectric metasurfaces [26], lowering the entry barrier for researchers who lack machine-learning background. Lu et al. fine-tuned ChatGPT 3.5 on various details of prompts and temperatures for the design of metamaterials [27], Liu et al. used LLMs for design recommendation of phosphorescent materials [28], and, by optimizing and stitching wavelength-scale superpixels, Lupoiu et al. introduced a multi-agent LLM framework paired with a surrogate Maxwell solver that autonomously designs metasurfaces in near-real time [29]. However, scaling these approaches from parameterized meta-atoms to arbitrarily shaped metasurfaces is of great importance to numerous applications but remains largely unexplored [30], [31]. Token-wise attention is intrinsically one-dimensional, whereas free-form surfaces require rich spatial reasoning. Emerging hybrids that couple LLM backbones with graph or vision transformers, or that embed topology as structured tokens, may be a possible solution [32]. However, a generally applicable framework for LLM-accelerated design of complex metasurfaces has yet to be reported, as accuracy of vision language models (VLMs) still lags behind LLM-level reliability while their implementation is cost-intensive and fragile [33].

Here, we present a workflow using LLMs to accelerate both forward and inverse design of arbitrarily shaped metasurfaces. We note that, in our study, an “arbitrarily shaped” meta-atom refers to a planar structure with a non-canonical or free-form shape, rather than a fully unparameterized one. Although limited to one-dimensional token streams, our results show that sequence-based LLMs are capable of capturing the physics required to predict optical responses for arbitrarily shaped metasurfaces. Also, for the inverse design section, our workflow addresses the designs of high-degree-of-freedom, randomly shaped 2D unit cells, which cannot be solved by existing image-generation or multimodal LLM approaches. This method eliminates DNN engineering and therefore further lowers the barrier for researchers with limited expertise in machine learning. Finally, cross-model benchmarks that exploit state-of-the-art LLMs in this design task are provided, establishing reference baselines to guide future work on LLM-accelerated photonic design.

## 2 Methods

Figure 1a outlines the workflow we use to generate arbitrarily shaped meta-atoms, a successfully verified parameterization approach adopted from [30]. First, a  $4 \times 4$  control-point grid was randomly generated, where each element ranges within  $[0, 1]$ . The grid was



**Figure 1:** Mapping arbitrarily shaped metasurface geometries to language sequences and training an LLM for rapid optical prediction. (a) A  $4 \times 4$  matrix of a randomly sampled grid of control-points is replicated by four-fold rotational symmetry, interpolated into a  $256 \times 256$  scalar field, binarized at a fixed threshold of 0.5, and regularised by iterative morphological opening/closing that removes isolated features smaller than 8,192 pixels and seals internal voids. The resulting binary mask is then extruded into a 200 nm-thick silicon layer on a 1  $\mu$ m-pitch glass substrate and analysed with FDTD, establishing paired grid-spectrum data. (b) Fine-tuning and inference process for forward prediction. Each grid-spectrum pair is rewritten as a natural-language prompt that encodes the control-point grid and a target output that lists the 31 transmission values between 1,050 nm and 1,600 nm. Moreover, parameter-efficient fine-tuning (LoRA) of a pre-trained LLM minimises cross-entropy between predicted and ground-truth tokens, so that at inference the model returns an accurate spectrum within seconds from a single grid prompt, eliminating the need for labour-intensive network design.

replicated by a four-fold rotational symmetry, yielding a  $7 \times 7$  lattice of control values. Interpolation converts these discrete values into a  $256 \times 256$  surface. Binarization at a fixed threshold ( $t = 0.5$ ) converts this surface into a preliminary foreground-background mask. To ensure the pattern is fabrication-friendly, the mask undergoes iterative morphological opening and closing until no further topological changes occur [31]. This regularisation step eliminates isolated islands, fills holes, and enforces a minimum feature width and gap size compatible with standard fabrication processes. The final design is a  $1,000 \text{ nm} \times 1,000 \text{ nm}$  square unit cell comprising an arbitrarily shaped silicon pattern (refractive index = 3.5) generated using this approach sitting on a glass substrate (refractive index = 1.5).

We generated a dataset of 45,790 metasurface designs with randomly generated control-point grids illuminated by a left-handed circularly polarized (LCP) normally incident wave. These designs were simulated using the commercial software package Lumerical FDTD on a server with two Intel(R) Xeon(R) Gold 6258R CPUs and 1.5 TB memory. After every simulation, the transmission spectrum was recorded at 31 uniformly spaced wavelengths from 1,050 nm to 1,600 nm. The completed dataset was randomly partitioned into training and test sets with a 4:1 ratio, resulting in 36,632 training samples and 9,158 test samples.

To prepare the geometrical-optical pairs for the forward-prediction task using LLMs, each  $4 \times 4$  control-point grid is converted into a natural-language prompt and its 31-element transmission vector into the

corresponding completion. A typical prompt would be: “We have a 4-by-4 grid:  $[[g_{11}, \dots, g_{14}], \dots, [g_{41}, \dots, g_{44}]]$ , what is the transmission spectrum of the metasurface generated using this grid?” while the target completion would then be: “The transmission values sampled at 31 evenly spaced points between 1,050 nm and 1,600 nm for the metasurface generated using this grid are  $[t_1, \dots, t_{31}]$ ”. All numerical values are rounded to three decimal places, a choice that is not accuracy-limiting at the error scales considered, balancing GPU memory usage with predictive accuracy, and aligns with prior works [26], [27]. All prompts and expected outputs are tokenised with the same byte-pair encoder as the base model to ensure vocabulary consistency. Fine-tuning proceeds by feeding these prompt-completion pairs to the LLM and minimising the loss between the predicted tokens and the ground-truth completion, as illustrated in Figure 1b. This formulation re-casts spectrum prediction as language-sequence completion, allowing us to exploit the LLMs’ autoregressive training objective without architectural modifications.

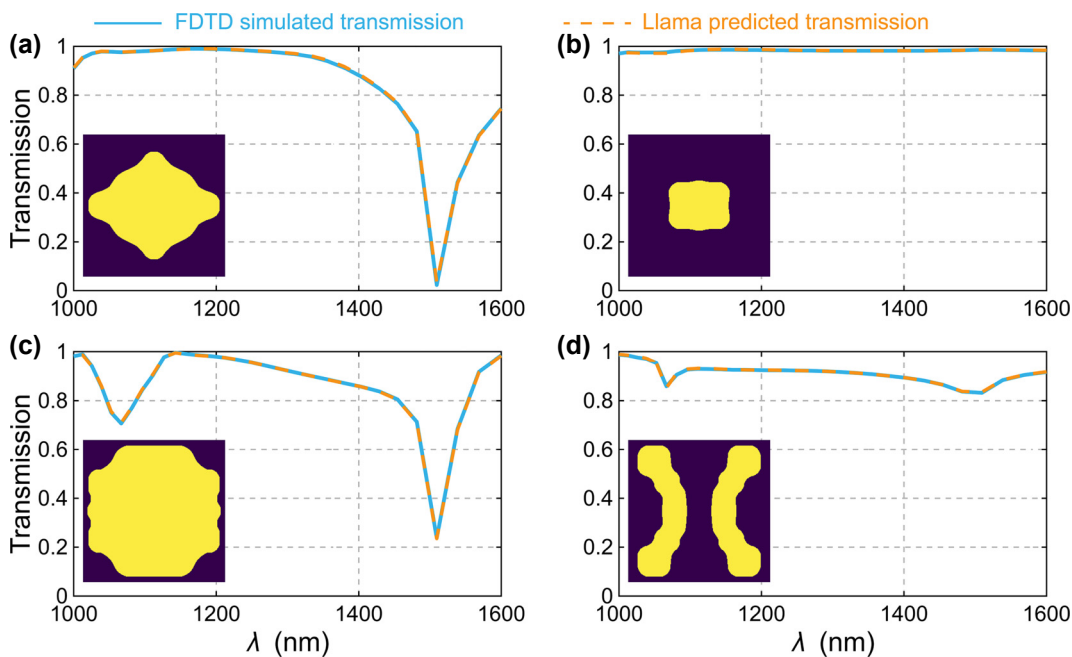
The LLM implemented in both the forward and inverse design process is Meta-Llama-3.1-8B-Instruct [25] quantised to 4 bit weights by Unsloth. This LLM is identical to that employed by Kim et al. [26], eliminating the need for neural network engineering. Parameter-efficient adaptation is

realised with low-rank adaptation (LoRA [24]), which injects low-rank adapters into all projection layers. The entire workflow is built using open-source libraries including Pytorch, HuggingFace, and Unsloth. All 7–9B parameter LLMs are trained on a single NVIDIA RTX 2080 Ti GPU, while the larger and smaller LLMs used in the benchmarking process are trained on one NVIDIA L40S GPU. The rank  $r$  and the scaling factor  $\alpha$  were both set to 32. Fine-tuning proceeds for 8 epochs with an effective batch size of 192, using the AdamW optimiser with an initial learning rate of  $4.0 \times 10^{-4}$  followed by linear decay and the standard cross-entropy objective for next-token prediction. This setup resulted in approximately 10 GB of GPU memory usage for 7–9B models, a requirement that is met by most contemporary commercially available consumer graphics cards.

## 3 Results and discussion

### 3.1 Forward design

To demonstrate the prediction accuracy, Figure 2 compares the spectra predicted by our fine-tuned Llama-3.1-8B with FDTD simulation results for four representative meta-atoms. The orange dashed curves (Llama) track the blue solid curves (FDTD) almost perfectly across

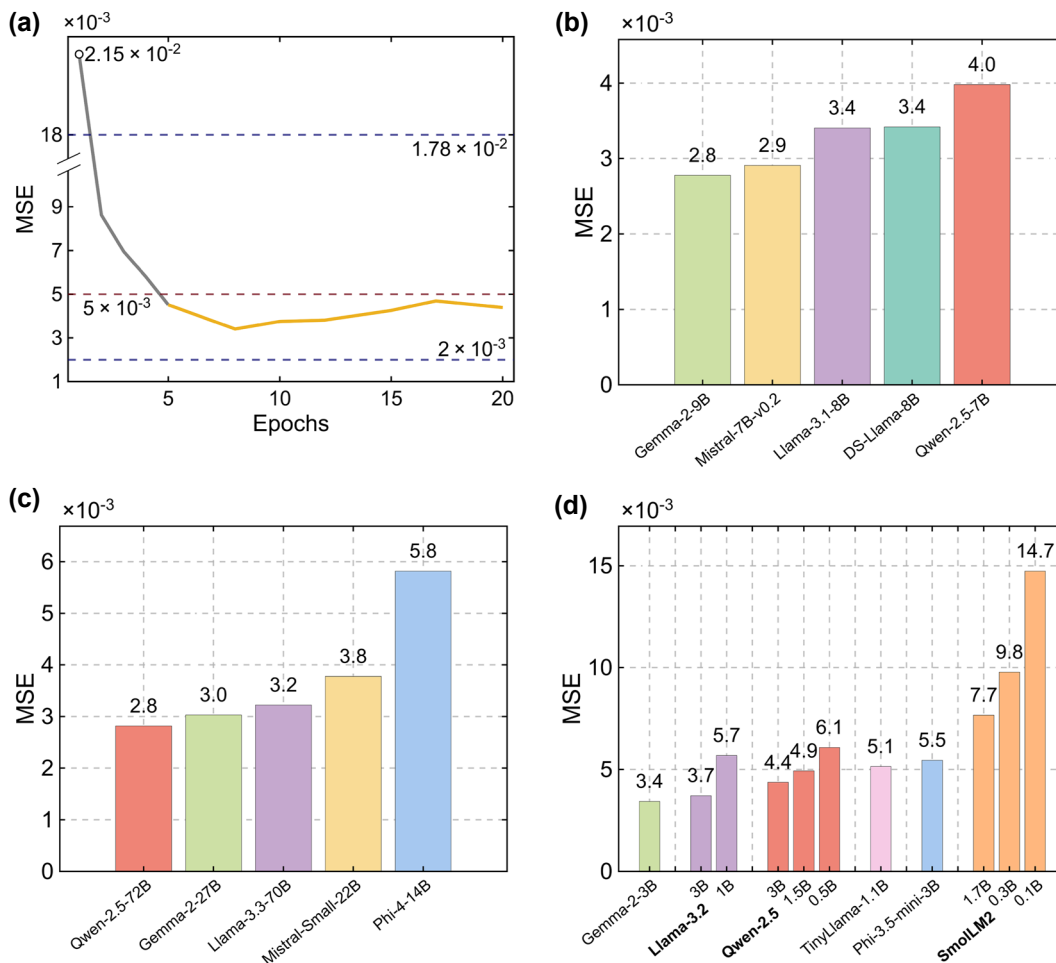


**Figure 2:** Predicted and simulated transmission spectra for four grids from the test set. The corresponding control-point grids and MSE are: (a)  $[[0.411, 0.795, 0.126, 0.233], [0.876, 0.187, 0.209, 0.911], [0.318, 0.479, 0.998, 0.826], [0.555, 0.820, 0.238, 0.058]]$ ,  $\text{MSE} = 7.8 \times 10^{-6}$ . (b)  $[[0.156, 0.485, 0.350, 0.248], [0.391, 0.476, 0.083, 0.444], [0.041, 0.419, 0.524, 0.511], [0.695, 0.026, 0.690, 0.560]]$ ,  $\text{MSE} = 2.6 \times 10^{-6}$ . (c)  $[[0.203, 0.155, 0.608, 0.655], [0.682, 0.541, 0.924, 0.898], [0.660, 0.610, 0.193, 0.065], [0.145, 0.508, 0.538, 0.098]]$ ,  $\text{MSE} = 3.6 \times 10^{-6}$ . (d)  $[[0.049, 0.881, 0.405, 0.843], [0.288, 0.836, 0.375, 0.149], [0.736, 0.211, 0.728, 0.012], [0.471, 0.181, 0.914, 0.007]]$ ,  $\text{MSE} = 4.1 \times 10^{-6}$ .

the 1,000–1,600 nm band, faithfully reproducing both plateaus and sharp resonances. Querying the model is straightforward: copy and paste the  $4 \times 4$  control-point matrix into a prompt, as discussed previously, and then read out the 31-point spectrum returned by the LLM. On a single RTX 2080 Ti GPU, this prediction takes approximately 2 s, about 60 times faster than the corresponding full-wave simulation on our CPU cluster. Across the entire 9158 sample test set the mean squared error (MSE) is  $3.4 \times 10^{-3}$  when trained for 8 epochs, matching specialised DNNs reported in the literature [31], [34]. In other words, a lightly fine-tuned (*i.e.*, only the number of epochs needs tuning), off-the-shelf LLM delivers turnkey, high-fidelity forward modelling without any bespoke network design or hyper-parameter

sweeps, demonstrating a practical “no code” path to rapid metasurface prototyping.

Furthermore, to quantify the influence of fine-tuning epochs, we tested how different training epochs impact the prediction accuracy. Figure 3a confirms that fine-tuning length is a minor knob for this prediction task. Specifically, when the Llama-3.1–8B model is fine-tuned for between 5 and 20 epochs, a typical range for fine-tuning LLMs, its test-set MSE meanders within  $3.4\text{--}4.7 \times 10^{-3}$ , comfortably below the  $5 \times 10^{-3}$  marked by the red dashed line. In contrast, a hand-built eight-layer fully connected network swings from a best-case  $2.0 \times 10^{-3}$  to  $1.78 \times 10^{-2}$  after enlarging each layer from 512 to 895 neurons, marked by the two dark-blue dashed lines. Custom DNNs can still edge out the LLMs by a small margin, but only at the cost of exhaustive architecture



**Figure 3:** Effect of fine-tuning duration and model size on forward-prediction MSE. (a) Test-set MSE for Llama-3.1–8B versus fine-tuning epochs. Although the MSE exceeds the  $5 \times 10^{-3}$  tolerance line (red dashed line) during epochs 1–4 (grey curve), once fine-tuning reaches epoch 5 the orange curve remains consistently below this tolerance and only marginally above the  $2 \times 10^{-3}$  benchmark reached by the best hand-tuned eight-layer DNNs (blue dashed line), indicating that predictive accuracy is largely insensitive to training length within a certain range. (b–d) MSE after eight-epoch LoRA fine-tuning for open-weight models grouped by size: (b) mid-size checkpoints (7–9B parameters). DS-Llama-8B stands for DeepSeek-distilled Llama-3.1–8B; (c) large models (> 9B); (d) small models (< 7B).



searches. Hence, a pragmatic workflow is to deploy LLMs for rapid evaluation and reserve heavyweight solvers or customized networks for final, high-precision refinement stages. We note that the proposed workflow is excitation-agnostic: adapting to other illumination conditions such as different polarization or incidence angle simply requires changing the simulation setup and regenerating the corresponding dataset with no architectural changes required [35].

However, not all instruction-tuned LLMs are well-suited for numerical regression. When we fine-tuned Qwen-3B, a reasoning model, it repeatedly abandoned the target format and instead produced a verbose diagnostic such as: *“To determine the transmission spectrum of the metasurface generated using the given 4-by-4 grid, we need to analyze the optical properties of the grid. Each value in the grid represents the transmission coefficient for a specific wavelength or frequency. However, the provided grid does not include information about the wavelength range or the specific frequencies corresponding to each value. Additionally, the transmission spectrum typically requires information about the incident light, the material properties, and the geometry of the metasurface. Without this information, it is not possible to accurately compute the transmission spectrum. Please provide additional details about the wavelength range, incident light parameters, and material properties to proceed with the calculation.”* Other similar reasoning LLMs, such as Phi-4-Reasoning, Llama-4, gpt-oss, and Gemma-3, exhibit similar behaviours during our fine-tuning. Such chain-of-thought digressions reveal that strong conversational priors can overshadow the supervised objective, prompting the model to seek more information from the user rather than produce the requested 31 transmission values. Therefore, reasoning-centric LLMs may demand additional engineering before they serve reliably as high-throughput, numeric predictors in scientific design loops.

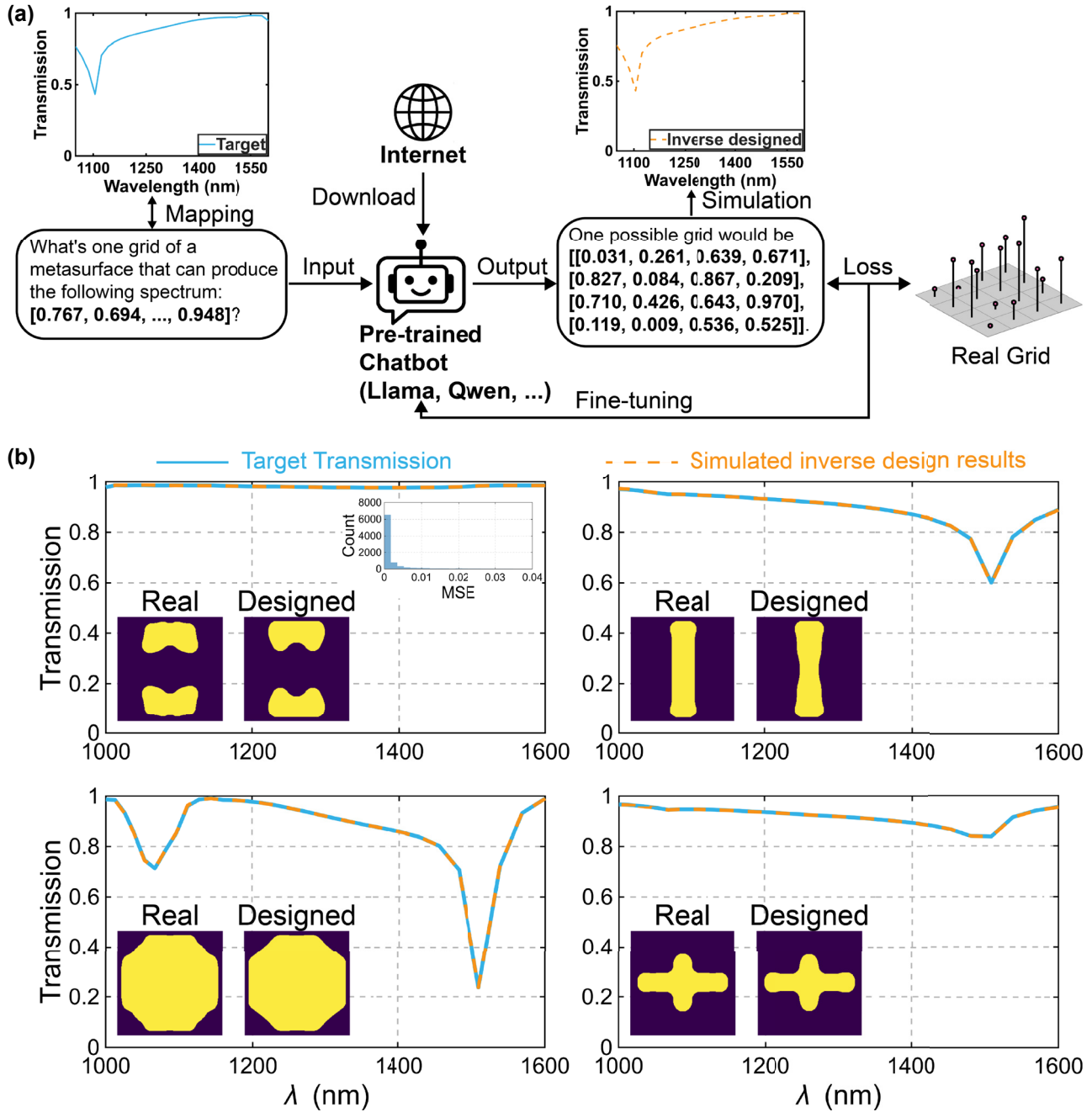
## 3.2 Benchmarking

To assess how sensitive our workflow is to model choice, we fine-tuned eleven open-weight LLMs spanning three parameter bands “small” ( $< 7$  B), “mid-size” ( $7\text{--}9$  B), and “large” ( $> 9$  B) on the same training-test split and fine-tuning setup, and summarized the resulting test-set MSEs in Figure 3b–d. Note that these regions are defined solely to show the feasibility of our method based on commonly used consumer-grade GPUs, rather than to align with definitions used in the machine learning community. Larger models are more sample-efficient during fine-tuning [36], and increasing epochs or fine-tuning data for larger models

leads to diminishing returns [37]. Thus, the fine-tuning configuration used for mid-sized models is sufficient for other regions. In the mid-size model regime (Figure 3b), accuracy generally improves with increasing size but not strictly monotonically: the 7B Qwen checkpoint reaches  $4.0 \times 10^{-3}$ , and the 9B Gemma variant levels off at  $2.8 \times 10^{-3}$ , indicating that entry-level GPUs can deliver spectra of acceptable fidelity. But Mistral 7B showed better MSE than the 8B Llama variant, illustrating that architecture and internal design can outweigh simple parameter count increase. Accordingly, the size-accuracy gains discussed in Figure 3b are best viewed as a trend rather than a strict rule. The small-model sweep (Figure 3d) reinforces this point: Gemma-2-3B achieves  $3.4 \times 10^{-3}$ , whereas the tiny SmoLLM2-0.1 B variants drift above  $14.7 \times 10^{-3}$ . However, scaling further yields diminishing returns. In particular, enlarging Qwen-2.5 from 7B to 72B shaves only  $1.2 \times 10^{-3}$  off the MSE yet stretches inference to almost 35 s and consumes the full 48 GB memory of a single NVIDIA L40S GPU (Figure 3c). Gemma models rank first or second across all size bands, further suggesting that architectural priors outweigh raw parameter count in certain ranges. A plausible reason Gemma advances across models in our task is its larger, digit-friendly tokenizer, which represents decimals more regularly. Note that no deeper architectural tests are investigated here because the goal of our study is to provide a clear, out-of-the-box workflow that lets non-AI practitioners accelerate photonics design. Taken together, these benchmarks show that: (i) model selection can be guided by simple size thresholds rather than exhaustive hyper-parameter searches: changing the model size within the LLM family produces only modest accuracy shifts. In contrast, the DNN baseline shows a much larger spread across sizes. (ii) Gemma variants currently offer the best accuracy-to-cost ratio for rapid prototyping, and (iii) future gains are likely to come from designs that embed stronger numerical priors or VLMs rather than from continued parameter scaling alone.

## 3.3 Inverse design

Inverse metasurface design is fundamentally many-to-one: distinct geometries produce near-identical spectra, so a deterministic inverse network receives conflicting labels and its gradients cancel, stalling training, leading to non-convergence problems. Conventional remedies such as tandem networks, where an inverse generator is optimized through a frozen forward model [38], ease convergence but often collapse to a single prototype [10] and inherit the surrogate’s biases, thereby limiting design diversity [39]. Leveraging the intrinsic stochasticity of LLMs circumvents this problem. As sketched in Figure 4a, we encode a 31-point



**Figure 4:** LLM-based inverse design. (a) Workflow of the inverse-design stage. A target 31-point transmission spectrum is fed to the fine-tuned Llama-3.1-8B as a natural-language query of a corresponding grid; the model autoregressively returns a control-point grid that defines a candidate meta-atom. (b) Representative results for four unseen targets. The orange dashed lines are FDTD simulated results of inverse-designed metasurfaces. The corresponding inverse-designed grids and MSE are: top-left:  $[[0.550, 0.073, 0.906, 0.559], [0.324, 0.326, 0.831, 0.708], [0.916, 0.060, 0.517, 0.120], [0.023, 0, 0.249, 0.263]]$ ,  $MSE = 2.0 \times 10^{-7}$ ; top-right:  $[[0.360, 0.903, 0.903, 0.822], [0.419, 0.386, 0.377, 0.962], [0.744, 0.397, 0.391, 0.742], [0.890, 0.048, 0.259, 0.686]]$ ,  $MSE = 1.2 \times 10^{-6}$ ; bottom-left:  $[[0.460, 0.289, 0.513, 0.473], [0.199, 0.641, 0.932, 0.866], [0.757, 0.956, 0.755, 0.282], [0.9120, 0.571, 0.547, 0.876]]$ ,  $MSE = 1.4 \times 10^{-6}$ ; bottom-right:  $[[0.964, 0.207, 0.656, 0.287], [0.777, 0.548, 0.192, 0.460], [0.181, 0.202, 0.218, 0.812], [0.303, 0.866, 0.496, 0.582]]$ ,  $MSE = 3.0 \times 10^{-7}$ . The histogram within the top-left figure depicts the inverse-design test-set MSE distribution, showing that over 88 % of samples achieve an MSE below  $1 \times 10^{-2}$ .

transmission vector into the prompt "What's one grid of a metasurface that can produce the following spectrum:  $[t_1, \dots, t_{31}]$ ", invite the model to return "One possible grid would

be  $[[g_{11}, \dots, g_{14}], \dots, [g_{41}, \dots, g_{44}]]$ " and parse the tokens into the control-point grids. The deliberate phrasing "one possible" signals the multiplicity of valid answers explicitly,

allowing the fine-tuned Llama-3.1–8B to learn that several candidates can generate similar spectra. After fine-tuning for 8 epochs, the Llama proposes a grid in about 0.9 s on a single RTX 2080 Ti GPU. Figure 4b (and more examples in Figure S1 of Supplementary Information) demonstrates four such inverse-designed meta-atoms: their FDTD-validated spectra (orange dashed) closely track the targets (blue solid) while their geometries differ markedly, confirming both fidelity and diversity without the need for techniques typically used in customized DNN approaches to mitigate non-convergence problems. Collectively, these results position LLMs as a fast, versatile alternative for inverse electromagnetic design. To compare to simple inverse baselines, we also implement a classical tandem inverse network, where an inverse network (spectrum to control points) is trained through a frozen forward network. Architecture details and representative results are provided in Figure S2 and S3 of Supplementary Information, with detailed observations from the comparison.

## 4 Conclusions

In summary, this work demonstrates that one-dimensional token-wise LLMs can serve as a practical “chat-to-chip” solution for both forward and inverse design of two-dimensional arbitrarily shaped metasurfaces without the need for vision models. Systematic benchmarking across widely used open-weight LLM checkpoints not only quantifies performance but also supplies a clear reference for future research. Collectively, these findings lower the barrier to entry for nanophotonic researchers who lack machine learning expertise and foreshadow a design paradigm in which LLMs drive rapid, automated exploration of increasingly complex metasurfaces and multifunctional electromagnetic devices.

**Research funding:** This work was supported by the John L. and Genevieve H. McCain endowed chair professorship at the Pennsylvania State University.

**Author contribution:** All authors have accepted responsibility for the entire content of this manuscript and consented to its submission to the journal, reviewed all the results and approved the final version of the manuscript. HZ and LK conceived the idea. HZ designed the experiments, developed the model code, performed the simulations. All authors contributed to the preparation of manuscript.

**Conflict of interest:** Authors state no conflict of interest.

**Data availability:** The datasets generated and analysed during the current study are available from the corresponding author upon reasonable request.

## References

- [1] T. J. Cui, *et al.*, “Roadmap on electromagnetic metamaterials and metasurfaces,” *J. Phys. Photonics*, vol. 6, no. 3, p. 032502, 2024.
- [2] M. Khorasaninejad and F. Capasso, “Metalenses: Versatile multifunctional photonic components,” *Science*, vol. 358, no. 6367, p. eaam8100, 2017.
- [3] L. Huang, S. Zhang, and T. Zentgraf, “Metasurface holography: From fundamentals to applications,” *Nanophotonics*, vol. 7, no. 6, pp. 1169–1190, 2018.
- [4] Y. Ding, *et al.*, “Waveguide-based augmented reality displays: Perspectives and challenges,” *eLight*, vol. 3, no. 1, p. 24, 2023.
- [5] Y. Hao and R. Mittra, *FDTD Modeling of Metamaterials: Theory and Applications*, Boston London, Artech House, 2009.
- [6] P. P. Silvester and R. L. Ferrari, *Finite Elements for Electrical Engineers*, 3rd ed., New York, Cambridge University Press, 1996.
- [7] M. M. R. Elsayy, S. Lanteri, R. Duvinneau, J. A. Fan, and P. Genevet, “Numerical optimization methods for metasurfaces,” *Laser Photon. Rev.*, vol. 14, no. 10, p. 1900445, 2020.
- [8] Z. Li, S. Yu, and G. Zheng, “Advances in exploiting the degrees of freedom in nanostructured metasurface design: From 1 to 3 to more,” *Nanophotonics*, vol. 9, no. 12, pp. 3699–3731, 2020.
- [9] S. D. Campbell, and D. H. Werner, Eds. *Advances in Electromagnetics Empowered by Artificial Intelligence and Deep Learning*, Hoboken, New Jersey, Wiley-IEEE Press, 2023.
- [10] A. Khairah-Walieh, D. Langevin, P. Bennet, O. Teytaud, A. Moreau, and P. R. Wiecha, “A newcomer’s guide to deep learning for inverse design in nano-photonics,” *Nanophotonics*, vol. 12, no. 24, pp. 4387–4414, 2023.
- [11] D. Lee, W. Chen, L. Wang, Y. Chan, and W. Chen, “Data-driven design for metamaterials and multiscale systems: A review,” *Adv. Mater.*, vol. 36, no. 8, p. 2305254, 2024.
- [12] S. An, *et al.*, “Deep convolutional neural networks to predict mutual coupling effects in metasurfaces,” *Adv. Opt. Mater.*, vol. 10, no. 3, p. 2102113, 2022.
- [13] Z. Zhang, C. Yang, Y. Qin, H. Feng, J. Feng, and H. Li, “Diffusion probabilistic model based accurate and high-degree-of-freedom metasurface inverse design,” *Nanophotonics*, vol. 12, no. 20, pp. 3871–3881, 2023.
- [14] S. D. Campbell, R. P. Jenkins, P. J. O’Connor, and D. Werner, “The explosion of artificial intelligence in antennas and propagation: How deep learning is advancing our state of the art,” *IEEE Antennas Propag. Mag.*, vol. 63, no. 3, pp. 16–27, 2021.
- [15] I. Malkiel, M. Mrejen, A. Nagler, U. Arieli, L. Wolf, and H. Suchowski, “Plasmonic nanostructure design and characterization via deep learning,” *Light Sci. Appl.*, vol. 7, no. 1, p. 60, 2018.
- [16] S. An, *et al.*, “Deep learning modeling approach for metasurfaces with high degrees of freedom,” *Opt. Express*, vol. 28, no. 21, p. 31932, 2020.
- [17] W. Chen, *et al.*, “Broadband solar metamaterial absorbers empowered by transformer-based deep learning,” *Adv. Sci.*, vol. 10, no. 13, p. 2206718, 2023.
- [18] H. Zhang, L. Kang, S. D. Campbell, K. Zhang, D. H. Werner, and Z. Cao, “Fixed-attention mechanism for deep-learning-assisted design of high-degree-of-freedom 3D metamaterials,” *Opt. Express*, vol. 33, no. 9, p. 18928, 2025.
- [19] H. Zhang, L. Kang, S. D. Campbell, J. T. Young, and D. H. Werner, “Data driven approaches in nanophotonics: A review of AI-enabled



- metadevices,” *Nanoscale*, 2025, <https://doi.org/10.1039/D5NR02043C>.
- [20] D. Lakhmiri, S. L. Digabel, and C. Tribes, “HyperNOMAD: Hyperparameter optimization of deep neural networks using mesh adaptive direct search,” 2019, *arXiv*: arXiv:1907.01698, <https://doi.org/10.48550/arXiv.1907.01698>.
- [21] Y. Chang, *et al.*, “A survey on evaluation of large language models,” *ACM Trans. Intell. Syst. Technol.*, vol. 15, no. 3, pp. 1–45, 2024.
- [22] H. Naveed, *et al.*, “A comprehensive overview of large language models,” 2024, *arXiv*: arXiv:2307.06435, <https://doi.org/10.48550/arXiv.2307.06435>.
- [23] T. Dinh, *et al.*, “LIFT: Language-interfaced fine-tuning for non-language machine learning tasks,” 2022, *arXiv*: arXiv:2206.06565, <https://doi.org/10.48550/arXiv.2206.06565>.
- [24] E. J. Hu, *et al.*, “LoRA: Low-rank adaptation of large language models,” 2021, *arXiv*: arXiv:2106.09685, <https://doi.org/10.48550/arXiv.2106.09685>.
- [25] A. Grattafiori, *et al.*, “The Llama 3 herd of models,” 2024, *arXiv*: arXiv:2407.21783, <https://doi.org/10.48550/arXiv.2407.21783>.
- [26] M. Kim, H. Park, and J. Shin, “Nanophotonic device design based on large language models: Multilayer and metasurface examples,” *Nanophotonics*, vol. 14, no. 8, pp. 1273–1282, 2025.
- [27] D. Lu, Y. Deng, J. M. Malof, and W. J. Padilla, “Learning electromagnetic metamaterial physics with ChatGPT,” *IEEE Access*, vol. 13, pp. 51513–51526, 2025.
- [28] X. Liu, *et al.*, “Design of circularly polarized phosphorescence materials guided by transfer learning,” *Nat. Comm.*, vol. 16, no. 1, 2025, <https://doi.org/10.1038/s41467-025-60310-6>.
- [29] R. Lupoiu, Y. Shao, T. Dai, C. Mao, K. Edee, and J. A. Fan, “A multi-agentic framework for real-time, autonomous freeform metasurface design,” 2025, *arXiv*, <https://doi.org/10.48550/ARXIV.2503.20479>.
- [30] E. B. Whiting, S. D. Campbell, L. Kang, and D. H. Werner, “Meta-atom library generation via an efficient multi-objective shape optimization method,” *Opt. Express*, vol. 28, no. 16, p. 24229, 2020.
- [31] R. P. Jenkins, S. D. Campbell, and D. H. Werner, “Establishing exhaustive metasurface robustness against fabrication uncertainties through deep learning,” *Nanophotonics*, vol. 10, no. 18, pp. 4497–4509, 2021.
- [32] J. Tang, *et al.*, “GraphGPT: Graph instruction tuning for large language models,” 2024, *arXiv*: arXiv:2310.13023, <https://doi.org/10.48550/arXiv.2310.13023>.
- [33] S. Yin, *et al.*, “A survey on multimodal large language models,” *Natl. Sci. Rev.*, vol. 11, no. 12, p. nwae403, 2024.
- [34] Y. Dong, *et al.*, “Advanced deep learning approaches in metasurface modeling and design: A review,” *Prog. Quant. Electron.*, vol. 99, p. 100554, 2025.
- [35] J. Jiang and J. A. Fan, “Global optimization of dielectric metasurfaces using a physics-driven neural network,” *Nano Lett.*, vol. 19, no. 8, pp. 5366–5372, 2019.
- [36] J. Kaplan, *et al.*, “Scaling laws for neural language models,” 2020, *arXiv*, <https://doi.org/10.48550/ARXIV.2001.08361>.
- [37] B. Zhang, Z. Liu, C. Cherry, and O. Firat, “When scaling meets LLM finetuning: The effect of data, model and finetuning method,” 2024, *arXiv*, <https://doi.org/10.48550/ARXIV.2402.17193>.
- [38] D. Liu, Y. Tan, E. Khoram, and Z. Yu, “Training deep neural networks for the inverse design of nanophotonic structures,” *ACS Photonics*, vol. 5, no. 4, pp. 1365–1369, 2018.
- [39] J. Chen, *et al.*, “An improved tandem neural network architecture for inverse modeling of multicomponent reactive transport in porous media,” *Water Resour. Res.*, vol. 57, no. 12, 2021, <https://doi.org/10.1029/2021WR030595>.

---

**Supplementary Material:** This article contains supplementary material (<https://doi.org/10.1515/nanoph-2025-0343>).